VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

ARTICLES

DEEPFAKE FIGHT: AI-POWERED DISINFORMATION AND PERFIDY UNDER THE GENEVA CONVENTIONS

Major D. Nicholas Allen

DEEPFAKE FIGHT: AI-POWERED DISINFORMATION AND PERFIDY UNDER THE GENEVA CONVENTIONS*

MAJOR D. NICHOLAS ALLEN**

All that we are not stares back at what we are.

- W.H. Auden[1]

INTRODUCTION

On February 24, 2022, missiles began hitting major cities across the country: Kyiv, Kharkiv, Chernihiv.[2] Russian infantry, armor, mechanized fighting vehicles, mobile artillery, aviation, trucks, and supply assets charged over Ukraine's border at every point of the compass except west. The war the world feared for years would happen, that had actually *been* happening but on a smaller, deniable scale, started.[3]

---

* The views, opinions, and assertions provided in this article, notwithstanding those cited, are the views, opinions, and assertions of the author alone. This article does not necessarily reflect the views or positions of the United States Army, the Department of Defense, or the United States government.

**Judge Advocate, United States Army. Presently assigned as Chief of National Security Law, 25th Infantry Division, Schofield Barracks, Hawaii. L.L.M. in Military Law, 2021, The Judge Advocate General's School, United States Army, Charlottesville, Virginia; J.D., 2010, University of Baltimore School of Law; B.A., 2006, University of Florida. Previous assignments include Command Judge Advocate, United States Army Security Assistance Training Management Organization, Fort Bragg, North Carolina, 2018-2020; Defense Counsel, Fort Bragg Trial Defense Service Field Office, Fort Bragg, North Carolina, 2016-2018; Battalion Judge Advocate, 2nd Battalion, 3rd Special Forces Group (Airborne), Fort Bragg, North Carolina, 2014-2016; Trial Counsel, Fort Jackson, South Carolina, 2013-2014; Legal Assistance Attorney, Office of the Staff Judge Advocate, Fort Jackson, South Carolina, 2012-2013. Member of the bar of Maryland. The author wishes to thank the editors and staff of the Notre Dame Journal on Emerging Technologies as well as the myriad mentors, colleagues, and friends who assisted with this article. Most of all the author thanks his wife Anna and his children Jackson and Finley for their boundless love and support.

[1] W.H. AUDEN, THE SEA AND THE MIRROR 204 (1944).

[2] *See e.g.* Madeline Fitzgerald, *Russia Invades Ukraine: A Timeline of the Crisis*, U.S. NEWS & WORLD REP. (Feb. 25, 2022, 5:49 PM), https://www.usnews.com/news/best-countries/slideshows/a-timeline-of-the-russia-ukraine-conflict; John Psaropoulos, *Timeline: The First 100 Days of Russia's War in Ukraine*, AL JAZEERA (Jun. 3, 2022), https://www.aljazeera.com/features/2022/6/3/timeline-the-first-100-days-of-russias-war-in-ukraine.

[3] *Id.* Deniability was a key component of Russia's hybrid military involvement in Ukraine when it invaded the Crimean Peninsula in 2014, doing so with troops sent from its territory and armed with its weapons and equipment but lacking any

But the expected quick Russian victory did not materialize. In the following days the Ukrainian military fought harder and better than Russia had planned for, resulting in thousands of Russian troops killed, hundreds of Russian combat vehicles destroyed, and almost none of Russia's apparent major military objectives achieved.[4] Russian forces also slogged through self-inflicted logistics woes which further degraded Russian forces' abilities to maneuver, caused many Russian crews to abandon their vehicles across Ukraine, and quickly became a point of tremendous embarrassment for Russian military leaders.[5]

In the public relations sphere Russia would be in arguably its deepest hole. Worldwide condemnation of its invasion would feed an enormous sanctions regime,[6] a strengthening among NATO alliances as

---

identifying features or flags. The troops became known pejoratively around the world as "Little Green Men." Russian President Vladimir Putin eventually admitted the obvious shortly after his forces secured the Crimean Peninsula. *See e.g.* Silvia Aloisi & Frank Jack Daniel (eds.), *Timeline: The Events Leading up to Russia's Invasion of Ukraine*, REUTERS (Feb. 28, 2022, 11:03 PM), https://www.reuters.com/world/europe/events-leading-up-russias-invasion-ukraine-2022-02-28/; Vitaly Shevchenko, *"Little Green Men or Russian Invaders?",* BBC (Mar. 11, 2014), https://www.bbc.com/news/world-europe-26532154; Steven Pifer, *Watch Out for Little Green Men*, BROOKINGS (Jul. 7, 2014), https://www.brookings.edu/opinions/watch-out-for-little-green-men/. These same hybrid forces would also aid separatists in the eastern Ukrainian Luhansk and Donetsk regions during years of fighting against the armed forces of Ukraine prior to Russia's all-out invasion in 2022. *Id.*

[4] *See supra* note 2; *see also* Paul D. Shinkman, *Russia Abandons March on Kyiv, Focuses Embattled Troops Instead on Donbas*, U.S. NEWS & WORLD REP. (Mar. 25, 2022 at 3:29 PM), https://www.usnews.com/news/world-report/articles/2022-03-25/russia-abandons-Mar.-on-kyiv-focuses-embattled-troops-instead-on-donbas.

[5] *See supra* note 2; *see also* Anna Ahronheim, *Fuel and Logistics Problems Frustrate Russian Advance*, JERUSALEM POST (Feb. 27, 2022 at 2:39 PM), https://www.jpost.com/international/article-698800; Bonnie Berkowitz & Artur Galocha, *Why the Russian Military is Bogged Down by Logistics in Ukraine,* WASH. POST (Mar. 30, 2022 at 10:17 AM), https://www.washingtonpost.com/world/2022/03/30/russia-military-logistics-supply-chain/; Brad Lendon, *What Images of Russia's Trucks Say About its Military's Struggles in Ukraine*, CNN (Apr. 14, 2022 at 12:06 AM), https://www.cnn.com/2022/04/14/europe/ukraine-war-russia-trucks-logistics-intl-hnk-ml/index.html; Ann Marie Dailey, *What's Behind Russia's Logistical Mess in Ukraine? A US Army Engineer Looks at the Tactical Level*, ATL. COUNCIL (Mar. 21, 2022), https://www.atlanticcouncil.org/blogs/new-atlanticist/whats-behind-russias-logistical-mess-in-ukraine-a-us-army-engineer-looks-at-the-tactical-level/.

[6] *See* Chad P. Bown, *Russia's War on Ukraine: A Sanctions Timeline*, PETERSON INST. FOR INT'L. ECON. (Jul. 1, 2022 at 12:45 PM), https://www.piie.com/blogs/realtime-economic-issues-watch/russias-war-ukraine-sanctions-timeline; *see also List of Sanctions Against Russia After it Invaded Ukraine*, AL JAZEERA (Mar. 3, 2022 at 12:04 PM), https://www.aljazeera.com/news/2022/2/25/list-of-sanctions-on-russia-after-invasion.

well as potential expansion of NATO,[7] and the growth of Ukrainian President Volodymyr Zelenskyy as an international hero figure.[8] Even at home Moscow would have to confront a significant counter swell among the Russian people, leading Moscow to resort to Soviet-style tactics of mass arrests, severe free speech restrictions, and intimidations to suppress the dissent movement.[9]

On March 16, 2022, a new tactic emerged. Ukraine 24, a major television news network in Ukraine, broadcast a quixotic video of Ukrainian President Zelenskyy imploring his troops, not to push to victory, but to surrender.[10] In a motif similar to his daily press briefings and which would have been familiar to his daily viewers, President Zelenskyy appeared behind a podium with short-crop hair, a thin growth of beard, wearing an olive-green shirt, and with presidential symbols in the background. However, instead of his usual remarks encouraging

---

[7] *See supra* note 2; *see also Finland and Sweden Submit Applications to Join NATO*, N. ATL. TREATY ORG. (May 18, 2022 at 9:08 AM), https://www.nato.int/cps/en/natohq/news_195468.htm; A. Wess Mitchell, *Putin's War Backfires as Finland, Sweden Seek to Join NATO,* U.S. INST. OF PEACE (May 26, 2022), https://www.usip.org/publications/2022/05/putins-war-backfires-finland-sweden-seek-join-nato. While Türkiye initially opposed Finland and Sweden's joining NATO, significantly slowing full acceptance, Türkiye has since dropped its opposition by signing a tripartite agreement with Finland and Sweden which now paves the way for the two countries to become NATO's newest member states. George Wright, *Turkey Supports Finland and Sweden NATO Bid*, BBC (Jun. 29, 2022), https://www.bbc.com/news/world-europe-61971858.

[8] *See* Laura King, *Waging War, Wielding Words: Zelenksy's Speeches Have Made Him a Folk Hero*, LOS ANGELES TIMES (Mar. 16, 2022 at 1:28 PM), https://www.latimes.com/world-nation/story/2022-03-16/ukraine-zelensky-speeches-have-made-him-folk-hero; Nidhi Razdan, *Volodymyr Zelensky: From TV Star to War Hero*, NEW DELHI TELEVISION (Mar. 31, 2022 at 6:02 PM), https://www.ndtv.com/world-news/volodymyr-zelensky-from-tv-star-to-war-hero-full-transcript-2840813.

[9] *See* Courtney Subramaniam & Anna Nemtsova, *In Russia Thousands Defy Police Threats to Protest the Invasion of Ukraine. Can it Make a Difference?*, USA TODAY (Mar. 7, 2022 at 12:00 PM), https://www.usatoday.com/story/news/politics/2022/03/04/russia-ukraine-war-protests/9351061002/?gnt-cfr=1; Anton Troianovski & Valeriya Safronova, *Russia Takes Censorship to New Extremes, Stifling War Coverage*, THE NEW YORK TIMES (Mar. 4, 2022), https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html; Marko Milanovic, *The Legal Death of Free Speech in Russia*, EUR. J. INT'L. L.: EJIL TALK! (Mar. 8, 2022), https://www.ejiltalk.org/the-legal-death-of-free-speech-in-russia/ (comparing current laws in Russia criminalizing the characterization of the Russian invasion of Ukraine as either an "invasion" or a "war" to similar laws from the Soviet Union).

[10] Bobby Allyn, *Deepfake Video of Zelenskyy Could be 'Tip of the Iceberg' in Info War, Experts Warn*, NPR (Mar. 16, 2022 at 8:26 PM), https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia; Jane Wakefield, *Deepfake Presidents Used in Russia-Ukraine War*, BBC (Mar. 18, 2022), https://www.bbc.com/news/technology-60780142.

Ukrainians to remain strong and detailing his armed forces' needs to the world, President Zelenskyy claimed instead that "[b]eing the president was not so easy," that "[i]t didn't work out," "[t]here is no tomorrow," and finally "I advise you to lay down your arms and return to your families. It is not worth dying in this war."[11] A chyron also ran at the bottom of the news broadcast claiming that Ukraine had surrendered.[12]

News agencies and social media companies around the world sped to analyze the video and quickly determined that this realistic video was not actually real at all.[13] Instead it was the most recent employment of a still-young technology – a deepfake.

---

[11] Samantha Cole, *Hacked News Channel and Deepfake of Zelenskyy Surrendering is Causing Chaos Online,* VICE (Mar. 16, 2022 at 7:08 AM), https://www.vice.com/en/article/93bmda/hacked-news-channel-and-deepfake-of-zelenskyy-surrendering-is-causing-chaos-online (providing a rare uncommented version of the entire video). While the entire, unaltered video is otherwise difficult to find due to being removed from social media sites or being flagged for false content, a transcript in Ukrainian of the purported remarks is available on the Way Back internet archive. WAYBACK MACH., https://web.archive.org/web/20220316142015/https://u24.ua/ (last visited Jul. 5, 2022)(Ukrainian-to-English translation provided via Google translate and compared to translation provided in Cole, *id.*).

[12] Cole, *supra* note 11.

[13] *Id.*; *see also* James Pearson & Natalia Zinets, *Deepfake Footage Purports to Show Ukrainian President Capitulating*, REUTERS (Mar. 17, 2022 at 2:16 AM), https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/; Joshua Rhett Miller, *Deepfake Video of Zelensky Telling Ukrainians to Surrender Removed from Social Platforms*, THE NEW YORK POST (Mar. 17, 2022 at 12:20 PM), https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/; Tom Simonite, *A Zelensky Deepfake was Quickly Defeated. The Next One Might Not Be.*, WIRED MAG. (Mar. 17, 2022 at 1:30 PM), https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/.

[*Fig. 1.* Side-by-side stills contrasting the deepfake Zelensky video on the left with a genuine video of President Zelensky on the right making remarks at a news conference days prior.][14]

As of the writing of this article the video has had no discernible direct impact on the battlefield or Ukraine's war effort, likely due to its relatively poor quality.[15] But the confusion it sowed, even if temporary, provided immediate and worldwide effects in the information space[16] and demanded priceless time and attention from President Zelenskyy and members of his administration to rebut.

The episode remains a clarion call to those who contemplate the future of media manipulation and digital deception. The evolutionary march of digital deception leads straight to the battlefield, and few capabilities when at their highest potential are better primed to cause confusion and chaos in the battlefield's information space than deepfake technology.

"Deepfake" is the term associated with ultra-realistic video and audio images created not by human actors but by artificial intelligence. Originally associated with salacious pornography videos that depicted

---

[14] Images at Graham Cluley, *Deepfake President Zelensky Calls on Ukraine to Surrender, as TV Station Hacked*, BITDEFENDER (Mar. 17, 2022), https://www.bitdefender.com/blog/hotforsecurity/deepfake-president-zelensky-calls-on-ukraine-to-surrender-as-tv-station-hacked/.
[15] Simonite, *supra* note 13.
[16] Cole, *supra* note 11.

unwitting victims participating in sex acts,[17] people have used the technology to create perceptually perfect fake videos of such figures as President Barack Obama, celebrities like Emma Watson and Nicolas Cage, or even Russian President Vladimir Putin as early as 2018.[18] The technology has manipulated images of weather patterns and even depicted the life cycle of a daisy without needing human input for guidance.[19]

The Zelenskyy deepfake is also not the first time that a deepfake has made a mark during a time of crisis. In 2019, a deepfake-caused crisis instigated an attempted coup in Gabon, which nearly caused a civil war.[20] Supporters of Gabonese President Ali Bongo Ondimba became convinced that, after the President had not been seen for several days, a video purporting to show President Bongo alive, astute, and on the job was not real but instead was a deepfake. In support of this assumption, citizens pointed to differences in the President's demeanor, physical appearance, his apparent inability to use a hand, and even raised skepticism about the video's lighting.[21] Local newspapers had also speculated about deepfake, and on January 7, 2019, military officers from

---

[17] *See* Thanh Thi Nguyen, et al., Deep Learning for Deepfakes Creation and Detection 2 (Jul. 28, 2020, 17:54 UTC), https://arxiv.org/pdf/1909.11573.pdf; Yisroel Mirsky & Wenke Lee, The Creation and Detection of Deepfakes 1-2 (Sep. 13, 2020, 22:44 UTC), https://arxiv.org/pdf/2004.11138.pdf. This derogatory use of deepfake technology has caused significant harm for hundreds if not thousands of victims since its inception. However, this impact is beyond the scope of this article. For a devoted analysis of deepfake technology and its role in revenge pornography or other related victimizing activities, *see, e.g.*, Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 7-8 (2020); Danielle Citron, *Sexual Privacy*, 128 YALE L.J. 1870, 1898-1902 (2019) (detailing how nonconsensual deepfake pornography videos violate sexual privacy rights); Rebecca A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, 88 FORDHAM L. REV. 887, 895-99 (2019) (discussing the ways that deepfake pornography is used, the harm it causes, and the problems with finding recourse in the law for victims); Russell Spivak, *'Deepfakes': The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 345-48 (2019) (discussing the history of deepfake proliferation from a Reddit user who posted the first deepfake videos to use in nonconsensual pornographic content to comparatively benign modifications of movie and television clips, and describing how private companies financially benefit from evolutions in media manipulation).

[18] *See* Bloomberg Quicktake, *It's Getting Harder to Spot a Deepfake Video* (Sep. 27, 2018), https://www.youtube.com/watch?v=gLoI9hAX9dw/

[19] *Id.*

[20] *See* Sarah Cahlan, *How Misinformation Helped Spark an Attempted Coup in Gabon*, WASH. POST (Feb. 13, 2020, 3:00 AM), https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/; Ali Breland, *The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bring an African Nation to the Brink*, MOTHER JONES (Mar. 15, 2019), https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

[21] Breland, *supra* note 20.

the Gabonese armed forces attempted a coup d'état by forcibly seizing a broadcast station and sending messages in an effort to "restore democracy."[22]

While the coup did not succeed[23] and the video was most likely real,[24] the impact of the episode is enough to give skeptics of deepfake manipulation further pause. No actual manipulation was necessary. Deepfake technology's existence alone brought the country to the edge of non-international armed conflict.[25]

With media manipulation at such new heights, international actors must not neglect its technical and legal impact on the battlefield. This Article therefore attempts to assess the current state of deepfake technology, look ahead to its potential future applications in armed conflict, process the ways in which current law contemplates such deception, and distill recommendations for improving governance where needed.

First, the Article will examine the origins of media manipulation and warfare in order to provide context for the later analysis of where deepfake deception fits in today's information arsenal. Second, the Article will detail the current state of deepfake technology. This discussion will explore the technology's structural roots, in both variational autoencoders and the more popular method via generative

---

[22] The Associated Press, *Gabon's Government Quashes Coup Attempt, Killing 2, Officials Say*, CBC (Jan. 7, 2019, 2:00 AM), https://www.cbc.ca/news/world/gabon-coup-attempt-1.4968177.

[23] Two of the officers were killed in a resulting raid and the others captured. *Id.*

[24] The President as it turns out had suffered a stroke and needed treatment, both of which likely explained his differences in mannerisms and appearance. Cahlan, *supra* note 20; *see also* Janosch Delcker, *Welcome to the Age of Uncertainty*, POLITICO (Dec. 17, 2019, 7:50 PM), https://www.politico.eu/article/deepfake-videos-the-future-uncertainty/. In a cryptic follow-up, later analysis of the video concluded both that the video was "likely" real but also could not rule out that it still could have been a deepfake. *Id.*

[25] While threatening to expand into a non-international armed conflict, this episode would not likely qualify as one under the Tadic Factors as the conflict, while involving a clash between government forces and an armed, uniformed, organized non-governmental force, was not "protracted," having started and ended in a day. *See Prosecutor v. Tadić*, Case No. IT-94-1, Decision on Defence Motion for Interlocutory Appeal on Jurisdiction, ¶ 70 (Int'l Crim. Trib. for the Former Yugoslavia Oct. 2, 1995) (finding that an armed conflict exists "whenever there is a resort to armed force between States or protracted armed violence between governmental authorities and organized armed groups or between such groups within a State."). Furthermore, in finding that the conflict in the Balkans qualified as "protracted," the International Criminal Tribunal for the former Yugoslavia observed that conflict between State and non-State forces had existed for years and involved "large-scale violence." *Id.* Neither of those facts presented in Gabon, though nothing about deepfake technology mitigated those possibilities.

adversarial networks, to show how deepfake technology can be complex yet accessible to trends and expected future advances. Third, the Article will detail abilities and limits for detecting deepfake manipulations and will analyze methods for determining attribution for an act of deepfake-derived deception both in a technological sense and a legal sense. Fourth, the Article will discuss the laws that may impact uses of deepfake technology in armed conflict. This discussion will look chiefly through a *jus in bello* lens to confront the conflict that arises when international humanitarian laws which permit misinformation may have to thwart misinformation. Fifth, the Article will distinguish uses of deepfake manipulation that would require enforcement of the laws against perfidy or violations of honor from uses which would qualify as lawful ruse. Finally, the Article will conclude with recommendations on how to improve the governance of deepfake technology even as the technology continues to evolve and its deception capabilities become sharper.

I. MEDIA MANIPULATION AND WAR

A. Genesis and the First Fake

In 1838, Louis Daguerre captured the first photograph of a human[26] – accomplished almost by accident. Attempting to use his photography process to capture a picture of a Parisian street, he could not capture humans or any other mobile items such as horse carriages because his process required seven minutes of light exposure and seven corresponding minutes of no movement. Apparently unaware that the photograph was happening, nobody on the street had any reason to stand still that long. Nobody except, as luck would have it, a distant man standing at a corner having his shoes shined (the shoe-shiner would be captured as well).[27] This photograph, and other similar tin-plate "daguerreotypes" that followed, were revolutionary, heralded at the time

[26] Adam Withnall, *This is the First Ever Photograph of a Human – and how the Scene it was Taken in Looks Today*, INDEP., (Nov. 5, 2014, 4:45 PM), https://www.independent.co.uk/news/world/world-history/first-ever-photograph-human-and-how-scene-it-was-taken-looks-today-9841706.html; Robert Krulwich, *First Photo of a Human Being Ever?*, NAT'L PUB. RADIO, (Oct. 25, 2010, 10:17 AM), https://www.npr.org/sections/krulwich/2011/03/31/130754296/first-photo-of-a-human-being-ever (comparing the 1838 daguerreotype photograph with an 1848 photograph made in Cincinnati, Ohio).

[27] *See* Withnall, *supra* note 26.

for their "truthful likeness,"[28] and soon Mr. Daguerre would seek official recognition of his direct positive photographic printing process from the French Academy of Sciences.[29]

Mr. Daguerre, however, had a rival. Hippolyte Bayard was a fellow Frenchman who created his own photography process while Louis Daguerre was developing his.[30] Mr. Bayard hoped to beat Mr. Daguerre and achieve recognition from the French Academy of Sciences as the first claimant to the direct positive photographic printing process. When, however, Mr. Daguerre instead submitted his work in the first week of 1839 on what would become known as the daguerreotype process, he beat Mr. Bayard, torpedoing Mr. Bayard's ambitions and relegating him to the status of a follow-behind.[31]

Severely chafed and eager to continue to prove himself, Mr. Bayard chose to pioneer a different kind of first – the first fake photograph. It was morbid. In his 1840 photograph entitled "Self Portrait as a Drowned Man,"[32] Mr. Bayard spliced a self-portrait of his

---

[28] LIBR. OF CONGRESS, THE DAGUERREOTYPE MEDIUM, https://www.loc.gov/collections/daguerreotypes/articles-and-essays/the-daguerreotype-medium/ (last visited Oct. 20, 2020).

[29] *See* LOUISE JACQUES MANDÉ DAGUERRE, HISTORY AND PRACTICE OF PHOTOGENIC DRAWING ON THE TRUE PRINCIPLES OF THE DAGUERRÉOTYPE, WITH THE NEW METHOD OF DIORAMIC PAINTING 1-6 (J.S. Memes, LL.D. trans., Smith, Elder and Co. ed. 1839) (also available online at https://archive.org/details/historyandpract00memegoog/page/n8/mode/2up (Jul. 15, 2008 at 10:12 AM)) (detailing the submission made to the French Academy of Sciences as well as both the acceptance of the submission by the French government and the purchase of the process from Mr. Daguerre); *see also* Randy Alfred, *Aug. 19, 1839: Photography Goes Open Source*, WIRED, (Aug. 19, 2010, 7:00 AM) (discussing Louis Daguerre's advancement of direct positive photography and his efforts to have the process officially acknowledged and shared, resulting in the publication of his work in Aug. of 1839).

[30] Michal Sapir, *The Impossible Photograph: Hippolyte Bayard's "Self-Portrait as a Drowned Man"*, 40 MOD. FICTION STUD., no. 3, 1994, at 619-29. It should also be noted that William Henry Fox Talbot was also simultaneously working in England to develop his own photographic process and that Mr. Talbot, though not within the same professional circles as Mr. Daguerre and Mr. Bayard, was also a contemporary competitor of Mr. Bayard at the French Academy of Sciences that year. However, Mr. Bayard's follow-on actions appear to have been most influenced by his disappointment in his competition against Mr. Daguerre. *Id*; *see also* THE GETTY MUSEUM, HIPPOLYTE BAYARD, http://www.getty.edu/art/collection/artists/1840/hippolyte-bayard-french-1801-1887/ (last visited Oct. 20, 2020).

[31] *Id*.

[32] *Id*. *See also* Sean O'Hagan, *Exposed: Photography's Fabulous Fakes*, THE GUARDIAN (Jan. 31, 2016, 1:00 PM), https://www.theguardian.com/artanddesign/2016/jan/31/exposed-photographys-fabulous-fakes (comparing the Bayard fake suicide photograph to later examples of faked photographic images); Michael Zang, *The First Hoax Photograph Ever Shot*,

face, eyes closed and cheeks lifeless, on to a different self-portrait of his pale upper torso and darkened hands, appearing to show that he had committed suicide.[33] On the back of the picture was even a purported suicide note in which Mr. Bayard wrote "the poor wretch has drowned himself," that "he has been at the morgue for several days, and no-one has recognized him or claimed him," and warning the viewer that "you'd better pass along for fear of offending your sense of smell . . . the face and hands of the gentleman are beginning to decay."[34]

While Mr. Bayard, who had not committed suicide, made the photograph as an expression of protest and not as an attempt to fake his own death,[35] his work has served as a predecessor for media manipulation. From nineteenth century presidential touch-ups and face-swaps,[36] to twentieth century fairies,[37] to historical re-writes,[38] to twenty-

---

PETAPIXEL (Nov. 15, 2012), https://petapixel.com/2012/11/15/the-first-hoax-photograph-ever-shot/.

[33] *See* Sapir, *supra* note 30.

[34] Quotes translated from the original French. *Id.*

[35] Mr. Bayard would actually go on to experience significant professional success and renown in the field of photographic technology, earning several accolades during his lifetime including in 1863 the *Légion d'honneur* – the highest award that can be bestowed in France. However, his fake suicide photograph has dominated his legacy. *See* Getty Museum, *supra* note 30.

[36] *See e.g.* Michael Waters, *The Great Lengths Taken to Make Abraham Lincoln Look Good in Portraits*, ATLAS OBSCURA (Jul. 12, 2017), https://www.atlasobscura.com/articles/abraham-lincoln-photos-edited (discussing efforts to make President Lincoln appear more virulent to the public during his 1860 presidential campaign by splicing a picture of his face on to the more commanding posture of John C. Calhoun).

[37] The "Cottingley Fairies" was a series of photographs taken in 1917 depicting two young girls, Frances Griffiths and Elsie Wright, playing with winged fairies. The girls made the photographs after the younger girl, Frances (then nine years old), had claimed that she actually had played with fairies in her garden but was not believed. The method of the trick was simple – the girls made hand-drawn cutouts of fairies, stuck them in the ground with hatpins, posed with them, and took the pictures. While it's questionable whether they intended for the photographs to be seen as real, their photographs eventually circulated widely among local societies and in the local news. They even grabbed the attention of famed author Sir Arthur Conan Doyle who wrote a book in defense of the photographs' authenticity. Unfortunately for the reputation of all involved, however, Elsie would confess shortly before her death in the 1980s that the photographs were fake. Hazel Gaynor, *Inside the Elaborate Hoax that made British Society Believe in Fairies*, TIME (Aug. 1, 2017, 9:15 AM), https://time.com/4876824/cottingley-fairies-book/; *see also* SIR ARTHUR CONAN DOYLE, COMING OF THE FAIRIES 13, 196 (1922).

[38] Fourandsix Technologies hosts a webpage entitled "Photo Tampering Throughout History" which provides an in-depth image-based historical profile of famous fake or doctored photographs. Several images reside there of political leaders, such as Joseph Stalin and Mao Tse-Tung, 'erasing' or removing disfavored people posing with the political leader from photographs after the individual fell out of favor with the leader. PHOTO TAMPERING THROUGHOUT HISTORY, http://pth.izitru.com/ (last visited Oct. 21, 2020).

first century Instagram,[39] deceptions in visual media have exploded from the product of a gifted few to an output today so large that by some estimates at least half—if not more—of internet content is artificially created by one means or another.[40] Furthermore, editing and doctoring have evolved from being visually distinct to virtually indistinguishable absent a dedicated professional forensic investigation or a happenchance sloppy edit.[41]

B. Photo Fraud Goes to War

Image and audio manipulation have been a part of war ever since photographers first lugged their equipment to the ravaged battlefields of the Crimean War in 1854. British photographer Roger Fenton, widely acknowledged to be the first war photographer for his work during that war, has been accused of staging his photograph "The Valley of the Shadow of Death," taken after the 1854 Battle of Balaclava, by pre-positioning cannonballs to make the shot more dramatic.[42] Scrutiny has also come down upon famed American Civil War photographers Alexander Gardner and Matthew Brady who purportedly captured the human wreckage at Antietam and Gettysburg but who also allegedly

---

[39] Today, Instagram, a photograph sharing platform, is ubiquitous with modern-day photograph fakes and forgeries where an entire cottage industry has bloomed of self-styled influencers earning income in many cases by having photographs of themselves either altered or invented entirely in order to earn followers. *See e.g.* Janine Puhak, *Instagram Influencer Slammed for 'Fake Traveling' Photos*, FOX NEWS (Dec. 19, 2018), https://www.foxnews.com/travel/instagram-star-slammed-for-fake-traveling-photos.

[40] *See* Max Read, *How Much of the Internet is Fake? Turns out, a Lot of It, Actually*, N.Y. MAG. (Dec. 26, 2018), https://nymag.com/intelligencer/2018/12/how-much-of-the-internet-is-fake.html.

[41] In their sweeping examination of deepfake technology implications, Professors Danielle Citron and Robert Chesney explain how digital forensic efforts to detect fake images have become more and more difficult, noting that the "field of digital forensics has been grappling with the challenge of detecting digital alterations for some time." Danielle K. Citron & Robert Chesney, *Deepfakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1759 (2019). This increasing difficulty has long been forecast. *See e.g.* Hany Farid, *Digital Forensics: How Experts Uncover Doctored Images*, SCI. AM. (Jun. 1, 2008), https://www.scientificamerican.com/article/digital-image-forensics/ (observing in 2008 that "today anyone with a computer can readily produce fakes that can be very hard to detect").

[42] *See* MUSÉE D'ORSAY, ROGER FENTON: THE VALLEY OF THE SHADOW OF DEATH, https://www.musee-orsay.fr/en/collections/works-in-focus/photography/commentaire_id/the-valley-of-the-shadow-of-death-16457.html?tx_commentaire_pi1%5BpidLi%5D=847&tx_commentaire_pi1%5Bfrom%5D=844&cHash=1613936201 (last visited Oct. 21, 2020) (discussing the nature of the allegation but dismissing it outright).

moved and propped up bodies in an effort to make the destruction of the war appear more gruesome, or their photographs appear more contemporaneous to the fight.[43]

Today's battlefields have been no exception. Aside from the examples from Ukraine and Gabon discussed earlier, China has been accused of creating false and incendiary content when one of its Twitter accounts posted a fabricated image of an Australian soldier slitting the throat of an Afghan child during the later years of Australia's fight in Afghanistan.[44] North Korea and Iran have also both in recent years distributed photographs purporting to demonstrate larger forces of landing craft[45] and missile launchers,[46] respectively, than they actually possessed. Consider also the 2014 case of the Associated Press having to sever ties with an esteemed combat photographer after editors discovered that the photographer had improperly altered images of an anti-Assad regime fighter in Syria.[47]

Now, thanks to the ever-increasing sophistication of artificial intelligence, technological capabilities to create fake content have experienced a bullet-speed rise in complexity and efficacy. As programmers and developers worldwide have competed voraciously to

---

[43] *See* Michael E. Ruane, *Alexander Gardner: The Mysteries of the Civil War's Photographic Giant*, WASH. POST (Dec. 23, 2011), https://www.washingtonpost.com/local/alexander-gardner-the-mysteries-of-the-civil-wars-photographic-giant/2011/12/12/gIQAptHhDP_story.html.

[44] Zhao Lijian (@zlj517), TWITTER (Nov. 29, 2020, 8:02 PM), https://twitter.com/zlj517/status/1333214766806888448. The tweet was sent by Mr. Zhao Lijian, deputy director of the Information Department of the Chines Ministry of Foreign Affairs. The tweet came on the heels of the Brereton Report conducted by the Australian government which detailed, among other things, apparent unlawful killings by its own troops in Afghanistan. The Australian government called the tweet "utterly outrageous" and demanded an apology which the Chinese government refused to provide, causing further strain in the countries' relationship. Kirsty Needham, *Australia Demands Apology from China After Fake Image Posted on Social Media*, REUTERS (Nov. 29, 2020, 9:59 PM), https://www.reuters.com/article/us-australia-china/australia-demands-apology-from-china-after-fake-image-posted-on-social-media-idUSKBN28A07Y.

[45] *See* Alan Taylor, *Is This North Korean Hovercraft-Landing Photo Faked?*, THE ATLANTIC (Mar. 26, 2013), https://www.theatlantic.com/photo/2013/03/is-this-north-korean-hovercraft-landing-photo-faked/100480/; Damien Mcelroy, *North Korea 'Photoshopped' Marine Landings Photograph*, THE TELEGRAPH (Mar. 27, 2013), https://www.telegraph.co.uk/news/worldnews/asia/northkorea/9956422/North-Korea-Photoshopped-marine-landings-photograph.html.

[46] *See* Adam Hadhazy, *Is that Iranian Missile Photo a Fake?*, SCI. AM. (Jul. 10, 2008), https://www.scientificamerican.com/article/is-that-iranian-missile/; David Folkenflik, *On the Smokey Trail of a Faked Missile Photo*, NAT'L PUB. RADIO (Jul. 11, 2008, 1:07 PM), https://www.npr.org/templates/story/story.php?storyId=92454193.

[47] *See* Associated Press, *AP Severs Ties with Photographer who Altered Work*, ASSOCIATED PRESS ONLINE (Jan. 22, 2014), https://www.ap.org/ap-in-the-news/2014/ap-severs-ties-with-photographer-who-altered-work.

improve artificial intelligence, they have made simultaneous advances in how artificial intelligence learns and performs. These advances, as explained below, have the battlefield poised for serious complexities.

II. THE TECHNOLOGY BEHIND DEEPFAKE

A. Defining the Device

To understand deepfake technology and thereby understand its legal ramifications, it is important to first understand what deepfake technology is not. Deepfake media are not works of total human invention. Unlike the copy-pasting of a missile battery, deepfake media does not necessitate human decisions at all stages.

What distinguishes deepfake media from other variants of falsified images, and what makes their nature so convincing, is that they are self-correcting. Deepfake technology is mathematically engineered from and through artificial intelligence. In particular, deepfake technology is a consequence of machine learning. Machine learning, defined generally as the ability of a computer to solve a problem without being explicitly programmed,[48] can take such primitive forms as a 1642 hand-dialed device that calculated taxes.[49] The earliest modern mathematical models for defining and developing machine learning explored the game of checkers to determine whether an IBM computer could learn from and defeat a human opponent. It did.[50] The next natural step was to see if an IBM computer could learn from and defeat a human opponent at chess. It did.[51]

---

[48] *See* JOHN R. KOZA ET AL., *Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming*, *in* ARTIFICIAL INTELLIGENCE IN DESIGN '96 151, 153 (John S. Gero & Fay Sudweeks eds., 1996) (paraphrasing the work of Arthur Lee Samuel, the inventor of modern machine learning applications); *see also* Arthur L. Samuel, *Some Studies in Machine Learning Using the Game of Checkers*, 3 IBM J. 211-29 (1959).

[49] Pascal's Arithmetic Machine, also known as the Pascaline, was an early calculator invented by French mathematician Blaise Pascal in 1642. Designed to help tax collectors like the inventor's father, it required Mr. Pascal to implement several mathematical equations into the Pascaline's design so that the device could produce accurate, arithmetically-derived tax figures with the simple turning of a few dials. *See* Paul A. Freiberger & Michael R. Swaine, *Pascaline*, ENCYCLOPAEDIA BRITANNICA (Apr. 26, 2019), https://www.britannica.com/technology/Pascaline.

[50] Samuel, *supra* note 42; *see also* Bernard Marr, *A Short History of Machine Learning – Every Manager Should Read*, FORBES (Feb. 19, 2016, 2:31 AM), https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#468739a915e7.

[51] Marr, *supra* note 50.

Today, machine learning has grown into several sub-disciplines, each guided in large part by algorithm design and designer intent. For example, logistic regression has helped as early as 1990 to recommend cesarean deliveries based on patient data provided by physicians.[52] Another algorithm, known as Naive Bayes, can help sort desirable emails from spam emails.[53] Algorithm-based programs such as these, however, rely on "representations," that is to say, collections of information provided by human input (whether a computer programmer or a user checking their email inbox)[54] which communicates within the algorithms what right looks like.[55] In other words, machine learning in these contexts continues to require human hand-holding.

While such a fact is not inherently problematic, it has, in some sense, posed a barrier to more advanced machine learning. From this conundrum came deep learning. The concept is cogently explained by researchers at the Massachusetts Institute of Technology who pioneered certain advances in machine learning, explaining:

> "The hierarchy of concepts enables the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers. For this reason we call this approach to AI deep learning."[56]

An artificial intelligence designed to build and perfect images based on an algorithmic infrastructure that through multiple efforts generates *its own* representations (as opposed to constantly requiring human inputs) demonstrates deep learning. Amazon's Alexa AI, for example, employs deep learning through Google's proprietary Natural Language Processing program that enables Alexa to swiftly scan virtually all recorded words in the English language in order to improve how it receives and responds to a person's command.[57] This way, if Alexa AI

---

[52] IAN GOODFELLOW ET AL., DEEP LEARNING 3 (2016).
[53] *Id.*
[54] Also known as a "feature." *Id.*
[55] *Id.* at 4.
[56] *Id.* at 2.
[57] *See* Alexandre Gonfalonieri, *How Amazon Alexa Works? Your Guide to Natural Language Processing (AI)*, TOWARDS DATA SCI. (Nov. 21, 2018), https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3; Brian Barrett, *The Year Alexa Grew Up*, Wired (Dec. 19, 2018, 10:00 AM), https://www.wired.com/story/amazon-alexa-2018-machine-learning/ (detailing how Alexa's NLP enables it to find a radio station when a person requests it by a station nickname).

issues an incorrect return the first time, it could issue a correct return on a second or third attempt without any human command to make a different attempt.[58] Thus, when deep learning began to enable artificial intelligence to fabricate media, the term "deepfake" grew from a recognition of the role of deep learning in the creation of otherwise unreal or nontruthful media.[59]

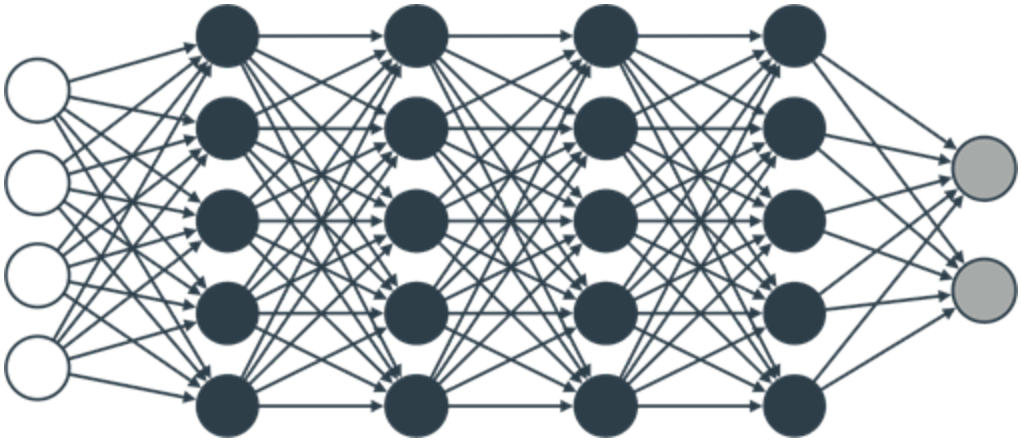
That the algorithmic function generates without human input, much less corrects without human input, is what fundamentally distinguishes deepfake from other methods of fabrication. How this occurs lies in the most basic component of information-gathering–the node –and the most basic component of computer activity.

B. Building Blocks

Merriam-Webster defines a "node" *inter alia* as a point at which other parts originate or center.[60] In the field of computer science, a node is, at its essence, a point of information.[61] A node can be either a device, such as a phone or computer, or a point of information input, such as a year, hair color, or height. A network occurs when two or more nodes become connected.[62] Thus, for example, a computer connected to the internet forms at least one network with the computer being one node and the internet[63] another. Additional computer connections then branch from this original network. Computer scientists sometimes represent clusters of nodes in what are called "trees" due to the fact that nodes will subordinate from a primary node (also called a "parent node") in a fashion that graphically represents a tree.[64] As they grow in complexity and function, producing even rudimentary thought patterns, these tree networks can be described as "neural networks," a nod to the similarly complex and hyper-connected nature of the human brain.[65]

---

[58] *Id.*
[59] *See* Riana Pfefferkorn, *"Deepfakes" in the Courtroom*, 29 B.U. PUB. INT. L. J. 245, 246 (2020) (describing the term "deepfake" as a "portmanteau of 'deep learning' and 'fake'.").
[60] *Node*, MERRIAM-WEBSTER ONLINE DICTIONARY, https://www.merriam-webster.com/dictionary/node (last visited Oct. 22, 2020).
[61] *See* BRIAN HARVEY & MATTHEW WRIGHT, SIMPLY SCHEME: INTRODUCING COMPUTER SCIENCE 299 (2nd ed. 1999); *see also* COMPUTER BUSINESS REVIEW, WHAT IS A NODE?, https://techmonitor.ai/what-is/what-is-a-node-4927877 (last visited Oct. 22, 2020).
[62] *Id.* (*see also* Harvey, *supra* note 61 at 306-07).
[63] Or, more accurately, servers hosting internet content.
[64] *See* Harvey, *supra* note 61 at 297.
[65] *See* Citron, *supra* note 41(citing Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 (explaining that the term "neural network" was first coined as far back



[*Fig. 2*. A demonstrative representation from Dr. Luis Serrano of a multi-layered neural network. For example, this particular network features five horizontal layers, left-to-right, not counting the first column of "input" nodes. Note that the four dark columns are "hidden," meaning that a person interacting with this network would see the input (for example, a Google search request for a local restaurant) and the output (a website link to a local restaurant) but would not see the various interconnected networks operating to filter out incorrect returns and find a correct return.][66]

Neural networks are the central infrastructure of artificial intelligence, serving as highways and byways along which machine learning, more complex representation learning, and eventually deep learning, occurs. While heavy research focus on neural networks waned during the first decade of the twenty-first century,[67] intensity of interest renewed with the advent of better computer processing abilities.[68] Then in the second decade, leaps in artificial neural network interaction theory

as 1944 by researchers at MIT)). *See also* GOODFELLOW, *supra* note 52, at 13 (observing that the early efforts to develop neural networks termed these networks "artificial neural networks" directly due to researchers' intent on using said networks to better understand the function of the human brain).

[66] LUIS SERRANO, GROKKING MACHINE LEARNING Ch. 10, fig. 10.1 (2020), https://livebook.manning.com/book/grokking-machine-learning/chapter-10/v-13/1. *See also* Jason Brownlee, *How to Configure the Number of Layers and Nodes in a Neural Network*, MACHINE LEARNING MASTERY (Jul. 27, 2018), https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network/.

[67] *See* Hardesty, *supra* note 65(describing how interest in neural networks rose and fell repeatedly during the 20th and 21st centuries).

[68] *Id.* (noting that advances in video game performance particularly fueled improvements in computer processing abilities which set the conditions for a neural network resurgence).

set the conditions for the deepfake technology that exists and evolves today.[69]

C. Two Tales of Two Networks

Leaps in artificial neural network interaction theory occurred in the evolution of variational autoencoders (VAEs) and most notably the pioneering development of generative adversarial networks (GANs).[70] Both disciplines use the relationship between two or often more networks to help train the networks to create a desirable output product, but in notably different ways.

1. Variational Autoencoders

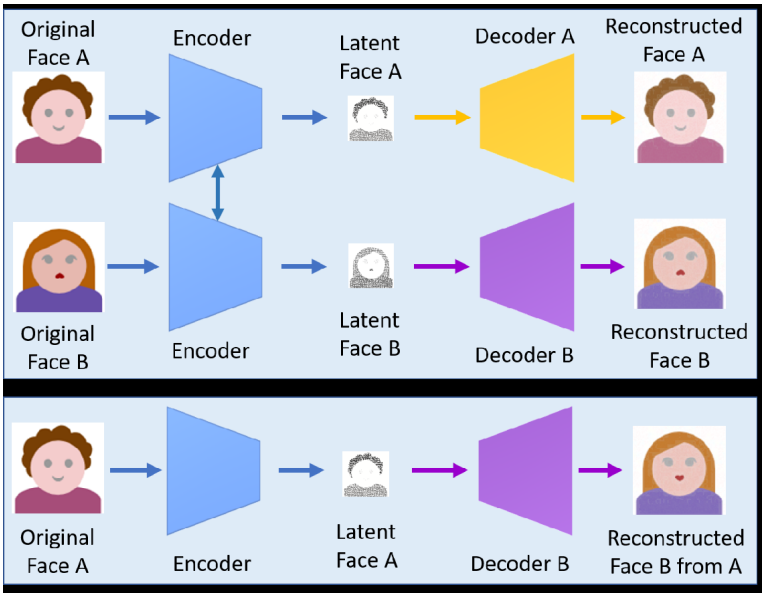
As alluded to in the introduction, the first widely known deepfake synthetic media creation was by a Reddit user who employed autoencoders to conduct a simple face swap to create pornographic content of female celebrities.[71] Today, given that most deepfake content relies on simple changes, such as face swaps, face editing, or face synthesis,[72] developers still often make deepfake content with autoencoders.

Autoencoders focus on two network players, an encoder and a decoder, which interact through an intermediary layer sometimes described as a "bottleneck" layer.[73] The encoder network receives an input, for example in the form of a picture of a person with dark hair (the source image).[74] The encoder identifies, categorizes, and condenses variables about that source image, such as jaw structure, hair color, lighting, etc. into the bottleneck. The decoder then extracts those variables from the bottleneck and recreates the source image. Once the autoencoder has accomplished this initial feat, the encoder then receives

---

[69] *Id.*; see also Michael Woolridge, A Brief History of Artificial Intelligence 139 (2020)(observing "In the second decade of the twenty-first century, AI has attracted more interest than any new technology since the World Wide Web in the 1990s.").
[70] Ian J. Goodfellow et al., Generative Adversarial Nets (Jun. 10, 2014, 6:58 UTC) (Neural Information Processing Systems conference paper), https://arxiv.org/abs/1406.2661; *see also* Citron, *supra* note 41, at 1760.
[71] *See* discussion *supra* note 17.
[72] *See* Mirsky, *supra* note 17 at 3; *see also* Andreas Rössler et al., FaceForensics++: Learning to Detect Manipulated Facial Images 1, 4 (Aug. 26, 2019, 17:59 UTC), https://arxiv.org/pdf/1901.08971.pdf.
[73] *See* Rössler, *supra* note 72at 14; *see also* Ben Dickson, *What are Deepfakes?*, TECHTALKS (Sep. 4, 2020), https://bdtechtalks.com/2020/09/04/what-is-deepfake/.
[74] *See* Nguyen, *supra* note 17 at 2; Rössler, *supra* note 72 at 14; Dickson, *supra* note 73.

a second input, for example a person with light hair (the target who, in the case of a face swap, the developer wants depicted in place of the face from the source image). The encoder, with some degree of guidance from the developer, distills variables about the target image into the bottleneck layer where the source image variables still reside, and the compression of data mitigates margins of error. The decoder extracts variables from both images and then attempts to construct the goal synthetic image as exemplified here:



[*Fig. 3.* Graphical representation of synthetic media creation via an encoder-decoder pair. The goal fake image is at the bottom right.][75]

Autoencoders predate deepfake technology so this advent is not new. What accelerated these neural networks towards deepfake-level capacity were *variational* autoencoders (VAE).[76] Prior autoencoders required users to comb laboriously through sometimes thousands of images in order to find useful variables for decoder use.[77] VAEs, on the other hand, use probabilistic generative modeling, meaning the decoder tries to predict from the information available in the bottleneck layer what the goal hybrid image should be.[78] The result has been described

[75] The image is from Nguyen, *supra* note 17 at 3.
[76] *See* Lars Ruthotto & Eldad Haber, An Introduction to Deep Generative Modeling 22 (Mar. 9, 2021, 02:19 UTC), https://arxiv.org/pdf/2103.05180.pdf.
[77] *See* Dickson, *supra* note 73 (describing the process of selecting images from a video and cropping each one to just portray a face).
[78] Diedrick P. Kingma & Max Welling, An Introduction to Variational Autoencoders 28-30 (Dec. 11, 2019, 17:33 UTC), https://arxiv.org/pdf/1906.02691.pdf (describing how VAE training can develop an "*importance sampling* technique" [emphasis original] to assist with VAE inferences).

as "elegant" and "simple to implement."[79] However, VAEs can still demand a significant amount of time and data,[80] and they suffer from distinct image blurriness[81] which has made other avenues more attractive.

2. Generative Adversarial Networks

The creation of GANs, by comparison, has been a game-changer in media manipulation . Employing the analogy of the counterfeiter and the cop [82] imagine a counterfeiter is trying to sneak a counterfeit picture past a cop who is diligently looking out for counterfeit pictures. Being a first attempt, the counterfeiter's first efforts are rudimentary. When the cop obtains the picture, the cop easily determines that the picture is a fake and discards it. The counterfeiter, however, learns that the cop has detected faults in the picture. The counterfeiter determines to avoid those faults, generates a new picture that does not include those faults, and tries again. The process continues, the counterfeiter removing one detected fault from the creation process after another, until the counterfeiter has removed so many faults that the cop can no longer detect the difference between an authentic picture and a fake picture.

Generative adversarial networks operate in the same way. A GAN consists essentially of a pair of neural networks that compete against each other.[83] One network, termed a "generator,"[84] will act as the counterfeiter, generating information that it has manufactured. The other network, termed a "discriminator,"[85] will act as the cop, filtering out information that does not match the parameters set for authenticity. A programmer will build the discriminator network first. In the process, the programmer will define the properties that characterize an authentic

---

[79] *See* Goodfellow, *supra* note 52 at 688.
[80] *See* Matthew Stewart, *GANs vs. Autoencoders: Comparison of Deep Generative Models*, TOWARDSDATASCIENCE (May 12, 2019), https://towardsdatascience.com/gans-vs-autoencoders-comparison-of-deep-generative-models-985cf15936ea. However, databases have proliferated online to facilitate such data collection. CelebFaces Attributes Dataset, for example, contains over 200,000 face images of over 10,000 public figures. *Id.*
[81] *See id.*; Kingma, *supra* note 78 at 32.
[82] This analogy is most commonly associated with Mr. Ian Goodfellow, an often-credited trailblazer of GAN development who also uses the analogy often to illustrate the concept. *See* Ian Goodfellow, *Introduction to GANs, NIPS 2016 | Ian Goodfellow, OpenAI* (Aug. 24, 2017). https://www.youtube.com/watch?v=9JpdAg6uMXs.

[83] *See* Goodfellow, *supra* note 70 at 1.
[84] *Id.*
[85] *Id.*

output—often by numerical valuative factors but sometimes simply by uploading authentic video images of a target individual or typing desired spoken text into a prompt in order to shape the desired synthetic output, the goal of the GAN. By defining conditions for success, the programmer has implicitly also begun defining conditions for failure as data that does not match the goal will eventually become waste, or "noise."[86] The programmer will then start building the generator. Through algorithmic inputs, some of which may be purposefully hidden or "latent,"[87] the programmer essentially sets the goalposts for the generator. The generator, once created, immediately begins to transmit data to the discriminator, and the adversarial back-and-forth starts.

Due to the nature of the exchange and the positions of the dueling networks, the discriminator will almost always lose.[88] In fact, arguably the best-case scenario for a discriminator is that the discriminator network will get to the point where it can accurately estimate that at least 50 percent of the data produced by the generator is noise.[89] The generator cannot produce such success without training. Sophisticated training, therefore, is the hinge-point for the effectiveness of deep learning and deepfake technology.

---

[86] *See generally* Serrano, *supra* note 66; *see also* Luis Serrano, *A Friendly Introduction to Generative Adversarial Networks (GANs)* (May 5, 2020), https://www.youtube.com/watch?v=8L11aMN5KY8.

[87] Diego Gomez Mosquera, *GANs from Scratch 1: A Deep Introduction. With Code in PyTorch and TensorFlow*, AI SOC. (Feb. 1, 2018), https://medium.com/ai-society/gans-from-scratch-1-a-deep-introduction-with-code-in-pytorch-and-tensorflow-cb03cdcdba0f.

[88] *See* Goodfellow, *supra* note 82; Serrano, *supra* note 86.

[89] *See e.g.* Jason Brownlee, *How to Identify and Diagnose GAN Failure Modes*, MACHINE LEARNING MASTERY (Jan. 21, 2021), https://machinelearningmastery.com/practical-guide-to-gan-failure-modes/. Google developers posit that this 50 percent assessment rate occurs at a tipping point which arrives when the generator becomes so good that the discriminator's success appears owed more to chance than calculation. *See GAN Training* in *Generative Adversarial Networks* (Jul. 12, 2019), https://developers.google.com/machine-learning/gan/training (last visited Jun. 25, 2022). If, however, the algorithm continues to generate images yet the discriminator begins to reflect an accuracy rate beyond 50 percent, rendering the accuracy rate artificial, this can indicate error in the discriminator which would unintentionally cause the generator to become less effective. *Id.*

D. Supervised, Unsupervised, and Semi-Supervised Training

1. Supervised Learning – Showing the Machine

Supervised learning is not only the original method of machine training but also the most common—so common actually that we all unwittingly participate in it every day. Supervised learning occurs when algorithms receive labeled or pre-defined information with the intention that the algorithm will use that information to achieve a preconceived target output. IBM describes it as the "use of labeled datasets to train algorithms that to [sic] classify data or predict outcomes accurately."[90] Put more directly, supervised learning involves actions by "an instructor or teacher who shows the machine learning system what to do."[91]

Anyone, however, can be an instructor or teacher for AI. We participate in supervised learning-style AI training whenever we ask an Amazon Alexa device to tell us the weather forecast, tap our brakes in vehicles with automated brake performance-enhancing technology[92], or ask Google Translate to convert a question from English to French.[93] Physicians can assist supervised learning by inputting patient data and treatment techniques into algorithmic-based programs to predict the likely journey of a COVID-19 infection and increase chances of successful recovery.[94] Data analysts use algorithms trained with various supervised learning techniques to improve face-recognition technology and predict stock market fluctuations.[95]

We all train artificial intelligence every day via supervised learning without really knowing it. However, pure supervised learning is really only useful for classification modeling (e.g., telling the difference

---

[90] IBM Cloud Education, *Supervised Learning*, IBM CLOUD LEARN HUB (Aug. 19, 2020), https://www.ibm.com/cloud/learn/supervised-learning#toc-what-is-su-d3nKa9tk.
[91] Goodfellow, *supra* note 70 at 103.
[92] *See* Alyssa Schroer, *Artificial Intelligence in Cars Powers an AI Revolution in the Auto Industry*, BUILTIN (Mar. 25, 2020), https://builtin.com/artificial-intelligence/artificial-intelligence-automotive-industry.
[93] Yonghui Wu et al., *Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation* (Oct. 8, 2016), https://arxiv.org/abs/1609.08144.
[94] *See* The Mount Sinai Hospital, *Developing Machine Learning Models to Predict Critical Illness and Mortality in COVID-19 Patients*, MEDICAL XPRESS (Nov. 10, 2020), https://medicalxpress.com/news/2020-11-machine-critical-illness-mortality-covid-.html.
[95] *See* JEREMY WATT ET AL., MACHINE LEARNING REFINED: FOUNDATIONS, ALGORITHMS, AND APPLICATIONS 1 (2016).

between a cat and a dog) or for regressive/predictive modeling (e.g., predicting the rate of student loan debt expansion over time).[96] Other learning techniques, therefore, become necessary to help sharpen AI.

2. Unsupervised Learning – Unbinding the Machine

Unsupervised learning deepens the AI talent pool and, ultimately, sets the conditions for deepfake technology to thrive. Whereas supervised learning occurs when an algorithm works within a set of labeled inputs, unsupervised learning removes the training wheels. In this case, a neural network will instead work with unlabeled inputs. Without the use of labeled inputs to communicate goal expectations, the network instead must identify patterns in order to deliver a goal output.[97]

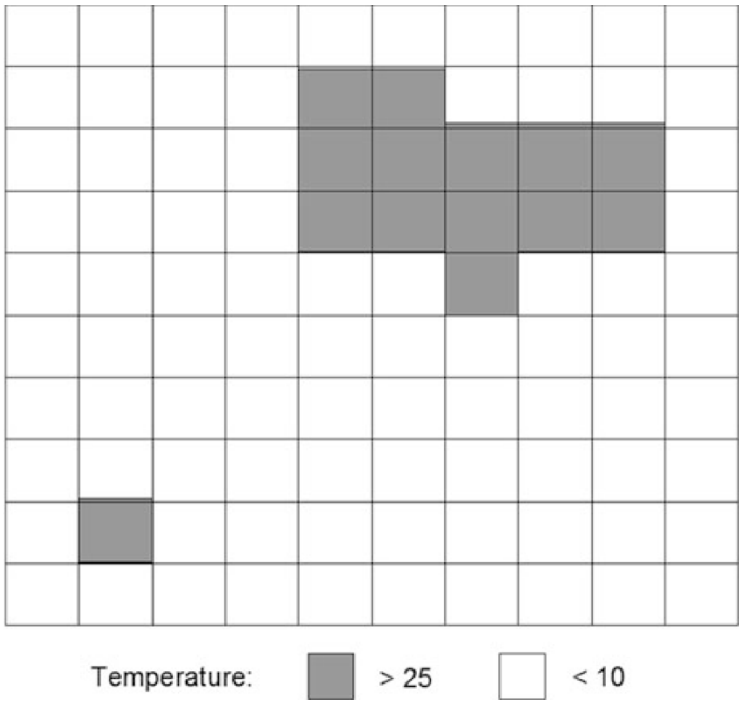
The learning that results is termed "unsupervised" because the human programmer minimizes their influence on the network so that the programmer can test the algorithm's independent ability to learn i.e., to adjust, discern, and identify, mathematically speaking.[98] The kinds of tasks that unsupervised learning tends to accomplish are generally those which group similar kinds of data or information, also known as "clustering."[99] In this way, an algorithm can identify one or several themes in a data group (e.g., pictures of men with dark hair v. men with gray hair v. men with no hair v. men with dark beards v. men with gray beards v. men with no beards) and compartmentalize each piece of data into groups, based on apparent patterns, to present a cluster of results (e.g., all pictures of men with beards) which a person can retrieve by requesting that particular cluster. On a larger scale, clustering assists with everything from data mining to data extraction to data analysis.

By logical and actual extension, unsupervised learning can also accomplish an implied task of clustering, that is to say, identify what data does *not* belong to a data cluster. Known as "anomaly detection"[100] or, in a related context, "denoising,"[101] this task identifies those data points

---

[96] *Id.* at 1-12.
[97] *See* Goodfellow, *supra* note 70 at 103.
[98] *See e.g.* UNSUPERVISED LEARNING ALGORITHMS V (M. Emre Celebi & Kemal Aydin eds., 2016) (observing that unsupervised learning algorithms "automatically discover interesting and useful patterns" in unlabeled data).
[99] *See* Goodfellow, *supra* note 70 at 103; *see also* Tülin İnkaya, Sinan Kayaligil, & Nur Evin Özdemirel, *Swarm Intelligence-Based Clustering Algorithms*, in UNSUPERVISED LEARNING ALGORITHMS 303 (M. Emre Celebi & Kemal Aydin eds., 2016).
[100] *See e.g.* P. Deepak, *Anomaly Detection for Data with Spatial Attributes*, in UNSUPERVISED LEARNING ALGORITHMS 1 (M. Emre Celebi & Kemal Aydin eds., 2016).
[101] Goodfellow, *supra* note 70 at 101, 507 (discussing how denoising autoencoders receive a "corrupted data point as input and [are] trained to predict the original, uncorrupted data point as [their] output.").

or characteristics which do not comport with the patterns already established during clustering. Whether identifying anomalies (i.e., groups of data which depart from what is generally regarded as common[102]) or outliers (i.e., an *individual* object which presents an uncommon characteristic[103]), the result is that the network is able to actively filter.



[*Fig. 4.* The above graphic, devised by Dr. Deepak Padmanabhan of Queen's University Belfast, depicts a hypothetical geographic region split into grids with each grid colored in accordance with its average temperature. As the largest pattern in this set is that most grid areas possess average temperatures, the cluster of dark squares to the top-right are an abnormality because they represent a sub-region that experiences higher-than-normal average temperatures. The dark square at the bottom left represents an outlier.[104] A network tasked with finding a place for someone to spend a weekend in a comfortable climate could, using unsupervised learning-devised anomaly detection, search in only the white grids in order to improve the chances of finding the most-desired vacation spot.]

This filter training, combined with immense computing power, is what makes the GANs discussed above work and by extension can make today's deepfake technology threat so potent. Because unsupervised learning principles help inject discrimination into machine learning, GANs receive the discriminator needed to enable the tasked program to

[102] P. Deepak, *supra* note 100 at 1-2.
[103] *Id.* at 2.
[104] *Id.*

constantly improve results. Researchers have seized on this strength by pairing GANs in unsupervised learning contexts with other neural networks such as VAEs to develop even more sophisticated content production,[105] resulting in more robust deepfake capabilities. However, sometimes requirements necessitate hybrid machine learning which is where semi-supervised learning gains purchase, and sometimes where deepfake content is best made.

3. Semi-Supervised Learning – Cooperating with the Machine

There is no chicken-egg, which-came-first conundrum about deepfake images. The person desiring to obtain a deepfake image comes first. This person supplies sometimes basic, sometimes sophisticated, parameters into a GAN in the hopes of getting a desired result. By doing so, the person has weighted and labeled at least some data sets. However, the GAN works to produce an image that is not only equivalent to the labels provided by the person but, for those features which do not carry an express label, is also consistent with patterns identified during the GAN's generate-and-reject volleys.

Deepfake images, therefore, can often be the product of semi-supervised machine learning. Consider the work built upon the pioneering GANs which have enabled today's deepfake technology. Within a year after Mr. Goodfellow published his work on GANs, advocates for a semi-supervised learning approach to GANs advanced the concept of a third network–known as a classifier–to deepen and improve the performance of the discriminator network, and thereby the generator network.[106]

A few years thereafter, researchers in China expounded upon the semi-supervised GAN approach with the development of the Margin

---

[105] *See e.g.,* Ming-Yu Liu et al., Unsupervised Image-to-Image Translation Networks 2 fig. 1 (Jul. 23, 2018, 3:39 AM), https://arxiv.org/pdf/1703.00848.pdf (proposing UNIT Networks as a combination of VAEs and GANs to leverage each network structure's strengths in order to better refine image generation accuracy and quality).

[106] *See e.g.,* Jost Tobias Springenberg, Unsupervised and Semi-Supervised Learning with Categorical Generative Adversarial Networks 2-4 (Nov. 19, 2015, 21:26 UTC) (presented at the 2016 International Conference on Learning Representations), https://arxiv.org/abs/1511.06390; Aug.us Odena, *Semi-Supervised Learning with Generative Adversarial Networks* 1 (Jun. 5, 2016, 11:42 PM), https://arxiv.org/pdf/1606.01583.pdf (citing Springenberg). This structure is more commonly known today as a "Triple-GAN." *See* Chongxuan Li et al., *Triple Generative Adversarial Nets* 2 (Dec. 20, 2019, 12:17 PM), https://arxiv.org/abs/1912.09784 (describing the growing utility of classifier or classifier-like networks in teaching GANs to produce more precise results).

Generative Adversarial Network (MarginGAN), a GAN which has a classifier network designed not only to help the discriminator sort data in order to identify fake images but also, by influence of "pseudo labels" provided to the generator, to increase margins of real images and decrease margins of fake images.[107] With additional proliferations of similar off-shoots such as CatGANs,[108] Triangle GANs,[109] and SGANs,[110] and the realization through semi-supervised learning that even greater network precision can occur by introducing further adversity between not only the generator and discriminator but also the generator and the classifier[111], semi-supervised learning has helped foster tremendous progress in synthetic content development. Combine these advances with developments in "reinforcement learning," described as a "crowning achievement of deep learning,"[112] in which the AI improves its output through a trial-and-error/reward-punishment system imposed by a programmer,[113] and it becomes easier to see how deepfake technology has arrived at its current sophisticated state.

III. IDENTIFYING VIOLATIONS AND VIOLATORS: CLASSIFICATION, ATTRIBUTION, AND AGENCY.

In order to know how to enforce the laws on deception in combat, a State must understand what a violation of those laws looks like and how to identify perpetrators. The first challenge in combating deepfake content is knowing when content is in fact fake. After swiftly notifying partners about the fake content, the second challenge is identifying the responsible actors as quickly as possible. The third and potentially most

---

[107] Tong Lin & Jinhao Dong, *MarginGAN: Adversarial Training in Semi-Supervised Learning*, *in* ADVANCES IN NEURAL INFO. PROCESSING SYS. (H. Wallach et al. eds., 32nd ed., 2019), https://papers.nips.cc/paper/2019.

[108] *Id.* at 2 (citing Jost Tobias Springenberg, Address at International Conference on Learning Representations: Unsupervised and Semi-Supervised Learning with Categorical Generative Adversarial Networks (Nov. 19, 2015)).

[109] *Id.* (citing Zhe Gan et al., Triangle Generative Adversarial Networks, (2017 Neural Information Processing Systems conference paper, 2017), https://arxiv.org/abs/1709.06548).

[110] *Id.* (citing Zhijie Deng et al., Structured Generative Adversarial Networks, (2017 Neural Information Processing Systems conference paper, 2017), https://arxiv.org/abs/1711.00889).

[111] *See* Wenyuan Li et al., *Semi-Supervised Learning Using Adversarial Training with Good and Bad Samples*, 31 MACH. VISION AND APPLICATIONS 49 (2020) (also available at https://doi.org/10.1007/s00138-020-01096-z).

[112] GOODFELLOW, *supra* note 70, at 25, 103.

[113] *Id.*; *see also* Surbhi Arora, *Supervised vs Unsupervised vs Reinforcement*, AITUDE (Jan. 29, 2020), https://www.aitude.com/supervised-vs-unsupervised-vs-reinforcement/.

sensitive challenge is determining whether the actions of any actors can be attributed to a State or an organization. Deepfake technology poses unique difficulties in all three efforts.

A. Chasing and Catching Mirages

The persistent problem among those who wish to regulate digital and cyber activities is that human ingenuity often has the appearance of staying one step ahead. The same is true for those currently hoping to find ways to quickly identify deepfake content. As Dr. Alexa Koenig has observed, detecting deepfakes presents several challenges including the "increasing sophistication and decreasing costs of deep learning technologies," an "information ecosystem" degraded by a continuous influx of misinformation, and a lack of legal professionals trained to verify fakes—a skill Dr. Koenig describes as "a first line of defense against being duped."[114]

Although these challenges exist, several projects are nonetheless underway to combat AI-enhanced deception—and some of these projects employ just as much ingenuity as their adversaries. The most common intuition is to design automated deepfake detection systems—i.e., combat AI with AI—in order to maximize detection timing, sophistication, and capacity while reducing the potential for human error.[115] To this end, hosts of computer scientists and engineers have researched various methods that can algorithmically detect deepfake-enabled content.[116] Diverse research has competed to develop machine learning algorithms that detect deepfakes by the various subtle errors that today's technology still exhibits, such as co-motion patterns,[117] the

---

[114] Alexa Koenig, *"Half the Truth is Often a Great Lie": Deepfakes, Open Source Information, and International Criminal Law*, 113 AJIL UNBOUND 250, 252 (2019). Dr. Koenig is the Executive Director of the Human Rights Center at the University of California Berkley School of Law.

[115] *See* Alex Engler, *Fighting Deepfakes When Detection Fails*, BROOKINGS INSTITUTE (Nov. 14, 2019), https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/.

[116] *See e.g.*, *id.* (citing, *inter alia*, Yuezun Li & Siwei Lyi, Exposing DeepFake Videos by Detecting Face Warping Artifacts, (Nov. 2018), https://arxiv.org/abs/1811.00656; David Guera & Edward J. Delp, DeepFake Video Detection Using Recurrent Neural Networks (Nov. 2018), https://engineering.purdue.edu/~dgueraco/content/deepfake.pdf).

[117] Gengxing Wang, Jiahuan Zhou, & Ying Wu, Exposing Deep-Fake Videos by Anomalous Co-Motion Pattern Detection (Aug. 11, 2020), https://arxiv.org/abs/2008.04848. "Co-Motion Patterns," as described by the authors here, occur when a person's face in a deep-fake video exhibits slight movements (a.k.a. landmarks) or lacks slight movements in a way that is atypical in genuine facial movements. These are more identifiable in deep-fake videos that have

lack of miniscule changes of skin color in a face that an actual normal heartbeat would presumably cause,[118] and atypical eye blinking.[119] Observers also acknowledge the early efforts of Gfycat to combat deepfake pornography through its Project Angora and Project Maru initiatives which scour the internet and find images of the depicted individual in order to compare facial features and make an analytical assessment about whether the concerned content is synthetic.[120] Recently, Facebook has also invested in deepfake detection technology through its 2020 Deepfake Detection Challenge (DFDC) which incentivized over 2,000 competitors to devise a program which would have the highest detection rate among a selection of video images.[121]

The United States government has also been a vigorous player in the effort to develop deepfake-combating AI. Through its Guaranteeing AI Robustness against Deception (GARD) Program, the Defense Advanced Research Projects Agency (DARPA) has a robust portfolio of approaches for identifying deepfake and other similarly faked content with the specific aim of creating "deception-resistant [machine learning] technologies"[122] which can competently defeat both current levels of deepfake technology and expected future evolutions.[123] Finding biological inspiration in the immune system, GARD looks to develop a defense system that "identifies attacks, wins and remembers the attack to create a more effective response during future engagements."[124]

---

both real and deep-fake content (for example, where a genuine video of a President giving a real speech is altered to make the President say only a few things that he or she did not actually say). A similar focus influenced some of the first work on counter-deepfake AI. *See e.g.* Darius Afchar, et al., MesoNet: A Compact Facial Video Forgery Detection Network (Sep. 4, 2018), https://arxiv.org/abs/1809.00888.

[118] Hua Qi et al., DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms (Aug. 26, 2020), https://arxiv.org/pdf/2006.07634.pdf.

[119] Yuezun Li et al., In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking (Jun. 11, 2018), https://arxiv.org/abs/1806.02877.

[120] *See* Louise Matsakis, *Artificial Intelligence is Now Fighting Fake Porn*, WIRED (Feb. 14, 2018), https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/; Koenig, *supra* note 114, at 254; Citron, *supra* note 41, at 1787n.145.

[121] *Deepfake Detection Challenge Results: An Open Initiative to Advance AI*, FACEBOOK AI, https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/ (last visited Feb. 25, 2021).

[122] *Defending Against Adversarial Artificial Intelligence*, DEF. ADVANCED RSCH. PROJECTS AGENCY (Feb. 6, 2019), https://www.darpa.mil/news-events/2019-02-06.

[123] Anticipated future evolutions in deep-fake technology include "multi-sensor and multi-modality variations" as well as generative AI capable of making predictions, decisions, and adaptations. *Id.*; *see also* Bruce Draper, *Guaranteeing AI Robustness Against Deception (GARD)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception (last visited Feb. 24, 2021).

[124] *Defending Against Adversarial Artificial Intelligence*, *supra* note 122.

Many of the programs intended to bring GARD's immune system-inspired deception-defeat capability to life show high creative and practical potential, and equally high ambition. The Reverse Engineering of Deceptions (RED) program seeks to employ AI capable of reverse engineering media content's algorithmic toolchains (i.e. the sequential series of steps in a machine's operation from start to finish) not only to determine if content is fake but also to determine the content's point of origin–enabling the U.S. to actually identify the adversary sending the deepfake.[125] The Media Forensics (MediFor) program builds on work already done in the fields of digital and other media forensics by developing an "end-to-end" platform which can employ techniques relevant across the media spectrum to detect expected manipulations, explain how the programmers made the manipulations, and quantify the likelihood that target content is actually fake.[126] Finally, the Semantic Forensics (SemaFor) program would train AI to latch on to semantic errors such as problems with facial structure, coloration, or eye-blinking discussed above to develop a catalogue of errors which would impose a burden on creators to "get every semantic detail correct, while defenders only need to find one, or a very few, inconsistencies."[127] Additionally, the SemaFor program would also train AI, like the MediFor program, to determine not only that content is fake but also where the content originated in order to aid in attribution.[128]

While the combined results of these efforts, both within DARPA and within the larger computer sciences communities, show tremendous progress in combating deepfake, many of these approaches still have inherent weaknesses. Gfycat's Projects Maru and Angora, for instance, would appear useless when faced with videos that do not have any source content from the internet. The DeepRhythm methodology, which would look for semantic errors if an image's facial coloration did not correlate to a normal heartbeat,[129] does not appear immediately able to account for

---

[125] Matthew Turek, *Reverse Engineering of Deceptions (RED)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/reverse-engineering-of-deceptions (last visited Feb. 24, 2021) [hereinafter *RED]*.
[126] Matthew Turek, *Media Forensics (MediFor)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/media-forensics (last visited Feb. 24, 2021) [hereinafter *MediFor*].
[127] Matthew Turek, *Semantic Forensics (SemaFor)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/semantic-forensics (last visited Feb. 24, 2021) [hereinafter *SemaFor*].
[128] *Id.*; *see also Uncovering the Who, Why, and How Behind Manipulated Media,* DEF. ADVANCED RSCH. PROJECTS AGENCY, *https://www.darpa.mil/news-events/2019-09-03a* (last visited Jun. 25, 2022).
[129] Qi, *supra* note 118 at 1.

biological variables such as might present in a person with a heart condition or blood pressure issues. And the winner of the Facebook DFDC achieved an 82.56 percent accuracy[130]—certainly impressive but still allowing for an error rate that could permit significant harm in a national security or armed conflict scenario.

Potentially most problematic, even for all of the work and resources expended in deepfake detection efforts, is the "detection dilemma."[131] Simply put, this is the notion that the more work that goes in to detecting deepfake, the more deepfake creators learn how to avoid detection. As discussed, and cited to above, much of the research done into detection strategies is open source. For every publication that describes how a new set of algorithms can detect unnatural blinking patterns, deepfake developers learn to improve blinking. Even a mass-effort style approach by entities like DARPA can seem from afar like a Sisyphean task. A recent article on the subject by members of three highly-influential AI advancement enterprises called for an all-hands "multistakeholder" coalition effort among academia, media, technology, and civil society organizations in order to effectively counter the coalition of adversarial interests that can cause deepfake proliferation.[132] This kind of broad-based cooperability between government and non-governmental entities has also been proposed in seeking ways to confirm and counter GAN-enabled manipulation of satellite imagery.[133] A bulletproof, long-term solution may ultimately not be likely. The real best defense, and thereby best ability to detect deepfakes, may at least for now be the fact that we know they exist, that we continue to talk about them, and that major social players maintain dialogue to determine methods of cooperation as deepfake threats grow.

---

[130] *Deepfake Detection Challenge Results: An Open Initiative to Advance AI*, *supra* note 121.

[131] Claire Leibowicz et al., The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media (Feb. 11, 2021), https://arxiv.org/abs/2102.06109.

[132] *Id.* The three enterprises are The Partnership for AI, the XPRIZE Foundation, and the Thoughtful Technology Project. It is worth noting that this article did not list government explicitly as a "stakeholder" in this effort.

[133] *See* Patrick Tucker, *The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth*, DEFENSE ONE (Mar. 31, 2019), https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/.

B. Agency and Attribution: Technical Analysis

Once a deception has been revealed, and authorities have spread the word to assure that the deception or similar variants of it do not continue to succeed, the next task is to identify the source of the deceptive action. In the context of armed conflict, this determination must occur immediately in order to stop the bleeding, both figuratively and possibly even literally, as well as to determine follow-on responses.

Specifically, the targeted force must be able to identify persons and belligerents. Like with other forms of cyber warfare, this task can be difficult, as a cyberattack does not often leave a literal trail of smoke. Furthermore, the asset which deploys the attack does not have to even be in the same hemisphere as the target.

Many of the deception identification efforts discussed above seek not only to confirm that content is fake but also to begin to detect agency i.e., the confirmation that human actors are involved, their identities, and their level of responsibility. Once agency is established, attribution of the concerned people or entities to States or non-State actors can begin. DARPA's RED program, for example, acknowledges that "identifying an adversary" is one of many desired outcomes from its automated toolchain reverse engineering approach.[134] Their SemaFor program specifically seeks to employ "attribution algorithms" in order to help determine if the content originated from an individual or an organization.[135]

Significant academic research over the past three years has produced a steady stream of analyses helpful for finding actors and entities employing deepfake. One such study, financed in part by DARPA's MediFor program, has developed attribution algorithms which train on and identify "GAN fingerprints" in images in order to increase a classification network's ability to specifically identify GANs and conduct image and model attribution.[136] Their classifiers, even when tested against attribution defenses, often demonstrated accuracy rates well in excess of 90%.[137]

Currently, however, it is unclear whether any of these efforts are effective enough to solve the Gordian Knot that has become cyberspace attribution. First, many attribution methods still suffer from exploitable

---

[134] *RED*, *supra* note 125.
[135] *SemaFor*, *supra* note 127.
[136] Ning Yu et al., *Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints*, 7555, 7556 (Feb. 27, 2020), https://ieeexplore.ieee.org/document/9010964.
[137] *Id.* at 7562.

vulnerabilities. The GAN fingerprint-detecting attribution algorithm discussed above, for example, is only trained for scouring images of still faces.[138] It does not analyze videos, audio data, or any other data other than still human faces. Also, it can only "attribute" the image to being a GAN construct or not being a GAN construct—it is not yet sufficiently sophisticated to attribute a GAN to a server.[139]

Second, several means currently exist for mal-intended actors to hide their involvement. Tor is a popular anonymity platform which prevents IP address tracking.[140] However, other utilities such as Mixmaster, Onion Routing, and AN.ON further complicate the picture because they use anonymity networks via proxy servers to code, re-code, and re-order (scramble) data in order to make a content's route or source strenuously difficult to track.[141] While robust work and resources have been invested in developing anonymity network hacks and have seen some success,[142] to the delight of privacy advocates none so far have proven effective enough to lift the veil.

Third, none of the known automated attribution systems are anywhere near the level of sophistication necessary to identify fakes and find actors to the level needed to be truly effective real-time. If a skilled developer constructs a deepfake video appearing to show U.S. soldiers mocking the Quran and cursing the Prophet Muhammed and then manages to release it anonymously on the internet claiming it came from an area of active operations in a Muslim-majority region, forensics work

---

[138] *Id.*

[139] *Id.*; however, *cf.* Tianyun Yang et al., *Deepfake Network Architecture Attribution* 1 (Mar. 14, 2022), https://arxiv.org/abs/2202.13843 (providing an architecture-based approach to deepfake "fingerprint" detection as opposed to model-based detection).

[140] *See* Citron, *supra* note 41, at 1792. Tor pre-dates deepfake technology, having been used famously in 2009 by Iranians trying to protest the elections there and the subsequent crushing of popular unrest by then-President Mahmoud Ahmadinejad. *See* Cyrus Farivar, *Geeks Around the Globe Rally to Help Iranians Online*, FRONTLINE (Jul. 8, 2009, 3:56 pm), https://www.pbs.org/wgbh/pages/frontline/tehranbureau/2009/07/geeks-around-the-globe-rally-to-help-iranians-online.html.

[141] *See* Simone Fischer-Hbner & Stefan Berthold, *Privacy-Enhancing Technologies*, in COMPUTER AND INFORMATION SECURITY HANDBOOK 759-78 (John R. Vacca ed., 3d ed. 2017) (discussing the mix net concept).

[142] *See e.g.*, Zhongxiang Wei et al., Fundamentals of Physical Layer Anonymous Communications: Sender Detection and Anonymous Precoding (Oct. 18, 2020), https://arxiv.org/abs/2010.09122 (discussing the ability of signaling patterns and channel characteristics to provide inferences which can help identify a sender, while also acknowledging that precoding can help defeat sender identification efforts); Wenlin Han & Yang Xiao, *Privacy Preservation for V2G Networks in Smart Grid: A Survey*, 91 COMPUT. COMMC'N 17, 17-28 (2016) (concluding that adversarial algorithms can detect individuals otherwise clouded in an anonymity network by compiling various data outside the anonymity network which provides inferences about the individual item's presence in the anonymized group).

may reveal that the image is fake and even begin to point towards a particular country or even individuals inside a country. However, this will take a few days to confirm. By the time this task is complete, the realistic-looking video will have already done its damage.

Additionally, combatants are in even more trouble if the deepfake content is not a photo or video image of a person. If, for example, an enemy unit devises a deepfake-enabled voice recording or voice masker and manages to call their adversary unit's commander directly to make the unit commander think his superior is instructing him to surrender to the enemy (a capability made progressively more real today by such developers as WellSaid Labs[143] and Google),[144] nothing in the arsenal of computer science research can currently combat this tactic.

Certainly, the computer sciences would not be alone in any of these scenarios to help reveal a fraud. Sophisticated deepfake content still requires extremely skilled developers. So, the synthetic content in both of these scenarios may demonstrate enough imperfections to trigger quick scrutiny and provide signs, along with various degrees of intelligence collection, that can point to a responsible office or even person.[145] Also, the use of deepfake in several armed conflict scenarios, such as in an international armed conflict between two states or a non-international armed conflict between long-time familiar enemies, will logically facilitate finger-pointing before digital forensics can even tie its proverbial shoes. However, while progress has proceeded quickly, we still remain quite a long way from having automated networks which can detect deepfakes across the media spectrum and quickly attribute them to human actors.

---

[143] WELLSAID LABS, https://wellsaidlabs.com/ (last visited Feb. 25, 2021).

[144] Google's artificial intelligence development group DeepMind, for example, works expressly to "solve intelligence" by replicating the brain's physiological and mathematical progressions in order to imitate and train human-like thought processes in artificial intelligence. Part of DeepMind's various programs is one called "WaveNet" which seeks to train artificial neural networks to develop realistic-sounding text-to-speech audio capabilities in order to assist the disabled. DEEPMIND, https://deepmind.com/ (last visited Jan. 5, 2021); *see also* Yutian Chen et al., *Using WaveNet Technology to Reunite Speech-Impaired Users with Their Original Voices*, DEEPMIND (Dec. 18, 2019), https://deepmind.com/blog/article/Using-WaveNet-technology-to-reunite-speech-impaired-users-with-their-original-voices.

[145] This is essentially how the Zelenskyy deepfake was so quickly debunked. Its production value was relatively low likely owing to the hasty nature of its creation. Viewers were able make out lighting inconsistencies, odd head-to-body proportionality, image blurriness, and could perhaps most easily tell that the video was fake due to the poor quality of Mr. Zelenskyy's depicted voice. *Supra* note 13.

C. Agency and Attribution: Legal Analysis

Legally attributing an act of deepfake-deception to an individual, an organization, or a country is also complicated. Because deepfake is an act of artificial intelligence that utilizes node-based neural networks within cyberspace often to achieve objectives through cyberspace, deepfake invokes legal equities related to cyberspace operations.[146] Deepfake technology can also pair with other classic examples of cyber activities, such as ransomware, to conduct a cyberattack.[147] However, deepfake does not have the same purpose as typical cyberspace operations such as distributed denial of service attacks on servers supporting an adversary's headquarters. Deepfake is a means of deception and hence also bears legal equities related to information operations.[148]

The current best source for modern perspectives on attribution for cyber activities conducted prior to or during an armed conflict are not in a law, but in a manual. Published in 2017, the Tallinn Manual 2.0[149]

---

[146] *See* U.S. DEP'T OF DEF., LAW OF WAR MANUAL ¶ 16.1.2 (May 2016) [hereinafter DOD LAW OF WAR MANUAL] (citing JOINT PUBLICATION 3-0, *Joint Operations* (Aug. 11, 2011)); JOINT PUBLICATION 3-12, *Cyberspace Operations*, GL-4 (Feb. 5, 2013) (defining cyberspace as a "global domain within the information environment consisting of interdependent networks of information technology infrastructures and resident data, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers.").

[147] *See e.g.*, Jovi Umawing, *The Face of Tomorrow's Cybercrime: Deepfake Ransomware Explained*, MALWAREBYTES LABS (Jun. 26, 2020), https://blog.malwarebytes.com/ransomware/2020/06/the-face-of-tomorrows-cybercrime-deepfake-ransomware-explained/.

[148] *See* JOINT PUBLICATION 3-13, *Information Operations*, GL-3 (Nov. 27, 2012 (incorporating Change 1, Nov. 20, 2014)) (defining information operations as the "integrated employment . . . of information-related capabilities . . . to influence, disrupt, corrupt, or usurp the decision-making of adversaries and potential adversaries while protecting our own."). While U.S. Department of Defense doctrine, which is currently silent on deepfake, would not organize deep-fake operations into cyberspace operations, this perspective does not seem to be universal. *Compare* DOD LAW OF WAR MANUAL, *supra* note 146 at ¶¶ 16.1.2.1 and 16.1.2.2 (stating cyber operations "use computers to disrupt, deny, degrade, or destroy information resident in computers and computer networks" but that "operations to distribute information broadly using computers would generally not be considered cyber operations.") *with* Citron, *supra* 41, at 1801 (highlighting domestic liability for deepfake-based crimes in federal cyberstalking laws under 18 U.S.C. § 2261A); DANIELLE CITRON, HATE CRIMES IN CYBERSPACE (2014); Mike Faden, *Malicious Deepfake Technology: A Growing Cyber Threat*, MIMECAST (Jul. 13, 2020), https://www.mimecast.com/blog/malicious-deepfake-technology-a-growing-cyber-threat/; *see also* Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224, 227 (2011) (discussing impacts from "cyberspace harassment").

[149] TALLINN MANUAL ON INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (Michael N. Schmitt ed., 2d ed. 2017) [hereinafter TALLINN MANUAL 2.0].

absorbed the first iteration of the Tallinn Manual issued in 2013[150] and, despite being only a statement by international legal experts and not a law itself, it is nonetheless an influential source in an armed conflict-related field that has virtually no subject-specific multilateral treaties[151] and is otherwise light on expressions of customary international law.[152]

The Tallinn Manual 2.0 articulates several elements of international law that are vital for ascertaining attribution. It recognizes for instance that international law provides States with sovereignty in cyberspace.[153] States also have a duty to exercise due diligence to ensure they do not allow their territory or cyber infrastructure to be used to produce "serious adverse consequences" for other States.[154] States may exercise territorial and extraterritorial jurisdiction over cyber activities or persons involved in cyber activities that cause substantial effects in the State.[155] Taking inspiration from the Draft Articles on Responsibility of States for Internationally Wrongful Acts,[156] the Tallinn Manual 2.0 also

---

[150] TALLINN MANUAL ON INTERNATIONAL LAW APPLICABLE TO CYBER WARFARE (Michael N. Schmitt ed., 2013) [hereinafter TALLINN MANUAL 1.0].

[151] Some multinational treaties that might have application in a non-armed conflict context include the 2000 Palermo Convention and the 2001 Budapest Convention. United Nations Convention Against Transnational Organized Crime, Nov. 15, 2000, 2225 U.N.T.S. 209 [hereinafter Palermo Convention]; Convention on Cybercrime, Nov. 8, 2001, E.T.S. 185 [hereinafter Budapest Convention]. The United States has ratified and is party to both treaties.

[152] For a discussion on how the Tallinn Manual iterations accompany expressions of international law, *see* Eric Talbot Jensen, *The Tallinn Manual 2.0: Highlights and Insights*, 48 GEO. J. INT'L L. 735, 738 (2017). Mr. Jensen was a member of the International Group of Experts who met at the NATO Cooperative Cyber Defense Center of Excellence in Tallinn, Estonia to develop the Tallinn Manuals.

[153] TALLINN MANUAL 2.0, *supra* note 149, at 11 r. 1., 16 r. 3, 17 r. 4 (providing that "[t]he Principle of Sovereignty applies to cyberspace" and that "it is a violation of territorial sovereignty for an organ of a State, or others whose conduct may be attributed to the State, to conduct cyber operations while physically present on another State's territory against that State or entities or persons located there."). The Manual acknowledges that it is not settled international law as to whether violation of sovereignty in cyberspace by itself constitutes an internationally wrongful act or whether sovereignty just acts as a mere rule. *See* Jensen, *supra* note 152, at 741-42 (citing Gary Corn, *Tallinn Manual 2.0 – Advancing the Conversation*, JUST SECURITY (Feb. 15, 2017, 8:41 am), https://www.justsecurity.org/37812/tallinn-manual-2-0-advancing-conversation/#more-37812).

[154] TALLINN MANUAL 2.0, *supra* note 149, at 30 r. 6. As Mr. Jensen points out, this rule does not prohibit all harm – just that harm which results in serious adverse consequences. Jensen, *supra* note 152, at 744.

[155] TALLINN MANUAL 2.0, *supra* note 149, at 51 r. 8. Particularly, Rule 9 provides that States can exercise jurisdiction over "cyber infrastructure and persons engaged in cyber activities on its territory," cyber activities "originating in, or completed on, its territory," or cyber activities causing "substantial effect" in its territory. *Id.* at 55 r. 9.

[156] For related discussion *see* Jensen, *supra* note 152, at 750.

provides that States "bear international responsibility for a cyber-related act that is attributable to the State . . ."[157]

But who is "the State"? Rule 15 seeks to clarify that "[c]yber operations conducted by organs of a State, or by persons or entities empowered by domestic law to exercise elements of governmental authority, are attributable to the State."[158] This clarification of course only raises more questions about what qualifies as an "organ," what domestic law would need to do to show empowerment, where does the divide lay between element and non-element, and so forth.

The Manual explains that the term "State organ" has "broad meaning to ensure that States do not escape responsibility by asserting an entity's non-status as its organ in domestic law."[159] It provides that the "clearest case" occurs when State military or intelligence agencies commit the acts, listing U.S. Cyber Command and Israel's Unit 8200 as examples.[160] In order to cast a wide net, however, the Manual adopts the perspective of the Draft Articles on Responsibility as well as the International Court of Justice, stating:

> "[P]ersons, groups of persons or entities may, for the purposes of international responsibility, be equated with State organs even if that status does not follow from internal law, provided that in fact the persons, groups or entities act in 'complete dependence' on the State, of which they are ultimately merely the instrument."[161]

But despite this language, quickly the net begins to narrow. The burden to show that a person or entity must act in "complete dependence" of the State is not low. There must be a showing that a "particularly great degree of State control" exists over the person or entities concerned[162] and that when determining this, the key factors are "the function of the entity" and the "State's intention" concerning the person or entities because even State ownership of an entity is not

---

[157] *Id.* (citing TALLINN MANUAL 2.0, *supra* note 149, at 84 r. 14).
[158] TALLINN MANUAL 2.0, *supra* note 149, at 87 r. 15.
[159] *Id.* at 87-88 ¶ 3 (referencing Int'l Law Comm'n, Draft Articles on Responsibility of States for Internationally Wrongful Acts, Rep. of the Int'l Law Comm'n on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10, at art. 4(2) (2001) [hereinafter Draft Articles on Responsibility]).
[160] *Id.* at 87 ¶ 1.
[161] *Id.* at 88 ¶ 4 (quoting Application of Convention on Prevention and Punishment of Crime of Genocide (Bosn. & Herz. v. Serb. & Montenegro), Judgment, 2007 I.C.J. 47 ¶ 392 (Feb. 2007) [hereinafter I.C.J. Genocide Case]).
[162] *Id.* (citing I.C.J. Genocide Case, *supra* note 161, at ¶ 393).

enough to demonstrate requisite State control.[163] While responsibility may still attach for *ultra vires* acts that exceed State grants of authority, the actor must nonetheless still appear "under colour of authority."[164]

The purpose of this narrowing is that the Tallinn Manual 2.0—like international law when it comes to activities in general in cyberspace—only sees international legal support for holding *states* responsible for internationally wrongful acts. Individuals or entities in most contexts would only be subject to the jurisdiction of domestic law or certain specific treaties. This is why Rule 17 provides that cyber operations by non-state actors are only attributable to states if the non-state actor acts "pursuant to [the state's] instructions or under its direction or control" or if the state "acknowledges and adopts" the non-state actor's activities as their own.[165] Thus even if a state gave malware to a terrorist organization and the terrorist organization then decided on its own to independently plan and execute an offensive cyber operation with that malware, the Manual would not legally attribute the cyber operation to the state unless the state later adopted the cyber operation as its own.[166]

The result is that in the context of cyberspace operations, *lex generalis* provides that acts are legally attributable to only one entity—states—and therefore in such a case only states would need to be concerned about countermeasures. By this view, terrorists, insurgents, stateless militias, hacktivists, non-governmental organizations, Silicon Valley titans, Silicon Valley start-ups, protestors, risk-inclined college students, and bored teenagers would not face jeopardy under international law for deploying deepfake deception which, in times of peace, is not a *per se* international crime. However, as the doctrine of *lex specialis derogat legi generali* explains, specified international laws override general law.[167] The laws of armed conflict are precisely the *lex specialis* which might bridge the gap in legal attribution for deepfake-derived deception—in both cyber and information operation contexts—when it may not seem to otherwise exist.

---

[163] *Id.* at 88 ¶ 5.
[164] *Id.* at 89 ¶ 7.
[165] *Id.* at 94 r. 17.
[166] *Id.* at 97 ¶ 8.
[167] *See id.* at 80 ¶ 5 (citing commentary to Draft Articles on Responsibility at art. 55); *see also* Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. Rep. 226, at 25 (Jul. 8); Anja Lindroos, *Addressing Norm Conflicts in a Fragmented Legal System: The Doctrine of Lex Specialis*, 74 NORDIC J. INT'L L. 27, 35-39 (2005) (tracing the history of the doctrine back to Roman law and its later development from Hugo Grotius to the International Court of Justice).

IV. DISINFORMATION AND THE LAWS OF ARMED CONFLICT (LOAC).

A. Ruse

The law generally categorizes any lawful deception as a ruse.[168] The U.S. Department of Defense (DoD) broadly defines a ruse as a "trick of war designed to deceive the adversary, usually involving the deliberate exposure of false information to the adversary's intelligence collection system."[169] Joint Publication 3-13.4, which provides baseline DoD policy on military deception activities, characterizes a ruse as a "cunning trick designed to deceive the adversary to obtain friendly advantage."[170]

Both of these definitions derive primarily from two sources of international law—the Hague Conventions and the 1977 Additional Protocol I to the Geneva Conventions. The regulations featured in the 1907 Hague Convention concerning the Laws and Customs of War on Land ("Hague Convention IV") provide at Article 24 that generally

---

[168] *See e.g.*, U.S. DEP'T OF ARMY, FIELD MANUAL 6-27, THE COMMANDER'S HANDBOOK ON THE LAW OF LAND WARFARE, ¶ 2-171 (7 Aug. 2019) [hereinafter FM 6-27]. This is also true in a domestic law sense though the use of deception, particularly in law enforcement circumstances, can encounter significantly more skepticism than in a combat scenario. *Compare e.g.* Nadia B. Soree, *Thank You All the Same, but I'd Rather not be Seized Today: The Constitutionality of Ruse Checkpoints Under the Fourth Amendment*, 66 BUFFALO L. REV. 385, 433-34 (2018) (arguing that the use of "ruse checkpoints" violates the Fourth Amendment) *with* Daniel R. Dinger & John S. Dinger, *Deceptive Drug Checkpoints and Individualized Suspicion: Can Law Enforcement Really Deceive its Way into a Drug Trafficking Conviction?*, 39 IDAHO L. REV. 1, 29-55 (2002) (arguing that deceptive checkpoints can be just as lawful a manner of ruse as the use of undercover techniques and are not per se violative of the Fourth Amendment). This paper, however, does not seek to explore domestic impacts of deep-fake technology outside of the context of armed conflict.

[169] JOINT CHIEFS OF STAFF, JOINT PUB. 1-02, DEPARTMENT OF DEFENSE DICTIONARY OF MILITARY AND ASSOCIATED TERMS 207 (8 Nov. 2010, as amended through 15 Feb. 2016); *see also* NAT'L SEC. LAW DEP'T, THE JUDGE ADVOCATE GEN.'S LEGAL CTR. & SCH., OPERATIONAL LAW HANDBOOK (2018) (this publication was amended in 2020 at which time the publication opted instead to lean more on the definition of ruse provided in the Hague Regulations discussed *infra*).

[170] JOINT CHIEFS OF STAFF, JOINT PUB. 3-13.4, MILITARY DECEPTION, ¶ 11(c)(3) (26 Jan. 2012). JP 3-13.4 takes the additional step of distinguishing ruses from other similar acts of deception such as feints ("an offensive action involving contact with the adversary conducted for the purpose of deceiving the adversary as to the location and/or time of the actual main offensive action") and displays ("the simulation, disguising, and/or portrayal of friendly objects, units, or capabilities in the projection of the MILDEC story"). *Id.* at ¶ 11(c)(1),(4). Notably this regulation, promulgated at the same echelon and near in time to JP 1-02, avoids the JP 1-02 narrowing of the definition of ruse to those acts which interact with "[an] adversary's intelligence collection system," focusing instead on the objective of employing a ruse, namely, to obtain "friendly advantage." With respect to the Army, however, FM 6-27 goes to significant effort to encompass both dynamics so that the definition of ruse is not limited in either respect. FM 6-27, *supra* note 168, at ¶¶ 2-172, 2-173.

speaking "ruses of war . . . are considered permissible."[171] However, this article does not attempt to redefine ruse—instead, it implies that a ruse is anything that is not expressly forbidden by the regulations.[172]

Additional Protocol I to the Geneva Conventions ("API"), however, provides clarity to the concept of ruse that still controls today.[173] Bearing in mind that API applies to Common Article 2 international armed conflicts only,[174] Article 37 of API provides at Section 2 that "[r]uses of war are not prohibited."[175] Article 37 defines a ruse as those acts "intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict."[176] Furthermore, it distinguishes a ruse from an act of perfidy by explaining that a ruse "[does] not invite the confidence of an

---

[171] Convention (IV) Respecting the Laws and Customs of War on Land and its Annex: Regulations Concerning the Laws and Customs of War on Land art. 24, Oct. 18, 1907, 36 Stat. 2277 [hereinafter 1907 Hague Convention IV]. The Hague Conferences, both the 1907 meeting and the earlier 1899 meeting, were themselves inspired by preceding benchmark regulations on armed conflicts, most notably the famous Lieber Code promulgated by the Lincoln Administration during the American Civil War as well as the 1874 Brussels Declaration and the 1880 "Oxford Manual" on the laws of war on land by the Institute of International Law. *See* Sean Watts, *Law-of-War Perfidy*, 219 MIL. L. REV. 106, 125-37 (2014).

[172] 1907 Hague Convention IV, *supra* note 171, at art. 24. This is most likely due to the fact that this language came from the 1899 Hague Conventions which themselves borrowed enormously from the 1874 Brussels Declaration. Convention (II) Respecting the Laws and Customs of War on Land and its Annex: Regulations Concerning the Laws and Customs of War on Land art. 24, Jul. 29, 1899, 32 Stat. 1803 ("1899 Hague Convention II"); *see also* Watts, *supra* note 171, at 137 n.103.

[173] While the United States is not a party to Additional Protocol I to the Geneva Conventions, the United States acknowledges that several portions of the Protocol are customary international law and therefore seeks to abide by those portions. As for Article 37, the United States recognizes it in its entirety to be customary international law. Michael J. Matheson, *Remarks in Session One: The United States Position on the Relation of Customary International Law to the 1977 Protocols Additional to the 1949 Geneva Convention*, 2 AM. U. J. INT'L L. & POL'Y 419, 425 (1987).

[174] Protocol Additional to the Geneva Conventions of 12 Aug. 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 1, Jun. 8, 1977, 1125 U.N.T.S. 3 [hereinafter Additional Protocol I]. Article 1(4) seeks to apply API beyond the scope of international armed conflicts to conflicts involving efforts to throw off colonial domination, alien occupation, or fight racist regimes (a.k.a. conflicts of "national liberation"), circumstances which often involve non-international armed conflict. Many countries including the United States expressly reject this expansion. *See* Matheson, *supra* note 173. Hence, as is discussed *infra*, the discussion of perfidy to follow would not apply in Common Article 3 conflicts. *Compare also* Additional Protocol I at arts. 37-39 *with* Protocol Additional to the Geneva Conventions of 12 Aug. 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts, Jun. 8, 1977, 1125 U.N.T.S. 609 [hereinafter Additional Protocol II] (pertaining solely to Common Article 3 non-international armed conflicts yet omitting any focused discussion on perfidy, misuse of recognized emblems, or misuse of emblems of nationality).

[175] Additional Protocol I, *supra* note 174, at art. 37(2).

[176] *Id.*

adversary with respect to protections under that law."[177] Article 37 then provides a short list of acts which would qualify as a ruse to include "the use of camouflage, decoys, mock operations, and misinformation."[178]

The Article's explicit reference to "misinformation" as a lawful example of deception accepts the reality that trickery has and to some degree should be a part of war. The Diplomatic Conference which promulgated API expressly recognized this when it considered how to draft Article 37. As the International Committee of the Red Cross (ICRC) Reading Commission wrote in its Commentary on the Additional Protocols of 8 June 1977 when discussing the Conference's perspective on ruse, "[t]he art of warfare is a matter, not only of force and of courage, but also of judgment and perspicacity. In addition, it is no stranger to cunning, skill, ingenuity, stratagems and artifices, in other words, to ruses of war, or the use of deception."[179] The Commentary even goes so far as to concede that, while it can cause significant problems, deception is "a just and necessary means of hostility."[180]

At the same time, however, Article 37's drafters acknowledged that setting parameters on lawful deception was harder to do than setting parameters on unlawful deception. Its list of examples of ruse is purposefully broad and non-exclusive because the Conference understood it would be a fool's errand to try to predict the limits of human creativity.[181] This appears to have everything to do with why the Article defines perfidy, discussed more below, first[182] and then defines ruse in contradistinction of perfidy.

The Commentary does offer an affirmative definition of ruse by explaining that a ruse "consists either of inducing an adversary to make a mistake by deliberately deceiving him, or of inducing him to commit an imprudent act, though without necessarily deceiving him to this end."[183]

---

[177] *Id.*
[178] *Id.*
[179] INT'L COMM. RED CROSS, Commentary, ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949, at 439-440 (Yves Sandoz et al. eds, 1987) [hereinafter API Commentary] (referencing the use of the term 'ruses of war' by Carl Von Clausewitz).
[180] *Id.* at 440 n.49 (citing Adjutant Gen.'s Office, U.S. Dep't of War, Gen. Orders No. 100, *Instructions for the Government of Armies of the United States in the Field*, Art. 101 (Apr. 24, 1863) [hereinafter *Lieber Code*]).
[181] *Id.* at 443 (explaining "It was impossible to enumerate in the Protocol all the operations described under this heading . . ."). The Commentary also noted that the examples proposed, and ultimately included, did not provoke any debate at the Conference. *Id.*
[182] Additional Protocol I, *supra* note 174, at Art. 37(1).
[183] API Commentary, *supra* note 179, at 441.

However, its discussion quickly branches again into contrasting this characterization from acts that would not qualify as a ruse.[184]

More helpful insight on what could qualify as a ruse under Article 37 comes from the Commentary's list of examples. This includes such "commonly described" ruses of war as ambushes, simulated operations on land, air, or sea, camouflaging troops or positions "in the natural or artificial environment," and even laying dummy mines.[185] Most notably for the purposes of this article, the Commentary also listed acts which remarkably seem to foretell the evolution of 20th century electronic and cyber warfare. These ruses included:

> ". . . [T]ransmitting misleading messages by radio or in the press; knowingly permitting the enemy to intercept false documents, plans of operations, despatches [sic] or news items which actually bear no relation to reality, using the enemy wavelengths, passwords and wireless codes to transmit false instructions; pretending to communicate with reinforcements which do not exist . . . using signals for the sole purpose of deceiving an adversary; resorting to psychological warfare methods by inciting the enemy soldiers to rebel, to mutiny or desert, possibly taking weapons and transportation; inciting the enemy population to revolt against its government etc."[186]

In fact, in attempting to delineate the multiple ways that a ruse could lawfully occur, the Conference ultimately had to toss up its hands and declare "the imagination of man is too inventive for one to think that everything it could come up with can be covered in a list."[187] The drafters wisely conceded that the evolution of combat is "unforeseeable" and presciently that its nature "will always give rise to new ideas."[188]

B. Perfidy

For the above reasons, then, much more effort has gone into defining what a ruse is not rather than what it is, and if a lawful deception

---

[184] *Id.* The Commentary particularly here contrasts a ruse from a "prohibited ruse" discussed further *infra* which itself contrasts against perfidy.
[185] *Id.* at 443.
[186] *Id.* at 443-44.
[187] *Id.* at 444.
[188] *Id.*

is a ruse, its opposite is perfidy and treachery. International law strictly defines p—more so than is often realized. However, whether a use of deepfake-generated content would amount to perfidy is not correspondingly easy to define.

Article 37, Section 1 of API observes that perfidy is those actions "inviting the confidence of an adversary to lead him to believe that he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence."[189] As the Commentary explains about this definition, perfidy "consists of the deliberate use of international law protection[s] . . . to deceive the adversary."[190]

Just like in its discussion on ruse, Article 37 also provides a list of acts which would constitute perfidy. These acts include: feigning an intent to negotiate under a flag of truce or surrender, feigning incapacitation by wounds or sickness, feigning civilian, non-combatant status, and feigning protected status by the use of signs, emblems, or uniforms of the United Nations or of states not party to the conflict.[191]

Based on the language in Article 37, perfidy has four elements: (1) inviting the confidence of an adversary, (2) with the intent of betraying that confidence, (3) betraying that confidence through the claim or demonstration of an unmerited protection afforded by international law applicable in armed conflict, and (4) killing, injuring, or capturing an adversary as a result.[192] The first two elements tend to be somewhat straightforward, though cyber dynamics can muddy what it would take to prove an "inviting."[193] The third element can be very broad in practice. It does not confine perfidy to acts which invoke Article 37 protections, or API protections, or even protections under just the Geneva Conventions. The Commentary points out that Article 2 of API provides that the laws and rules applicable to armed conflict extend outside of the Geneva Conventions to also encompass not only bilateral agreements between parties to a conflict but also "generally recognized principles and rules of

---

[189] Additional Protocol I, *supra* note 174, at Art. 37(1).
[190] API Commentary, *supra* note 179, at 444.
[191] Additional Protocol I, *supra* note 174, at art. 37(1)(a)-(d). Acts which seek to use the flag, emblem, uniform, or other insignia of a State which is not a party to the conflict, or which seek to use such emblems of an adversary, are also prohibited independently by Article 39. *Id.* at art. 39(1)-(2).
[192] *Id.* at art. 37(1). *See also* API Commentary, *supra* note 179, at 435 (only enumerating the first three elements but recognizing the fourth in later discussions).
[193] For example, a line of malicious code which does not activate until a target clicks a link or downloads a file may or may not "invite" a particular confidence yet could theoretically still be perfidious in nature.

international law which are applicable to armed conflict."[194] In other words, the protection claimed can possibly have no direct correlations to the Geneva Conventions or the Hague Conventions but still constitute perfidy.[195]

The fourth element, however, imposes a quizzical limitation. By narrowing perfidy to only those situations that produce specific consequences i.e., "to kill, injure, or capture an adversary,"[196] Article 37 dramatically narrows the purpose of enumerating perfidy at all and threatens the ability to regulate other actions that would undercut trust in the laws of armed conflict. To be fair, this possibility was not lost on the API Conference. Questions arose immediately as to why the use of deception that disabuses protections under the laws of armed conflict to achieve any other objective other than killing, injuring, or capturing prisoners—such as seizing an enemy fighting position or delaying an attack—would not also constitute perfidy.[197] Additionally, delegates wondered openly whether an *attempt* to kill, injure, or capture an adversary through perfidy would also constitute perfidy since the Article and the rest of API were silent on inchoate offenses.[198]

Unfortunately, the Conference does not appear to have provided Jean Pictet and his peers drafting the Commentary much to answer these questions. The Commentary instead notes that the drafters of the Article considered that there remained "a sort of gray area of perfidy which is not explicitly sanctioned as such, in between perfidy and ruses of war," leading to a "permanent controversy in practice as well as in theory."[199]

The Commentary did manage, however, to construct an analysis on how to read Article 37 in relation to the rest of API which helps to broaden the application of perfidy to consequences beyond killing, injuring, or capturing prisoners. First, the Commentary observes that attempts or unsuccessful acts of perfidy also fall within the definition of

---

[194] API Commentary, *supra* note 179, at 435 (citing Additional Protocol I, *supra* note 174, at Art. 2(b)).
[195] The Commentary observes "the definition of perfidy extends beyond the prohibition formulated . . ." *Id.* It notes, for example, that there are protections at sea provided by the laws of armed conflict that API does not entertain.
[196] Additional Protocol I, *supra* note 174, at Art. 37(1). The Commentary observes that this finite, exclusive list was a direct adoption from the 1907 Hague Convention IV which sought to make it illegal "to kill or wound treacherously." API Commentary, *supra* note 179 at 432 (citing 1907 Hague Convention IV, *supra* note 171, at Art. 23(b)). The Conference decided to add "capturing" as an additional nod to the combat nature of perfidy but seem to have limited it there because agreements on expanding the definition to other acts had become too difficult.
[197] API Commentary, *supra* note 174, at 432-33.
[198] *Id.*
[199] *Id.* at 433.

perfidy, although without providing any explanation why other than "it seems evident."[200] Second, the Commentary reminds readers that the Vienna Convention on the Law of Treaties demurs against interpreting treaties to conflict with a peremptory norm of general international law, and that any related peremptory norms should be read into perfidy as a result.[201] Third, Articles 38 and 39 reinforce Article 37 by prohibiting the misuse of recognized emblems under the Geneva Conventions as well as the misuse of emblems of either non-party or adversary states, respectively,[202] thus capturing a large bulk of related concerning behavior.

This latter interpretation may not sit on firm ground. It implicitly relies on Articles 37-39 to become customary international law, if not universally ratified and adopted. It does not contend with the fact that powers such as the United States would not and still have not ratified API and, unlike Articles 37 and 38, does not consider Article 39 to be customary international law.[203] It is even less ready to resolve applicability in the face of inter-government disagreements about applicability, such as how the United States has accepted that Article 37 reflects customary international law[204] (and is therefore binding on the United States) but its own Department of Defense has declared that Article 37's "capture" is actually not a part of customary international law.[205] Nonetheless, these observations remain helpful for discerning a wider landscape in which to declare acts beyond those resulting in killing, injuring, or capturing individuals to be perfidious.

---

[200] *Id.*

[201] *Id.* (citing Vienna Convention on the Law of Treaties art. 53, May 23, 1969, 1155 U.N.T.S. 331; 8 I.L.M. 679 [hereinafter Vienna Convention]).

[202] *Id.*

[203] *See* Matheson, *supra* note 173, at 425.

[204] *Id.*

[205] DOD LAW OF WAR MANUAL, *supra* note 146, at ¶ 5.22.2.1. The DoD Law of War Manual does not provide any authority on which it bases its interpretation. FM 6-27 is careful to only define perfidy as "wounding or killing" the enemy while possessing a protected status. FM 6-27, *supra* note 168, at ¶¶ 1-82, 2-91, 2-109, 2-151, 2-152 (citing the DoD Law of War Manual; the latter paragraph only invites the reader to "consider" API, Art. 37). The only discussion of capture and perfidy in FM 6-27 instead relates that "any combatant who feigns death in the hope of evading capture has not engaged in perfidy," a notion which would not offend Articles 37-39. *Id.* at ¶ 2-152.

C. Treachery a.k.a. Violations of Honor

Scholars do not usually discuss this third category explicitly as "violations of honor" but instead in terms of "treachery" to touch on the concept's historical roots, usage, and proximity to perfidy.[206]

1. Chivalry and Honor

Battlefield concepts of "chivalry" referenced in international humanitarian law today are well-documented as originating in Europe during the Middle Ages[207] and have still managed to persist, albeit to varying degree, into the 21st century.[208] The API Commentary acknowledged the role of chivalry in fostering concepts of honor also found in contemporary laws of armed conflict, noting that "[t]his sense of honour, which was nourished during the Middle Ages of Europe by chivalry, particularly in tournaments and in jousting, has contributed to the establishment of the rules which finally became assimilated into the customs and practices of war . . ."[209] The Commentary characterizes the battlefields of Medieval Europe, or at least the Christian warriors at that time, as steeped in "rules for attack and rules for defence," and that undergirding conduct in battle was the notion that "the knight always trusted the word of another knight, even if he were an enemy."[210] This notion was so strong, the Commentary posited, that "[p]erfidy was

---

[206] *See generally* Watts, *supra* note 171. Whether treachery is actually a distinct concept from perfidy is still debatable. Significant historical evidence does support the position that they are substantively different with the former more concerned about violations of ethical or chivalrous expectations of good-faith behavior and the latter concerned about abuses of international law's protections in ways which could neutralize the law itself. *Id*. at 109, 113-14, 125-29, 134-37, 140-41 (discussing distinctions between the two concepts as found in, among other things, the Lieber Code, the 1907 Hague Regulations, the prosecution of defendants at the Tokyo International Military Tribunal, and the 2009 Military Commissions Act).

[207] *See* Watts, *supra* note 171 at 106, 157-58 (citing Geoffrey Parker, *Early Modern Europe*, and Robert C. Stacey, *The Age of Chivalry*, *in* THE LAWS OF WAR 29-31, 54 (Michael Howard, George J. Andreopoulos, & Mark Shulman eds., 1994)).

[208] *See e.g.*, JUDGE ADVOC. GEN., CANADIAN ARMED FORCED, LAW OF ARMED CONFLICT AT THE OPERATIONAL AND TACTICAL LEVELS, B-GJ-005-104/FP, 2-1 (2001) (declaring chivalry to be a core principle of the laws of armed conflict). Indeed for 63 years before it was updated in 2019, the U.S. Army's primary field manual on the laws of armed conflict expressly required that U.S. Soldiers abide by chivalry as a core principle of armed conflict. U.S. DEP'T OF ARMY, FIELD MANUAL 27-10, THE LAW OF LAND WARFARE, para. 3(a) (Jul. 1956) [hereinafter FM 27-10] (superseded by FM 6-27, *supra* note 168).

[209] API Commentary, *supra* note 179, at 434.

[210] *Id.*

considered a dishonour which could not be redeemed by any act, no matter how heroic."[211]

However, the Commentary had to acknowledge that notions of chivalry, honor, and good faith had no single origin, and that no particular culture could claim sole ownership over any of the concepts.[212] While chivalry *per se* may have originated in Christian Europe, it could not reliably inform modern notions of treachery. Medieval Christian warriors, of course, did not always abide by their own oaths of chivalry. Often, they were prone to abandon their chivalric code—with no legal or immediate political consequence—when facing non-Christian adversaries, such as during the First Crusade when Crusader armies brutally stormed Jerusalem in 1099, slaughtered many of the city's inhabitants, and desecrated several holy sites.[213]

However, in contrast to the limited application of chivalry, honor and good faith have been facets of armed conflict across the world. The Islamic warriors who opposed Crusaders at Jerusalem, for example, and in later Crusades had their own ethics of honor, founded in their own religious beliefs and world views rather than European or Christian culture.[214] The famous samurai warriors of Japan were required to prize

---

[211] *Id.* The view that acts of treachery by knights could incur a lifelong bounty has some support. *See, e.g.*, Watts, *supra* note 171, at 106 (citing Parker, *supra* note 207) (stating "medieval notions of honor and chivalry sanctioned unending blood feuds to avenge knights killed by treachery.").

[212] API Commentary, *supra* note 179, at 434 (observing "Perfidy is injurious to the social order which it betrays, regardless of the values on which this social order is founded.").

[213] *See, e.g.*, JAY RUBENSTEIN, ARMIES OF HEAVEN: THE FIRST CRUSADE AND THE QUEST FOR APOCALYPSE 290 (2011) (quoting Raymond of Aguiler, a French participant in the sack of Jerusalem, who boasted that in the city, "[s]ome had their heads cut from their bodies (which was fairly merciful) or were hit with arrows and forced to jump from towers. Others suffered for a long, long time, and were consumed and burned up in flames. Horses and men on public roads were walking over bodies. But these things I say are trifling. Let us go to the Temple of Solomon.").

[214] The distinguished 12th century Muslim warrior and writer Usama Ibn Munqidh, a native of modern-day Syria and witness to the Second Crusade, wrote extensively about the multi-cultural world of the contemporaneous Near East. His writings often contrasted his views on honor and good behavior, informed by his own Islamic beliefs, with the behavior and lack of honor he perceived of "Franks" (as he called all Europeans, even if they did not come from France) who he characterized as unintelligent and "[possessing] nothing in the way of regard for honour or propriety." USAMA IBN MUNQIDH, THE BOOK OF CONTEMPLATION: ISLAM AND THE CRUSADES 144, 148 (Paul M. Cobb trans., Penguin Group 2008) (1183). In fact, in one anecdote he conveys that the invitation from a close Christian friend for Usama's son to come to Europe to "acquire reason and chivalry" was kind but laughable and was carefully refused. *Id.* at 144. Historian Will Durant has observed that during the Crusades the Islamic forces, while themselves not strangers to inflicting suffering or division, on the whole "seem to have been better gentlemen than their Christian peers; they kept their word more frequently, showed more mercy to the defeated, and were seldom guilty of

honor as "more important than life itself."[215] Today, the famous *pukhtunwali* code in Afghanistan, revered most fervently by the country's Pashtun population who are also called to abide by it even in battle, maintains honor as one of its three pillars (the other two being hospitality and revenge/justice).[216]

While chivalry may not be as fashionable a concept today as it once was in international humanitarian law, honor and good faith are still relevant. The United States Army continues, as it has for several decades, to declare "honor" to be one of its seven Army Values along with related concepts such as "respect, duty, loyalty, selfless service, integrity, and personal courage, in everything Soldiers and Marines do."[217] Additionally, the Army very recently confirmed the continued legal relevance of honor in FM 6-27 which effectively sidelines "chivalry" in favor of "honor." Characterizing honor as a "core Army and Marine Corps value,"[218] FM 6-27 declares honor as a "basic LOAC principle" in line with the other four historically-accepted principles of distinction, proportionality, military necessity, and humanity.[219] FM 6-27 defines the concept as "[t]he LOAC principle [sic] that demands a certain amount of fairness in offense and defense and a certain mutual respect between opposing forces."[220] Honor "gives rise to rules that help enforce and give effect to LOAC"[221] and "provides legitimacy to the entire endeavor."[222] While the concept does not define its limits, FM 6-27 observes that the

---

such brutality as marked the Christian capture of Jerusalem in 1099." WILL DURANT, THE STORY OF CIVILIZATION: PART IV, THE AGE OF FAITH 341 (1950).

[215] *See* Nicholas W. Mull, *The Honor of War: Core Value of the Warrior Ethos and Principle of the Law of War*, 18 CHI.-KENT J. INTL'L & COMP. L. 1, 23 (2018) (citing YAMAMOTO TSUNETOMO, HAGAKURE: THE BOOK OF THE SAMURAI 30 (William Scott Wilson trans., Kodansha Int'l 1979) (1716)).

[216] *See, e.g.*, Ken Guest, *Dynamic Interplay Between Religion and Armed Conflict in Afghanistan*, 92 INT'L REV. RED CROSS 877, 886 (2010). Mr. Guest observes in his article that *pukhtunwali* "represents an ideal rather than an absolute—not dissimilar to Western concepts of chivalry." However, Mr. Guest also remarks that such similarity also leaves *pukhtunwali* susceptible to issues similar to chivalry, namely "it is subject both to personal interpretation (which can be very creative) and to common abuse." *Id. See also* ANDREA CHIOVENDA, CRAFTING MASCULINE SELVES: CULTURE, WAR, AND PSYCHODYNAMICS IN AFGHANISTAN 41-44, 46, 190 (2020)(discussing two separate concepts of honor in Afghan Pashtun male culture, particularly, *izzat* (masculine honor requiring revenge for insults) and *namus* (honor which demands modesty)).

[217] *See* FM 6-27, *supra* note 168, at ¶ 1-31.

[218] *Id.*

[219] *Id.* at ¶¶ 1-18 to 1-21.

[220] *Id.* at Glossary-3. The most treatment that chivalry gets from FM 6-27 is an equation to "honor" (in fact, the next sentence in the definition is "[a]lso called chivalry."). However, FM 6-27 does not treat chivalry as a separate concept either in definition or in consequence.

[221] *Id.* at para. 1-32.

[222] *Id.* at para. 1-21.

principle "require [sic] that parties accept . . . that certain legal limits exist."[223]

2. Good Faith

As for "good faith" in relation to the laws of armed conflict, sources today apply requirements and expectations of good faith almost as broadly as sources of yesteryear. As early as the sixteenth century, scholars on the laws of war such as the Dutch military jurist Balthazar Ayala observed that throughout history "there was no grander or more sacred matter in human life than good faith."[224] The 1863 Lieber Code instructed federal armies in the American Civil War that deception was permissible so long as it "does not involve the breaking of good faith either positively pledged . . . or supposed by the modern law of war to exist."[225]

In the twentieth century, good faith gained new *ius ad bellum* purchase as the post-World War II global order ardently embraced international law. The United Nations Charter now demands that Member States "shall fulfil in good faith the obligations assumed by them."[226] The Vienna Convention on the Law of Treaties provides at Article 26 that "[e]very treaty in force is binding upon the parties to it and must be performed by them in good faith."[227] In a *ius in bello* context, while the Hague and Geneva Conventions do not expressly use the term, the API Commentary shows that good faith often featured in the Additional Protocol's underlying philosophies, making pronouncements about the "rules on good faith"[228] and that these same rules "prohibit killing or wounding the enemy treacherously, as well as deceiving him by the improper use of the flag of truce, of national emblems or of enemy uniforms, and also by the improper use of the red cross emblem."[229] The Commentary even makes sure to stress that the obligation to think with good faith when engaged in armed conflict does not just sit with the lawyers "but is also imposed on those who enjoy a

---

[223] *Id.* at para. 1-32.
[224] Watts, *supra* note 171 at 174-75 (citing BALTHAZAR AYALA, 2 THREE BOOKS ON THE LAW OF WAR AND ON THE DUTIES CONNECTED WITH WAR AND ON MILITARY DISCIPLINE 55 (John P. Bate trans., 1912) (1582)).
[225] *Lieber Code*, *supra* note 180, at art. 15.
[226] U.N. Charter art. 2, ¶ 2.
[227] *See* Vienna Convention, *supra* note 201, at art. 26.
[228] API Commentary, *supra* note 179, at 382.
[229] *Id.* This is of course notwithstanding the objections made by the United States to the related provisions in Article 39 discussed *supra*.

certain degree of freedom of action in the field, even though the heat of battle does not favour an objective view of things."[230]

Today, both the DoD Law of War Manual and FM 6-27 make good faith a distinct concept in the United States Armed Forces, with FM 6-27 particularly declaring that "absolute good faith" is an essential component of armed conflict and its violation garners separate consequences.[231] Furthermore, these regulations impose expectations of good faith in several aspects of combat from assessing whether a person or object is a lawful target[232] to making agreements for the removal of vulnerable populations during a siege,[233] implying that good faith is a concept essential not only to actions done while interacting with the external enemy but also to decision-making situations requiring internal honesty.

V. ENFORCING THE LAWS ON DECEPTION IN ARMED CONFLICT

Consequences can vary widely for an actor's violation of the laws and principles related to deception in armed conflict. Distinctions arise not just in what kind of deception is used but how it is used, for what purpose, where, and by whom, the final being the hardest question to answer due to attribution challenges.

A. Perfidy – Grave, Prohibited, and Simple

Perfidy is the act with the most immediate severity and consequences. As United States Military Academy Professor Sean Watts explains, "[p]erfidy and treachery are among the gravest law-of-war violations . . . perfidy and treachery provoke draconian and irreversible reactions."[234] Amassing an impressive survey of perfidy from its treatment over the centuries, Professor Watts correspondingly articulates three kinds of perfidy in existence today which have different roots and different enforceability. The first is simple perfidy, described as "all acts" that falsely invite an enemy to provide law-of-war protections and then betray that confidence.[235] The second is prohibited perfidy,

---

[230] *Id.*
[231] FM 6-27, *supra* note 168, at ¶¶ 2-146, 2-147, 2-148, and 2-149. See discussion on consequences *infra*.
[232] DoD Law of War Manual, *supra* note 146 at ¶ 11.18.2.1; FM 6-27, *supra* note 168, at ¶ 2-17.
[233] FM 6-27, *supra* note 168, at ¶ 2-102.
[234] Watts, *supra* note 171, at 106.
[235] *Id.* at 154.

described as perfidious acts that "result in death, injury, or capture of the betrayed enemy."[236] The third is grave perfidy, described as acts of prohibited perfidy that willfully use the recognized emblems under the Geneva Conventions such as the Red Cross or Red Crescent against persons protected under the Geneva Conventions.[237]

The source of simple perfidy in this interpretation appears to be customary international law (described here as "broad custom")[238] that informed notions of honor and prohibited corresponding acts of treachery, and which still remains the only source of international law to prohibit such acts when not covered by written laws. The source of prohibited perfidy, by contrast, is Article 37 of Additional Protocol I with its distinct consequence limitations. The source of grave perfidy is equally concrete, this time originating from Article 85(3)(f) of Additional Protocol I which declares the "perfidious use, in violation of Article 37" of recognized emblems or other protective signs as "grave breaches."[239] The enforcement mechanisms for prohibited perfidy and grave perfidy, however, are not equally concrete, and enforcement mechanisms for simple perfidy are difficult to define.

1. Grave Perfidy

Grave perfidy enjoys the largest degree of certainty. By declaring this very specific vein of Article 37 perfidy a "grave breach," states parties must automatically promulgate domestic legislation in accordance with the grave-breaches provisions in all four 1949 Geneva Conventions to enact "effective penal sanctions" to repress grave perfidy.[240]

---

[236] *Id.*
[237] *Id.*
[238] *Id.*
[239] Additional Protocol I, *supra* note 174, at art. 85(3)(f). While Article 85 declares the acts listed under section 3 to be grave breaches when "causing death or serious injury to body or heath," Professor Watts posits that because subsection (3)(f) explicitly nests into Article 37, even here only acts which misuse recognized emblems in order to cause "killing, injury, or capture" would constitute grave perfidy. Watts, *supra* note 171, at 153.
[240] Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field art. 49, Aug. 12, 1949, 6 U.S.T. 3114, 75 U.N.T.S. 31 [hereinafter Geneva Convention I]; Geneva Convention for the Amelioration of the Condition of the Wounded, Sick, and Shipwrecked Members of Armed Forces at Sea, art. 50, Aug. 12, 1949, 6 U.S.T. 3217, 75 U.N.T.S. 85 [hereinafter Geneva Convention II]; Geneva Convention Relative to the Treatment of Prisoners of War art. 129, Aug. 12, 1949, 6 U.S.T. 3316, 75 U.N.T.S. 135 [hereinafter Geneva Convention III]; Geneva Convention Relative to the Protection of Civilian Persons in Time of War art. 146, Aug. 12, 1949, 6 U.S.T. 3516, 75 U.N.T.S. 287 [hereinafter Geneva Convention IV].

The "grave breach" designation also triggers several other effects in API. Article 85 declares that grave breaches "shall be regarded as war crimes,"[241] Article 86 requires states parties to "repress" grave breaches[242], and Article 87 requires state parties both to direct military commanders under their control "to prevent and, where necessary, to suppress and to report . . . breaches of the Conventions and of this Protocol"[243] and to require any commander who is aware that a breach has or will occur "where appropriate, to initiate disciplinary or penal action against violators thereof."[244]

Furthermore, Article 88 encourages maximum cooperation in criminal proceedings between aggrieved state parties.[245] Article 90 International Fact-Finding Commissions can investigate grave breaches independent of state party efforts to investigate.[246] Finally, Article 91 provides a minimum requirement for a "[p]arty to the conflict which violates the provisions of the Conventions or of this Protocol" to "pay compensation" (although API does not discuss the method, amount, currency determination, or means for deciding compensation disputes).[247]

Additionally, grave breaches of the Geneva Conventions as well as API are subject to universal jurisdiction.[248] While observers and jurists have debated the actual extent of this jurisdiction,[249] the fact remains

---

[241] Additional Protocol I, *supra* note 174, at art. 85(5). This provision is notable because the Convention for the 1949 Geneva Conventions purposefully did not equate grave breaches (a novel term at the time) to war crimes. The Conference felt the term "crimes" had too many different meanings and so sought to avoid it. *See* Gary D. Solis, INTRODUCTION TO GENEVA CONVENTIONS 25 (Kaplan Publishing, 2010).
[242] Additional Protocol I, *supra* note 174, at art. 86(1).
[243] *Id.* at art. 87(1).
[244] *Id.* at art. 87(3).
[245] *Id.* at art. 88(3). This subsection provides that "the law of the High Contracting Party requested shall apply in all cases." *Id.* How this choice-of-law decision would resolve would likely revolve around political considerations, although legal considerations could certainly be determinative if, for example, a State could not muster the resources to conduct prosecutions because armed conflict had crippled its law enforcement infrastructure.
[246] *Id.* at art 90(2)(c)(i).
[247] *Id.* at art. 91.
[248] Universal jurisdiction comes from the 1949 Geneva Conventions' demand that the States Party are "under the obligation to search" for people accused of committing grave breaches and "shall bring such persons, regardless of their nationality, before its own courts." They may also exercise, through the principle of *aut dedere aut judicare*, the option to extradite the accused to the custody of another state party for trial. Geneva Convention I, *supra* note 240, at art. 49; Geneva Convention II, *supra* note 240, at art. 50; Geneva Convention III, *supra* note 240, at art. 129; Geneva Convention IV, *supra* note 240, at art. 146.
[249] *Compare, e.g.*, Roger O'Keefe, *The Grave Breaches Regime and Universal Jurisdiction*, 7 J. INT'L CRIM. J. 811 (2009) (arguing that universal jurisdiction only

that this jurisdictional component is unique to grave perfidy compared to the other two varieties discussed here.

2. Prohibited Perfidy

Prohibited perfidy, by comparison, enjoys only a portion of the codified support necessary for a state to embark on a prosecution or a related legal sanction. The biggest substantive distinction is that while prohibited perfidy captures the *actus reus* qualifications under Article 37 of API, it does not concern the misuse of recognized emblems which when paired with a perfidious act would elevate the crimes to the grave-breaches threshold. Because prohibited perfidy here does not rise to the level of a grave breach, prohibited perfidy does not automatically qualify as a "war crime" under API.[250] None of the States Party have to criminalize it distinctively as "grave breaches" in their domestic criminal codes. They are also not under an affirmative obligation to "repress" prohibited perfidy and none of the States Party are required to bring anyone to trial for committing prohibited perfidy.[251] Finally, universal jurisdiction does not apply to prohibited perfidy, meaning that a state not party to the conflict would have no unilateral ability[252] to prosecute an actor who the state felt committed perfidy (albeit not constituting a grave breach) under Article 37—an omission that can have real-world impacts on efforts to prosecute perfidious uses of deepfake technology.

On the other hand, states party to the API still have an affirmative obligation to assure under Article 87 that their military commanders understand and execute their duties to prevent, suppress, report, and,

---

applies in those circumstances when no other country has made a proper claim to jurisdiction), *with* Arrest Warrant of 11 Apr. 2000 (Democratic Republic of the Congo v. Belgium), Judgment, 2002 I.C.J. 1, 24-25, § 59 (Feb. 14) (finding that universal jurisdiction and the "customary international law" which informs it permits one country's court to have jurisdiction to issue arrest warrants and another country's court to have jurisdiction to afford immunity).

[250] *See supra* note 239 and discussion.

[251] *See* API Commentary, *supra* note 179, at 159 (providing "Although the Parties to the conflict are under the obligation to take measures necessary for the suppression of all acts contrary to the provisions of the Conventions and Protocol I, they are only bound to bring to court persons having committed grave breaches of these treaties . . .").

[252] This presumes no other treaties—bilateral or multilateral—exist at the time which would provide said third-party State with jurisdiction. Additionally, Jean Pictet reasoned in the API Commentary that customary law supports the application of universal jurisdiction to "serious violations of the laws of war" regardless of whether they qualify as grave breaches. API Commentary, *supra* note 179, at 1011. This position, however, may not reflect actual customary international law or even a consensus among States Party. *See, e.g.,* Matheson, *supra* note 173.

"where appropriate, to initiate disciplinary or penal action" against acts which violate the Geneva Conventions, to include prohibited perfidy.[253] Additionally, the Geneva Conventions impose on states party the particular obligation to take domestic legal measures to prevent and repress "at all times" acts by any entity, public or private, that make unlawful use of recognized emblems.[254] The call for investigatory cooperation in Article 88 also applies[255] as may also the requirement to cooperate with the United Nations during investigations of "serious violations" under Article 89.[256] While the Article 90 International Fact-Finding Commission does not have unilateral authority to investigate non-grave breaches, it may still conduct inquiries into "other situations" so long as both parties to the conflict consent to the investigation.[257] Finally, the minimum penalty under Article 91 still applies as well.[258] While the United States does not believe that Articles 90 and 91 reflect customary international law and is unlikely to enforce them, the United

---

[253] Additional Protocol I, *supra* note 174, at art. 87(1), (3).
[254] *See, e.g.*, Geneva Convention I, *supra* note 240, at arts. 53-54.
[255] *See supra* note 245 and discussion.
[256] Additional Protocol I, *supra* note 174, at art. 89. API does not define the term "serious violations" which is purposefully distinct from "grave breach" terminology. The Commentary provides that the Conference initially intended this section to address reprisals in order to avoid breaches being answered by more breaches. However, the revision process neutered that intent and resulted in a broad article simply requiring cooperation with the United Nations. The Commentary says that "[t]he terms 'violation' and 'breach' may be considered to be synonymous." The Commentary, though, does not equate "serious violations" with "grave breaches" which it acknowledges the Conference purposefully made distinct from all other violations. The Commentary states flippantly "[w]e do not need to have in mind exactly what conduct could fall under this definition" in order to avoid proposing a definition. Instead, it posits three categories of acts which would equate to a "serious violation": (1) non-grave isolated acts "of a serious nature," (2) non-grave acts which because of frequency or other circumstances "takes on a serious nature," and (3) "'global' violations" described as "acts whereby a particular situation, territory or a whole category of persons or objects is withdrawn from the application of the Conventions of the Protocol." API Commentary, *supra* note 179, at 1032-33. Research does not show any application of this three-tier definition of "serious violations." Instead, practice appears to show that declaring an act to be a serious violation can be more by feel and circumstance than adherence to a rigid definition. Furthermore, the finding of an act to constitute a serious violation does not garner any more resources or heightened sanctions under the Geneva Conventions or Additional Protocol I than any other non-grave violations. *See, e.g.*, Tadić, *supra* note 25, at ¶¶ 90-91 (finding "It is therefore appropriate to take the expression 'violations of the laws or customs of war' [found in Article 3 of the ICTY Statute] to cover serious violations of international humanitarian law" and that the intent to hitch "serious violations" to the broader concept of violations of laws or customs of war in the ICTY Statute (itself drafted very closely to the Geneva Conventions and Additional Protocol I) was to make the Tribunal's jurisdiction "watertight and inescapable.").
[257] Additional Protocol I, *supra* note 174, at art. 90(2)(d).
[258] *Id.* at art. 91.

States has expressed the intent to require "[a]ppropriate authorities" to take "all reasonable measures to prevent acts contrary to the applicable rules of humanitarian law," "to bring to justice all persons who have willfully committed such acts," and "to cooperate" with other States Party in related proceedings.[259]

3. Simple Perfidy

By comparison, the enforcement mechanism for prohibiting so-called simple perfidy would be unpredictable. Some acts may fall within prohibitions on the misuse of recognized emblems, such as in Article 53 of Geneva Convention I,[260] but not involve any killing, injury, or capture—acts described by Jean Pictet as "prohibited ruse."[261]

For example, some acts may deliberately make an adversary falsely believe they have law-of-war protections but not involve either the misuse of a recognized emblem or a killing, injury, or capture. This is aplausible scenario should deepfake technology proliferate in combat, for example, to convince a belligerent to send supplies to an adversary or to waste instead of to the intended recipient. Other acts may engage in the seemingly perfidious behavior but have no other intended and actual effect than to sow confusion and distrust—also equally plausible as a utility for deepfake. So long as an act of deception can invoke some portion of the Geneva Conventions or the Additional Protocols, these instruments can provide some mechanism to repress and punish those actions similar to the above discussion on prohibited perfidy. Where they do not invoke either document, however, the alleged simple perfidy is likely to blend into a correspondingly simple notion of violating honor—and encounter corresponding repression challenges.

---

[259] *See* Matheson, *supra* note 173, at 428.
[260] Geneva Convention I, *supra* note 240, at art. 53.
[261] API Commentary, *supra* note 179, at 441, 443. Mr. Pictet argues here that "a distinction should be made between a ruse, a prohibited ruse, and an act of perfidy," with a prohibited ruse constituting primarily those acts of deception which unlawfully employ recognized emblems but do not meet the requirements of Article 37 to constitute perfidy. Mr. Pictet goes on to surmise that "prohibited ruse" could also theoretically apply to acts involving delayed-action weapons such as mines and certain booby-traps. However, the extent to which international humanitarian law has adopted this suggestion is not clear. Notably, the 1999 Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction ("Ottawa Treaty") contains no reference to concepts of ruse, perfidy, or deception in general.

B. Violations of Honor and the Problem with Treachery

Violations of honor as described above are difficult to legally articulate, especially under international humanitarian law, particularly because they are not, by definition, perfidious. They can be offensive, cause tremendous confusion, inflict irreparable damage to property, and themselves also make peace harder to establish. But they do not enjoy express positive prohibition in the modern laws of armed conflict. As Professor Watts has observed, "claims to a complementary, broad-based perfidy prohibition derived from notions and principles of chivalry and honor are overstated. Such claims seem grounded in little more than nostalgia, hardly worthy of legal recognition."[262] However, these claims are more easily addressed in domestic laws and codes.

Today chivalry is, from an international humanitarian law perspective, dead letter.[263] Some modern efforts have attempted to re-define the legal notion of chivalry,[264] but the concept is instead often wrapped into discussions on honor and good faith.

Violations of honor and good faith, by comparison, enjoy more robust treatment in international law. The API Commentary, for example, notes particularly that "[Articles 37-39 of API] appeal to the good faith of the combatant which is a fundamental condition for the existence of law."[265] However, neither the Hague nor the Geneva Conventions, nor their Protocols define deceptive actions that constitute explicit violations of "honor" or "good faith." Instead, the laws offer general expressions that States Party must abide by their obligations honorably and/or in good faith[266] and that states remain bound to "principles of the law of nations" (presumably including principles of honor and good faith) as derived from "the laws of humanity and the

---

[262] Watts, *supra* note 171, at 174.
[263] *Id.* at 160 (observing "Chivalry as a principle . . . would be unlikely to actually regulate the conduct of hostilities or form a reliable basis for law-of-war enforcement efforts such as criminal prosecution.").
[264] *See, e.g.*, Evan J. Wallach, *Pray Fire First Gentlemen of France: Has 21st Century Chivalry Been Subsumed by Humanitarian Law?*, 3 HARV. NAT'L SEC. J. 431, 443-60 (2012) (seeking to define modern chivalry by the concepts of courage, trustworthiness, mercy, loyalty, and courtesy; also argues that "[c]hivalry mandates actions and punishes inaction that IHL can only recommend.").
[265] API Commentary, *supra* note 179, at 473.
[266] 1907 Hague Convention IV, *supra* note 171, at pmbl.(commending the instrument to, *inter alia*, the "dictates" of the public conscience); Additional Protocol I, *supra* note 174, at art. I(1) (requiring that States Party "respect" the Protocol "in all circumstances.").

dictates of public conscience" even if they try to withdraw from or denounce the Conventions.[267]

This is not to say that these aspirations are not important or do not present jeopardy for a potential violator. The latter aspirations particularly, reflective of the famous "Martens Clause" found in the preamble of the 1899 Hague Convention[268] and further codified in the Geneva traditions,[269] make it plain that states remain bound to customary international law even if they try to withdraw from international multilateral treaties and will remain stubbornly so especially when these treaties reflect customary international law. However, the fact that armies of scholars and international jurists have proclaimed that honor and good faith are central components of the laws of armed conflict does not guarantee that a perpetrator who employs deepfake technology in odious but not perfidious ways during armed conflict can easily face trial.

This does not mean, however, that an actor hoping to employ deepfake technology in such a manner does so free of any consequences. Uses of deepfake technology could make the actor a lawful target for non-lethal and even potentially lethal force. Consider, for example, a scenario in which a non-state actor in a Common Article 3 non-international armed conflict[270] has crafted a successful deepfake campaign which has contributed significantly to losses of vital war-fighting materiel for an opposing state force. The opposing state force has through various

---

[267] *See, e.g.*, Geneva Convention I, *supra* note 240, at art. 63; Geneva Convention II, *supra* note 240, at art. 62; Geneva Convention III, *supra* note 240, at art. 142; Geneva Convention IV, *supra* note 240, at art. 158; Additional Protocol I, *supra* note 174, at art. 1(2); 1907 Hague Convention IV, *supra* note 171, at pmbl.

[268] 1899 Hague Convention II, *supra* note 172, at pmbl. (declaring "[u]ntil a more complete code of the laws of war is issued, the High Contracting Parties think it right to declare that in cases not included in the Regulations adopted by them, populations and belligerents remain under the protection and empire of the principles of international law, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of the public conscience"). This provision, advocated for by Russian delegate Friedrich Martens at the Conferences to the 1899 Hague Conventions, was a compromise intended to keep disagreements about the Conventions' applicability and enforceability from scuttling the Conventions' creation. Today the Clause itself has received recognition and enforcement in the highest international forum. *See, e.g.*, Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, 257 (Jul. 8) (describing the Martens Clause as "an effective means of addressing the rapid evolution of military technology" and proclaiming that the Clause's "existence and applicability is not to be doubted".)

[269] Geneva Convention I *supra* note 240, at art. 63.

[270] Recall that Additional Protocol I and its prohibited perfidy could not apply here because Additional Protocol I only governs international armed conflicts. *See* discussion *supra* note 174; *see also* discussion about Tadić factors *supra* note 25.

means been able to positively identify the actor at a location which is in an area of active hostilities. The opposing state force believes that the actor's skill in deploying deepfake derived deceptions poses a threat to their lawful military objectives and to the personal safety of their own forces. If the opposing state force can correctly conclude that the actor is directly participating in hostilities, the opposing state force would be within its right under the laws of armed conflict to take lethal action against the actor.[271]

The next level of potential consequence would be prosecution in the opposing state force's domestic criminal system. If the actor, seized in a raid, were detained by the opposing state force, the opposing state force could prosecute the actor in a regularly constituted court[272] under domestic laws which assert personal jurisdiction over the actor. For example, if the opposing state force were the United States—which has a well-established (albeit it controversial) practice of employing military tribunals to try violations of the laws of war[273]— the actor could face trial by a United States military tribunal or even a general court-martial by way of Rule for Court-Martial 201 which applies personal jurisdiction over "any person" who "is subject to trial by military tribunal for any crime or offense against the law of war . . ."[274] In such a case, the actor faces severe legal jeopardy as the Uniform Code of Military Justice authorizes various punishments including the death penalty in cases where violations of the laws of war result in death.[275]

---

[271] This presumes that other non-lethal means, such as conducting a reciprocal malicious cyberattack against the actor's computer or servers or even a raid to arrest the actor, are not feasible. At any rate, if the actor is directly participating in hostilities at the time the actor is observed for targeting, the opposing state force would have no legal obligation to pursue non-lethal means first.

[272] *See* Geneva Convention I, *supra* note 240, at art. 3(2).

[273] *See, e.g*,. Ex parte Quirin, 317 U.S. 1, 45-46 (1942) (upholding the trial of German saboteurs by a U.S. military commission for violations of the laws of war); Hamdi v. Rumsfeld, 542 U.S. 507, 537 (2004) (plurality opinion) (observing that an enemy combatant detainee could be prosecuted by and have habeas corpus petitions entertained by a "properly constituted military tribunal"); for a perspective skeptical of the notion of using U.S. military tribunals to prosecute enemy combatant detainees, *see* Michael R. Belknap, *Alarm Bells from the Past: The Troubling History of American Military Commissions*, 28 J. SUP. CT. HIST. 300 (2003).

[274] MANUAL FOR COURTS-MARTIAL, UNITED STATES, R.C.M. 201(f)(1)(B)(i)(a) (2019) [hereinafter MCM]. This same subsection of R.C.M. 201 also declares that a general court-martial in such a case "may adjudge any punishment permitted by the law of war." *Id.* at R.C.M. 201(f)(1)(B)(ii); *see also id.* at R.C.M. 1003(d)(10)(explaining a general court-martial may, in cases tried under the law of war, adjudge any punishment "not prohibited by the law of war."); UCMJ art. 18.

[275] *See, e.g*., UCMJ art. 81(a)(b) (discussing the potential application of the death penalty in the case of a conspiracy to violate the laws of war that results in death).

Ultimately, whether and to what extent an actor may face prosecution for violations of honor in relation to a use of deepfake technology would be heavily fact dependent. Whether the forum is a U.S.-style military commission or court-martial, a domestic court, or even in an *ad hoc* international tribunal, if the governing code or tribunal charter does not carefully account for the distinctions discussed here then it will set its prosecutors up to fail.[276]

C. An Argument in Favor of Deepfakes: Lawful Ruse

Although much of this article has discussed circumstances in which deepfake manipulation would violate the laws of armed conflict, it is equally important to acknowledge that the employment of deepfake-based deception is not, by itself, illegal. Just like any other medium of deception, deepfake technology is not *per se* banned from war.

Deepfake deception can be as perfectly lawful a utility during the conduct of military operations as many acts of deception have been throughout history. For example, a belligerent could use deepfake-derived content to make an enemy think an attack was occurring on one outpost in order to create a distraction allowing the belligerent to attack a different outpost. In order to thwart an attack, a besieged belligerent could fake its numbers by broadcasting deepfake videos seeming to show hundreds of defenders at a base when in reality the base may only have a couple dozen defenders. As discussed below, even the deepfake

---

[276] A classic example of the folly inherent in trying to prosecute violations of honor without a concrete understanding of the offense occurred during the proceedings of the 1946 International Military Tribunal for the Far East (a.k.a. the Tokyo War Crimes Tribunal). There, prosecutors charged the Japanese defendants with, *inter alia*, violating Article 23(b) of the 1907 Hague Convention by committing Article 23(b) treachery which allegedly occurred when Japan attacked the United States at Pearl Harbor. The prosecutors and the Tribunal both failed to understand the fundamental divide in international law between *ius ad bellum* and *ius in bello*. The prosecutors confused *ius in bello* treachery under Article 23(b)—which would occur when the deception works to affect a hostile act while engaged in combat—with the *ius ad bellum* facts charged i.e., that Japan had engaged in diplomatic deception to affect a hostile act in furtherance of securing an advantage in a war that had not yet come, which the Hague Regulations are powerless to regulate. While the Tribunal did not rule against the prosecutors because of their erroneous charge, the result was the same—the Tribunal did not convict the defendants, applying a *ius in bello*-style rationale that the United States was in possession of too much information about Japan's intentions before the attack for the bombing of Pearl Harbor to constitute a violation of Article 23(b). *See* Watts, *supra* note 171, at 141 (citing NEIL BOISTER & ROBERT CRYER, THE TOKYO INTERNATIONAL MILITARY TRIBUNAL: A REAPPRAISAL 171 (2008)). Had the prosecutors and the Tribunal understood that they needed to apply *ius ad bellum* law to the *ius ad bellum* facts before them, the Tribunal's ruling on the Pearl Harbor attack may have been different.

manipulation of an enemy's satellite-based geo-spatial imagery could be done lawfully as part of a ruse. Synthetic content created and delivered by AI could support a "feint"[277] to deceive an adversary as to the time or place of a knockout assault, thereby winning a war or causing a belligerent to lose one.

To a military theorist, perhaps deepfake's most effective use would be to infiltrate an opponent's OODA Loop. The OODA Loop is the Observe-Orient-Decide-Act cyclic chain pioneered by the late U.S. Air Force Colonel (Ret.) John Boyd to describe the means to out-think, out-maneuver, and overwhelm an opponent's mental processing abilities and defeat them by getting "inside [their] decision cycle."[278] This occurs by using speed and unpredictability to create confusion in the enemy so severely that the enemy loses the mental ability to take in information and react in time to avoid losing. As Colonel Boyd's biographer described the effect, "the losing side rarely understands what happened."[279] A deepfake information and cyber campaign powered by algorithms designed precisely to hijack an opponent's OODA Loop could do just that with historic efficiency—and without legal ramification.

So long as these acts do not take advantage of or cause distrust in the protections under international law in order to achieve their objectives, and do not cause the enemy to unknowingly harm protected people or places, international law does not prohibit them. Such uses of deepfake technology more likely require political or military options, not legal recourse.

VI. CHALLENGES OF DEEPFAKE TECHNOLOGY ON PRESENT AND FUTURE CONFLICTS

A. Democratization

Although deepfake technology is still young, it has evolved quickly. The learning curve, which at first appeared too steep for most to

---

[277] *See* JP 3-13.4, *supra* note 170, at para. 11(c)(1).

[278] Colonel (Ret.) Boyd did not write a book or an article when creating the OODA Loop or its underlying concepts but instead featured them in a slide deck entitled "Patterns of Conflict" which he briefed to military leadership for decades. *See* John Boyd, Patterns of Conflict (Dec. 1986) (available at http://www.ausairpower.net/JRB/poc.pdf). The quote here, while often stated by Col. (Ret.) Boyd as a goal of the OODA Loop concept, actually comes from U.S. Army General Colin Powell as he described how coalition forces were able to secure a sweeping victory during Operation Desert Storm. ROBERT CORAM, BOYD: THE FIGHTER PILOT WHO CHANGED THE ART OF WAR 425 (2004).

[279] Coram, *supra* note 278, at 334.

handle, is barely visible today. Several programs now exist for creating deepfake content that anyone can buy. Applications such as Reface,[280] DeepFaceLab,[281] Descript,[282] and ZAO[283] which can produce high-quality deepfake content in under an hour, are widely accessible. Additionally, where previously a person would need some degree of training or programming experience to use these applications, YouTube now has several videos which seek to train people to create deepfakes using these applications, sometimes in under an often-claimed "10 minutes."[284]

The fruit of the feverish labor to democratize deepfake technology is, like the nature of the internet itself today, both entertaining and hazardous. Certainly, YouTube abounds with deepfake images composed for benign purposes such as depicting a *Star Wars* movie recast with a different actor or for satirical purposes.[285] However, the hazards of deepfake technology, which asserted themselves from the beginning, have evolved beyond the scatological.

Even before the Zelenskyy deepfake, actors had already used artificial intelligence to create synthetic content of world leaders for political and social purposes. For example, a January 2020 video by Alethea Group purports to show U.S. President Donald Trump and

---

[280] REFACE, https://play.google.com/store/apps/details?id=video.reface.app&hl=en_US&gl=US (last visited Feb. 28, 2021).

[281] *See* Ivan Perov, et al., *DeepFaceLab: A Simple, Flexible, and Extensible Face Swapping Framework* (May 12, 2020), https://arxiv.org/abs/2005.05535 (boasting that DeepFaceLab, an open-source deep-fake system, allows users to modify content "to achieve their customization purpose . . . with high fidelity and indeed indiscernible by mainstream forgery detection approaches . . .").

[282] DESCRIPT, https://www.descript.com/overdub (last visited Feb. 28, 2021).

[283] When ZAO became available on China's iOS App Store, it became China's most downloaded app overnight. *See, e.g.*, Zak Doffman, *Chinese Deepfake App ZAO Goes Viral, Privacy of Millions 'At Risk'*, FORBES (Sep. 2, 2019, 4:27 AM), https://www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-deepfake-app-zao-sparks-huge-faceapp-like-privacy-storm/?sh=3a391bf84700. Considerable controversy ensued when ZAO's privacy policy turned out to allow the Chinese government to retrieve data input through ZAO. *Id.*; *see also* Laura He, Jack Guy, & Serenitie Wang, *New Chinese "Deepfake" Face App Backpedals After Privacy Backlash*, CNN (Sep. 3, 2019, 6:33 AM), https://www.cnn.com/2019/09/03/tech/zao-app-deepfake-scli-intl/index.html.

[284] *See, e.g.,* Tom Baranowicz, *How to Make DeepFake in 10 Mins – Tutorial* (Aug. 12, 2020), https://www.youtube.com/watch?v=eq55Qy4RPiA; Amrit Aryal, *Create Deepfakes with Just One Picture in Under 10 Minutes* (Oct. 31, 2020), https://www.youtube.com/watch?v=TY2DEP-C-O4.

[285] *See, e.g.,* Shamook, *Harrison Ford in Solo: A Star Wars Story [DeepFake]*, YOUTUBE (Aug. 16, 2020), https://www.youtube.com/watch?v=bC3uH4Xw4Xo.

British Prime Minister Boris Johnson, among others, admitting they were wrong about denying climate change.[286]

Some videos like these can be discredited almost instantaneously because they depict globally-known figures. The January 2020 climate change deepfake completely contradicted the politicians' long-held and well-known positions as well as their own personalities. As a result, the content had no likelihood of convincing anyone that the 'speakers' had suddenly changed their views just minutes after delivering remarks to the contrary. They were easily and naturally identifiable as fake. In fact, the high-profile nature of political life is often the best utility for combating deepfake content depicting high-profile politicians, as the resolution of the 2022 Zelenskyy video incident also proved.[287]

The challenge grows, however, when confronting content that depicts relatively low-profile people, such as tactical-level military commanders, or people who otherwise do not have a large public profile and so the content cannot as quickly be disproven. Furthermore, content that is purposefully incomplete such as voice-only deepfake can make detection difficult and aggravate confusion, especially if transmitted in high-intensity situations.

This is no academic concern. If anyone thinks this technology could not reasonably fool someone into thinking that they are interacting with someone they personally know, much less effect any significant outcome, they should think again. It's already happened.

In 2019, a criminal enterprise used deepfake technology to make a U.K.-based CEO believe he was talking to the Germany-based CEO of his parent company.[288] The AI managed to perfectly mimic the German CEO's voice. As an insurance investigator for the company reported to the Wall Street Journal, the AI replicated the German CEO's "slight German accent" and even the "melody" of the German CEO's cadence.[289] It only took one phone call. The criminals used the AI to make the British CEO believe an emergency was occurring and that the British CEO

---

[286] Alethea Group created and posted the videos shortly after President Trump made comments at the 2020 World Economic Forum in Davos, Switzerland where he denied that the environment was an economic concern. While the quality of the images and audio produced in the faked videos was raw, the timing and swiftness of the videos were still remarkable. CBS News, *President's Words Used to Create "Deepfakes" at Davos*, YOUTUBE (Jan. 24, 2020), https://www.youtube.com/watch?v=4A9LAxhi68I.

[287] *Supra* note 13.

[288] Catherine Stupp, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*, WALL ST. J. (Aug. 30, 2019, 12:52 PM), https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[289] *Id.*

needed to transfer $243,000.00 to a Hungarian supplier's bank account in one hour. The British CEO, skeptical but nonetheless convinced he was talking with his boss, transferred the money. The money, however, went to a bank account in Mexico where it disappeared. A few moments afterwards, so did the criminals—needing mere minutes to succeed.[290]

The victims involved were not traditionally vulnerable. They were not under-resourced, under-educated, or over-leveraged. To the contrary, the AI fooled a European businessman, someone of presumably significant acumen entrusted with co-leading a multinational corporation.[291] Only hubris could argue that a military commander could not be fooled as well and be convinced in part by an AI-derived manipulation to surrender forces or even unknowingly commit a war crime themselves. Because of the democratization of deepfake technology, near-perfect media manipulation capabilities—and the resulting complications they can cause—are within reach of any actor, state or non-state, with a motivation, an internet connection, and some free time.

B. Satellite Imagery Manipulation

Another advent in deepfake proliferation that is growing quickly does not involve depicting people at all—but it is a serious threat to the multi-domain battlespace. GAN-powered manipulation has begun hitting satellite imagery.

The concept is both elegant and nefarious. An actor infiltrates an enemy's satellite link. The actor identifies the geographic area of an enemy's expected operations. The actor then uploads a deepfake-generating program that doesn't make major manipulations, such as wiping out mountains on a digital map, but makes subtle manipulations such as thinning a forest to make an area seem passable or depicting a small bridge over a stream where a bridge in reality does not exist. The satellite link transmits these manipulations throughout the enemy's formations who believe their convoy has a clear route to a waypoint on the other side of the stream. Only when the convoy reaches the stream and sees no bridge does the convoy realize the deception. Then the ambush begins.

---

[290] *Id.*

[291] While media has so far not published the business's name, as a possible sign of the business's robustness, the entire loss was swiftly covered by Euler Hermes Group, a multi-billion-dollar global insurance firm. *Id.*

This is the exact scenario which leaders in defense artificial intelligence development already acknowledge is here.[292] At a Genius Machines summit in 2019, Mr. Todd Myers, automation lead for the CIO-Technology Directorate at the U.S. National Geospatial-Intelligence Agency, publicly acknowledged this capacity and beyond even that, Mr. Myers conceded that "[t]he Chinese are well ahead of [the United States]."[293] Mr. Andrew Hallman, director of the C.I.A.'s Digital Directorate, speaking at the same summit, observed that "[w]e are in an existential battle for truth in the digital domain" and when asked if he felt that the C.I.A. was up to the task of defeating satellite imagery manipulation, responded "I think we are starting to. We are just starting to understand the magnitude of the problem."[294]

This vulnerability presents several problems beyond the one detailed above. The GANs which would manipulate geo-spatial imagery may also adversely influence the machine learning that other neural networks within the satellite are constantly conducting. If those neural networks lack effective classifiers to identify that a tree or a road is fake, they will exacerbate the manipulation by classifying the manipulation as authentic—thus causing allied neural networks to learn errantly and make the problem harder to detect. Also, defenses against infiltration and manipulation would be very expensive, requiring redundancies of all

---

[292] *See* Patrick Tucker, *The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth*, DEF. ONE (Mar. 31, 2019), https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/; *see also* Chunxue Xu & Bo Zhao, *Satellite Image Spoofing: Creating Remote Sensing Dataset with Generative Adversarial Networks*, Article No. 67, p. 1-6 (Jun. 10, 2018) (10th International Conference on Geographic Information Science paper), https://drops.dagstuhl.de/opus/volltexte/2018/9395/pdf/LIPIcs-GISCIENCE-2018-67.pdf (examining how satellite image manipulation works through the use of CycleGAN and Pix2Pix networks and advocating their use for urban planning purposes).

[293] *Id.* Russia may also be experienced in using AI to spoof navigation technology. In late June 2021 Russian naval and air forces stationed in the contested Crimean peninsula scrambled to confront UK and Dutch warships transiting the Black Sea on the basis of data it claimed showed the ships threatening Russian-claimed economic exclusion zones near the Crimean peninsula. The UK and Dutch ships and crews denied being close to the peninsula, arguing that they had been almost 200 nautical miles and 70 nautical miles away, respectively, from the peninsula – not the 12 nautical miles that Russia claimed. Fact-finding later appeared to show that Russia likely had spoofed the warships' radio transponders to give off an incorrect position to justify the subsequent show of Russian force. David Axe, *Harassing Ships and Spoofing Radios, Russia is Telling a Story – That Occupied Crimea is Russian*, FORBES (Jul. 1, 2021 at 8:00 AM), https://www.forbes.com/sites/davidaxe/2021/07/01/harassing-ships-and-spoofing-radios-russia-is-telling-a-story-that-occupied-crimea-is-russian/?sh=782ec8ba414b.

[294] *Id.*

imagery so that if one set is compromised, the compromise can be found by comparing the concerned set to a second set.[295] Furthermore, if a state's armed forces succeed in defending their satellite networks from deepfake manipulation, they would remain defenseless against infiltrations of privately-owned satellite mapping services unless they could enlist the help of the private entities who own the services.[296]

There is no question that this capability could impose mission- and possibly life-threatening costs. However, combating these costs is not just a matter of resource allocation or plan execution. These developments present moral, philosophical, and legal complications that stand to deeply challenge how and even whether actors observe certain facets of the laws of armed conflict.

C. The Liar's Dividend Weaponized, and the Competency Paradox

A unique challenge that AI-driven manipulation creates is the so-called "Liar's Dividend"[297] where someone actually does something or says something but then denies doing so, by falsely claiming that the content depicting the speech or action was a deepfake. All of these complications can impact the battlefield.

Cases of the Liar's Dividend have already impacted Gabon, as well as Sino-Australian relations.[298] Combatants could blatantly attack a civilian population, abuse protected emblems for military gain, execute prisoners of war, or commit a number of other offenses against the laws of armed conflict but take advantage of the Liar's Dividend to argue that even the clearest evidence of these crimes are just deepfake concoctions.

To be sure, bad faith actors can and often do argue regardless of basis that legitimate evidence against them is fraudulent or made-up and have done so long before the invention of deepfake technology. What makes the Liar's Dividend particularly nefarious is that it would arise not in a manicured court of law, where it could be strangled, but in a court of public opinion, where it could thrive and then deflect the trial that might strangle it.

In a court of law, a painstaking digital forensics evidentiary audit and related expert witness testimony, various degrees of corroborating evidence, the unique intensity of focus that trials muster, and procedural

---

[295] *Id.* (quoting Mr. Myers).
[296] *Id.* (detailing concerns about Google Maps or Tesla being infiltrated).
[297] Citron, *supra* note 41 at, 1785-86.
[298] *Supra* notes 20, 44, respectively.

rules that guide evidence presentation, credibility, challenge, ultimate acceptance, and fact-finder consideration could deliver a stout haymaker to a Liar's Dividend-style defense. That haymaker, however, requires an enormous wind-up. This delivery would only come after months if not years pass while the trial comes together. The court of public opinion never provides that time. It arraigns, holds trial, considers evidence, and delivers a verdict all before breakfast.

Furthermore, the Liar's Dividend takes advantage of a competency paradox. The most credible circumstance for a Liar's Dividend defense would be where the falsely accused party is in fact adept at engaging in deception themselves. In other words, the better a belligerent is at using deepfake technology or deception in general, the stronger the Liar's Dividend defense. In turn, as a state becomes more vulnerable to the Liar's Dividend, the actual perpetrator's platform becomes more powerful. The perpetrator can use that platform to make a trial appear unjust or evidence appear untrue.

Russia appears to have recently attempted both of these approaches. For example, in the early phase of its invasion of Ukraine, when its forces attacked from every point of the compass except west and attempted to seize Kyiv, it occupied the town of Bucha a short distance outside of the Ukrainian capital.[299] Ultimately Russian forces failed to take Kyiv and withdrew to focus on an offensive in the east. Almost immediately after Russian soldiers left Bucha, dozens of videos and photographs emerged showing that hundreds of Ukrainian citizens had been executed, many of them bound and tortured before the killing shot.[300]

Instead of launching an investigation or seeking to bring the perpetrators to justice, the Russian government launched a campaign declaring that the videos and photographs were fake.[301] Employing the state-run Russian Telegram (RT) network, Russia aired a piece to its viewers entitled "War on Fakes" which claimed that the images were "staged" by Ukrainian and Western media outlets, attempted to point out inconsistencies in the videos, and portrayed timelines involving the Russian occupation of Bucha to argue that the content was fake.[302] They

---

[299] Cara Anna, *War Crimes Watch: A Devastating Walk Through Bucha's Horror,* ASSOCIATED PRESS (Apr. 10, 2022), https://apnews.com/article/russia-ukraine-europe-war-crimes-7791e247ce7087dddf64a2bbdcc5b888.
[300] *Id.*
[301] Yevgeny Kuklychev, *Fact Check: Russia Claims Massacre in Bucha 'Staged' by Ukraine*, NEWSWEEK (Apr. 4, 2022 at 11:41 AM), https://www.newsweek.com/fact-check-russia-claims-massacre-bucha-staged-ukraine-1694804.
[302] *Id.*

even employed some of the same techniques used in Western media to discredit the Zelenskyy deepfake, labeling images of bodies as "fake" and holding "antifake" panel discussions purporting to inform viewers that they should not believe what they see.[303]

Various Western and Ukrainian media outlets have worked to debunk Russia's campaign, pointing to witness testimonies, drone footages, satellite images, and other means.[304] And while Ukrainian prosecutors have already begun war crimes trials to seek justice for the killings,[305] the victims and their families may have to agonizingly witness justice delayed and possibly denied for the very real crimes the perpetrators commited,[306] especially as prosecutors may need to exert significantly more time and resources to lay the evidentiary foundation for video or photographic evidence than would have been required in another age.

VII. RECOMMENDATIONS FOR IMPROVED GOVERNANCE OF DEEPFAKE

While the invasion of Ukraine has ushered deepfake technology into the records of war, as of the writing of this article, purely by the numbers, the vast majority of problems with deepfake media manipulation remains relegated to the domestic realm. However, like with other inventions such as barbed wire or the airplane that were not born for war but were nonetheless enlisted, there is no reason to believe that AI-derived media manipulation will not be further weaponized. It is important, therefore, to figure out now how to better handle its impact.

First, international agreements seeking to govern artificial intelligence or cyberspace operations in armed conflict must expressly

---

[303] Robert Mackey, *Russian TV is Filled with Images of Bucha's Dead, Stamped with the Word "Fake"*, THE INTERCEPT (Apr. 12, 2022 at 7:51 AM), https://theintercept.com/2022/04/12/bucha-massacre-russia-tv-fake-ukraine-war/.

[304] *Id.*; *see also* Aude Dejaifve, *Fresh Round of Fake Videos Claim the Bucha Massacre was Staged*, FRANCE24 (Jun. 4, 2022 at 6:40 PM), https://observers.france24.com/en/europe/20220408-fresh-round-of-fake-videos-claim-the-bucha-massacre-was-staged; Malachy Browne, *Satellite Images Show Bodies Lay in Bucha for Weeks, Despite Russian Claims*, THE NEW YORK TIMES (Apr. 4, 2022), https://www.nytimes.com/2022/04/04/world/europe/bucha-ukraine-bodies.html.

[305] Victor Jack, *Ukraine Files First War Crimes Charges Against Russia Over Bucha Killings*, Politico (Apr. 28, 2022 at 6:17 PM), https://www.politico.eu/article/ukraine-first-war-crimes-charges-against-russia-over-bucha-killings/.

[306] *See e.g.* Erika Kinetz, *War Crimes Watch: Hard Path to Justice in Bucha, Ukraine, Atrocities*, Frontline (Apr. 4, 2022), https://www.pbs.org/wgbh/frontline/article/bucha-ukraine-civilian-deaths-justice-tribunal-international-criminal-court/ (detailing the myriad difficulties in prosecuting Russian soldiers for the alleged killings).

address the use of artificial intelligence in deception or misinformation activities. These agreements, in whatever form they may take, should acknowledge the reality that AI can create synthetic content that seems to change reality. They should expressly govern such deployment of AI under a regime that criminalizes its use to engage in perfidy, whatever the style.

Second, such instruments should articulate perfidy definitions that not only align with Article 37 of Additional Protocol I but also build upon it. Article 37 has long been derided for being too narrow with its "kill, injure, or capture" limitation.[307] This list should remove consequence requirements all together, and replace them instead with a general intent *mens rea* of intending to secure a military advantage. If the wrongfulness of perfidy is that the abuse of protections afforded under international law will cause a destruction of trust necessary to secure peace, it should not matter whether that sin serves the purpose of killing or the purpose of confusing.

Third, and in assistance with the first two recommendations, U.S. Department of Defense doctrine on perfidy should align with the representations the U.S. government otherwise has made as expressed in the Matheson Memorandum.[308] If the Department of Defense believes it necessary not to acknowledge "capture," because the Department believes customary international law allows a combatant to fake a protected status in order to avoid capture, Article 37 does not conflict with this view. Article 37 only prohibits claiming a protected status in order to commit a capture. Updating this posture will be a net positive for the U.S. as it will foster intra-governmental unity of vision, intra-governmental unity of expectation, communicate to the rest of the world that the U.S. is of the same mind about Article 37, and better assure that its forces do not become subject to behavior that it would likely want to object to if such behavior occurred to U.S. forces.

Fourth, U.S. Department of Defense information operations and artificial intelligence doctrines should expressly address deepfake capabilities, threats, and counters, with corresponding training inserted into Information Operations and LOAC training to signal and military intelligence occupations and to senior leaders regardless of branch or occupation specialty on how to recognize and react to a deepfake ruse. Furthermore, deepfake technology implications should also be trained in

---

[307] *Cf.* U.S. Department of Defense refusal to recognize "capture" as part of customary international law discussed *supra* note 205.
[308] Matheson, *supra* note 173.

concert with the recently released DoD AI Ethics Principles.[309] Training in either context would not need to be overly-detailed – simply enough to apprise commanders and impacted subject matter experts of the issues and what they should and should not do in response.

Fifth and finally, given that deepfake manipulation is unlikely to lose its attractiveness anytime soon, and until counter-deepfake methods reach the same level of productivity as their opponents, the international community, spearheaded by the United States, should embark on a concerted public awareness and education campaign about deepfake technology problems. The best way to combat such sophisticated deception before it can do serious harm may be to just make sure everyone knows it exists and what it can really do. This approached proved itself when media outlets and the Ukrainian government identified and discredited the Zelenskyy deepfake almost as quickly as it was broadcast, with no reported surrenders or slackening in the Ukrainian war effort.[310] The Ukrainian government had even launched a deepfake public awareness campaign two weeks before the Zelenskyy deepfake broadcast, further aiding in the later content's quick debunking.[311] Without an awareness campaign, the resulting skepticism, while not without its own negative social impacts, may present a targeted entity with enough time to uncover the deception before anyone acts in a way that could achieve the deception's objectives.

CONCLUSION

Deepfake technology only promises to gain more traction in the affairs of armed conflict. Experts in artificial intelligence and armed conflict suggest that the technology has already in the short space of a couple years advanced from a first generation to a second-generation capability and that combating it now will require a "whole of society approach."[312]

However, despite its penchant for victimization and its clear potential to cause irreparable harm to notions of trust from the ballot box to the bunker, its growing uses in popular media have already endeared deepfake technology to an entire generation of consumers. These

---

[309] DEF. INNOVATION BD., AI PRINCIPLES: RECOMMENDATIONS ON THE ETHICAL USE OF ARTIFICIAL INTELLIGENCE BY THE DEPARTMENT OF DEFENSE (Oct. 2019).
[310] *Supra* note 13.
[311] Simonite, *supra* note 13.
[312] *See, e.g.,* Brigadier General R. Patrick Huston & Lieutenant Colonel M. Eric Bahm, *Deepfakes 2.0: The New Era of "Truth Decay,"* JUST SEC. (Apr. 14, 2020), https://www.justsecurity.org/69677/deepfakes-2-0-the-new-era-of-truth-decay/.

consumers may understandably cheer the sight of a circa-1980s Luke Skywalker appearing in 2021 *Star Wars* content, gush at the thought of swapping in themselves as the lead in their favorite movie, or adore the technology's capacity to engineer biting political satire. However, the legal community must remain vigilant to help the greater global community continue to always bear in mind that while deepfake technology may have harmless entertainment value in some contexts or even net positive effects in others,[313] as examples from Gabon and Ukraine show, it still bears a capacity for great harm and significant legal instability.

[313] *See, e.g.*, Jessica Silbey & Woodrow Hartzog, *The Upside of Deepfakes*, 78 MD. L. REV. 960, 962-64 (2019) (observing positive effects of deepfake technology such as creating new teaching utilities in education or strengthening journalistic integrity standards).

VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

ARTICLES

THE CENTRALITY OF DATA AND COMPUTE FOR AI INNOVATION: A BLUEPRINT FOR THE NATIONAL RESEARCH CLOUD

Daniel E. Ho *Jennifer King*

Russell C. Wald *Christopher Wan*

THE CENTRALITY OF DATA AND COMPUTE FOR AI INNOVATION: A BLUEPRINT FOR THE NATIONAL RESEARCH CLOUD[1]

Daniel E. Ho[2] *Jennifer King*[3]

Russell C. Wald[4] *Christopher Wan*[5]

---

[1] We thank Simran Arora, Sabina Beleuz, Nathan Calvin, Shushman Choudhury, Drew Edwards, Neel Guha, Tina Huang, Krithika Iyer, Ananya Karthik, Marisa Lowe, Kanishka Narayan, Diego Núñez, Tyler Robbins, Frieda Rong, Jasmine Shao, Sahaana Suri, Sadiki Wiltshire, and Daniel Zhang for their research and written contributions to this paper. We also would like to thank Jeanina Casusi, Celia Clark, Shana Lynch, Kaci Peel, Stacy Peña, Mike Sellitto, Eun Sze, and Michi Turner for their help in preparing this paper for publication. In our process, we also engaged many civil society leaders and advocates who have expressed many perspectives about building a National Research Cloud. We have incorporated their feedback where possible and are grateful for their shared thoughts and for helping us shape a better Section. These individuals include Erik Brynjolfsson, Isabella Chu, Jack Clark, John Etchemendy, Fei-Fei Li, Marc Groman, Eric Horvitz, Sara Jordan, Vince Kellen, Ed Lazowska, Naomi Lefkovitz, Brenda Leong, Amy O'Hara, Wade Shen, Suzanne Talon, Lee Tiedrich, and Evan White. We relied on extraordinary outside expert reviewers for feedback and guidance. We are grateful to Leisel Bogan, Jack Clark, John Etchemendy, Mark Krass, Marietje Schaake, and Christine Tsang for their thoughtful review of the full Section, and thank Isabella Chu, Kathleen Creel, Luciana Herman, Sara Jordan, Vince Kellan, Brenda Leong, Ruth Marinshaw, Amy O'Hara, and Lisa Ouellette for their subject expertise on specific sections. Authors are listed alphabetically and have equally contributed to this work.

[2] **Daniel E. Ho, J.D., Ph.D.,** is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, and Senior Fellow at the Stanford Institute for Economic Policy Research at Stanford University. He directs the Regulation, Evaluation, and Governance Lab (RegLab) at Stanford, and is a Faculty Fellow at the Center for Advanced Study in the Behavioral Sciences and Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). He received his J.D. from Yale Law School and Ph.D. from Harvard University and clerked for Judge Stephen F. Williams on the U.S. Court of Appeals for the District of Columbia Circuit.

[3] **Jennifer King, Ph.D.,** is the Privacy and Data Policy Fellow at the Stanford HAI. She completed her doctorate in information management and systems (information science) at the University of California, Berkeley School of Information. Prior to joining HAI, she was the Director of Consumer Privacy at the Center for Internet and Society at Stanford Law School from 2018 to 2020.

[4] **Russell C. Wald** is the Director of Policy for the Stanford HAI, leading the team that advances HAI's engagement with governments and civil society organizations. Since 2013, he has held various government affair roles representing Stanford University. He is a Term Member with the Council on Foreign Relations, Visiting Fellow with the National Security Institute at George Mason University, and a Partner with the Truman National Security Project. He is a graduate of UCLA.

[5] **Christopher Wan, J.D.,** is a graduate of Stanford Law School and an investor at Bessemer Venture Partners. He was the teaching assistant for the Stanford Policy Practicum: Creating a National Research Cloud and a research assistant for the Stanford HAI. He received his B.S. in computer science from Yale University and

INTRODUCTION

Artificial intelligence (AI) appears poised to transform the economy across sectors ranging from healthcare and finance to retail and education. What some have coined the "Fourth Industrial Revolution"[6] is driven by three key trends: greater availability of data, increases in computing power, and improvements to algorithm design. First, increasingly large amounts of data have fueled the ability for computers to learn, such as by training an algorithmic language model on all of Wikipedia.[7] Second, better computational capacity (often termed "compute") and compute capability have enabled researchers to build models that were unimaginable merely ten years ago, spanning billions of parameters (an exponential increase in scope from previous models).[8] Third, basic innovations in algorithms are helping scientists to drive forward AI, such as the reinforcement learning techniques that enabled a computer to defeat the world champion in the board game Go.[9]

Historically, partnerships between the government, universities, and industries have anchored the U.S. innovation ecosystem. The federal government played a critical role in subsidizing basic research, enabling universities to undertake high-risk research that can take decades to commercialize. This approach catalyzed radar technology, the internet, and GPS devices. As the economists Ben Jones and Larry Summers put it, "[e]ven under very conservative assumptions, it is difficult to find an average return below $4 per $1 spent" on innovation, and the social returns might be closer to $20 for every dollar spent.[10] Industry, in turn, scales and commercializes applications.

---

worked as a software engineer at Facebook and as a venture investor at In-Q-Tel and Tusk Ventures.

[6] *See generally*, KLAUS SCHWAB, THE FOURTH INDUSTRIAL REVOLUTION (2016).

[7] Tae Yano & Moonyoung Kang, Taking Advantage of Wikipedia in Natural Language Processing (Fall 2008) (unpublished term project report), 1-2, https://www.cs.cmu.edu/~taey/pub/wiki.pdf.

[8] *See, e.g.*, Anthony Alford, *Google Trains Two Billion Parameter AI Vision Model,* INFOQ (June 22, 2021), https://www.infoq.com/news/2021/06/google-vision-transformer/; Anthony Alford, *OpenAI Announces GPT-3 AI Language Model with 175 Billion Parameters*, INFOQ (June 2, 2020), https://www.infoq.com/news/2020/06/openai-gpt3-language-model/.

[9] *AlphaGo*, DEEPMIND (2021), https://deepmind.com/research/case-studies/alphago-the-story-so-far/.

[10] Benjamin F. Jones & Lawrence H. Summers, *A Calculation of the Social Returns to Innovation* (Nat'l Bureau of Econ. Rsch., Working Paper No. 27863, 2020); J.G. Tewksbury et al., *Measuring the Societal Benefits of Innovation*, 209 SCI. 658 (1980); *see also* NAT'L ACAD. OF SCI., ENG'G, & MED., RETURNS TO FEDERAL INVESTMENTS IN THE INNOVATION SYSTEM (2017).

A. Challenges to the AI Innovation Ecosystem

Yet this innovation ecosystem faces serious potential challenges. Computing power has become critical for the advancement of AI, but the high cost of compute has placed cutting-edge AI research in a position accessible only to key industry players and a handful of elite universities.[11] Access to data—the raw ingredients used to train most AI models—is increasingly limited to the private sector and large platforms,[12] since government data sources remain largely inaccessible to the AI research community.[13] As the National Security Commission on AI (NSCAI) has determined, "[t]he consolidation of the AI industry threatens U.S. technological competitiveness."[14]

Four interrelated challenges illustrate this finding: first, we are seeing a significant brain drain of researchers departing universities.[15] In 2011, AI PhDs were roughly as likely to go into industry as academia.[16] Ten years later, two-thirds of AI PhDs go into industry, and less than one-quarter go into academia.[17] Second, these trends indicate that many university researchers struggle to engage in cutting-edge science, draining the field of the diverse set of research voices that it needs. Third, the fundamental research that would guarantee the United States stays at the helm of AI innovation is being crowded out. By one estimate, 82 percent of algorithms used today originated from federally funded nonprofits and universities, but "U.S. leadership has faded in recent decades."[18] Fourth, government agencies have faced challenges in building compute infrastructure,[19]

---

[11] STUART ZWEBEN & BETSY BIZOT, COMPUTING RSCH. ASS'N, 2019 TAULBEE SURVEY: TOTAL UNDERGRAD CS ENROLLMENT RISES AGAIN, BUT WITH FEWER NEW MAJORS; DOCTORAL DEGREE PRODUCTION RECOVERS FROM LAST YEAR'S DIP (2019).
[12] Jathan Sadowski, *When Data is Capital: Datafication, Accumulation, and Extraction*, 2019 BIG DATA & SOC'Y 1 (2019).
[13] Amy O'Hara & Carla Medalia, *Data Sharing in the Federal Statistical System: Impediments and Possibilities*, 675 ANNALS AM. ACAD. POL. & SOC. SCI. 138, 140-41 (2018).
[14] NAT'L SEC. COMM'N ON A.I., FINAL REPORT 186 (2021).
[15] STAN. UNIV. INST. FOR HUM.-CENTERED A.I., 2021 ARTIFICIAL INTELLIGENCE INDEX REPORT 118 (2021).
[16] *Id.*
[17] *Id.*
[18] Neil C. Thompson et al., *Building the Algorithm Commons: Who Discovered the Algorithms that Underpin Computing in the Modern Enterprise?*, 11 GLOB. STRATEGY J. 17 (2020).
[19] *See, e.g.*, U.S. GOV'T ACCOUNTABILITY OFF., GAO-16-696T, FEDERAL AGENCIES NEED

and there are societal benefits to reducing the cost of core governance functions and improving government's internal capacity to develop, test, and hold AI systems accountable.[20] In short, a growing imbalance in AI innovation tilts toward industry, leaving academic and noncommercial research behind. Given the long-standing role of academic and non-commercial research in innovation, this shift has substantial negative consequences for the American research ecosystem.

B. The National AI Research Resource Task Force Act

Responding to these challenges, Congress enacted the National AI Research Resource Task Force Act as part of the National Defense Authorization Act (NDAA) in January 2021.[21] The Act forms part of the National Artificial Intelligence Initiative, which identifies further steps to increase research investments, set technical standards, and build a stronger AI workforce. The Act created a Task Force—the composition of which was announced on June 10, 2021[22]—to study and plan for the implementation of a "National Artificial Intelligence Research Resource" (NAIRR), namely "a system that provides researchers and students across scientific fields and disciplines with access to compute resources, co-located with publicly available, artificial intelligence-ready government and non-government data sets."[23] This research resource has also been referred to as the National Research Cloud (NRC) and was strongly endorsed by the NSCAI, which wrote that the NRC "will strengthen the foundation of American AI innovation by supporting more equitable growth of the field, expanding AI expertise across the country, and applying AI to

---

TO ADDRESS AGING LEGACY SYSTEMS (2016); U.S. GOV'T ACCOUNTABILITY OFF., GAO-19-58, CLOUD COMPUTING: AGENCIES HAVE INCREASED USAGE AND REALIZED BENEFITS, BUT COST AND SAVINGS DATA NEED TO BE BETTER TRACKED (2019).

[20] DAVID FREEMAN ENGSTROM ET AL., ADMIN. CONF. OF THE U.S., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 6, 71-72 (2020).

[21] William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, § 5106.

[22] *The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force*, THE WHITE HOUSE (June 10, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.

[23] William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, § 5106(g).

a broader range of fields."[24]

While other initiatives have sought to improve access to compute or data in isolation,[25] the NRC will generate distinct positive externalities by integrating compute and data, the two bottlenecks for high-quality AI research. Specifically, the NRC will provide affordable access to high-end computational resources, large-scale government datasets in a secure cloud environment, and the necessary expertise to benefit from this resource through a close partnership between academia, government, and industry. By expanding access to these critical resources in AI research, the NRC will support basic scientific AI research, the democratization of AI innovation, and the promotion of U.S. leadership in AI.

C. Themes

Stanford Law School's Policy Lab program convened a multidisciplinary research team of graduate students, staff, and faculty drawn from Stanford's business, law, and engineering schools to study the feasibility of and considerations for designing the NRC. Over the past six months, this group studied existing models for compute resources and government data, interviewed a wide range of government, computer science, and policy experts, and examined the technical, business, legal, and policy requirements. This Article was commissioned by Stanford's Institute for Human-Centered Artificial Intelligence (HAI), which originated the proposal for the NRC in partnership with 21 other research universities.[26]

Throughout our research, we observed three primary themes that cut across all areas of our investigation. We have integrated these themes into each section of our Article and have drawn on

---

[24] NAT'L SEC. COMM'N ON A.I., *supra* note 14, at 191.
[25] *See, e.g.*, *Cloudbank*, https://www.cloudbank.org (last visited Feb. 18, 2022); *Fact Sheet: National Secure Data Service Act Advances Responsible Data Sharing in Government*, DATA COALITION (May 13, 2021), https://www.datacoalition.org/fact-sheet-national-secure-data-service-act-advances-responsible-data-sharing-in-government/ [hereinafter *Fact Sheet*].
[26] Steve Lohr, *Universities and Tech Giants Back National Cloud Computing Project*, N.Y. TIMES (June 30, 2020), https://www.nytimes.com/2020/06/30/technology/national-cloud-computing-project.html; John Etchemendy & Fei-Fei Li, *National Research Cloud: Ensuring the Continuation of American Innovation*, STAN. UNIV. INST. FOR HUM.-CENTERED A.I., (Mar. 28, 2020), https://hai.stanford.edu/news/national-research-cloud-ensuring-continuation-american-innovation.

them to explain our findings.

- *Complementarity between compute and data.* As we evaluated the existing computing and data-sharing ecosystems, one of the systemic challenges we observed was a decoupling of compute resources from data infrastructures. High-performance computing is useless without data, and a major impediment to data sharing—particularly for high-value government data—lies in requirements for a secure, privacy-protecting computing environment.
- *Rebalancing AI research toward long-term, academic, and noncommercial research.* Presently, AI innovation is disproportionately dependent on the private sector. Public investment in basic AI infrastructure can both support innovation in the public interest and complement private innovation efforts. The NRC directs more resources toward AI development in the public interest and helps ensure long-term leadership by the United States in the field by supporting the kind of pure, basic research that the private sector cannot undertake alone.
- *Coordinating short-term and long-term approaches to creating the NRC.* Our research considers many near-term pathways for standing up a working version of the NRC by spelling out how to work within existing constraints. We also identify the structural, legal, and policy challenges to be addressed in the long term for executing the full vision of the NRC.

We summarize our main recommendations here.

D. Compute Model

The "Make or Buy" Decision. The main policy choice will be whether to build public computing infrastructure or purchase services from existing commercial cloud providers.

- It is well-established that, based solely on hardware costs, it is more cost-effective to own infrastructure

when computing demand is close to continuous.[27] The government also has experience building high-performance computing clusters, which are typically built by contractors and operated by national laboratories.[28] The National Science Foundation (NSF) has also supported many supercomputing initiatives at academic institutions.[29]

- The main countervailing concerns are that existing commercial cloud providers have software stacks and usability that AI researchers have widely adopted and may consider to be a more user-friendly platform. Commercial cloud providers offer a way to expand capacity expeditiously, although scale and availability will still be constrained by the availability of current graphics processing unit (GPU) computing resources.
- We recommend a dual investment strategy:
  - First, the compute model of the NRC can be quickly launched by subsidizing and negotiating cloud computing for AI researchers with existing vendors, expanding on existing initiatives like the NSF's CloudBank project.[30]
  - Second, the NRC should invest in a pilot for public infrastructure to assess the ability to provide similar resources in the long run. Such publicly owned infrastructure would still be built under contract or grant but could be operated

---

[27] Jennifer Villa & Dave Troiano, *Choosing Your Deep Learning Infrastructure; The Cloud vs. On-Prem Debate*, DETERMINED AI (July 30, 2020), https://determined.ai/blog/cloud-v-onprem/; *Is HPC Going to Cost Me a Fortune?*, INSIDEHPC, https://insidehpc.com/hpc-basic-training/is-hpc-going-to-cost-me-a-fortune/ (last visited July 23, 2021),

[28] *See, e.g., US Plans $1.8 Billion Spend on DOE Exascale Supercomputing*, HPCWIRE (Apr. 11, 2018), https://www.hpcwire.com/2018/04/11/us-plans-1-8-billion-spend-on-doe-exascale-supercomputing/; *Federal Government*, ADVANCED HPC, https://www.advancedhpc.com/pages/federal-government (last visited July 23, 2021); *United States Continues to Lead World In Supercomputing*, U.S. DEP'T OF ENERGY (Nov. 18, 2019), https://www.energy.gov/articles/united-states-continues-lead-world-supercomputing.

[29] *See NSF Funds Five New XSEDE-Allocated Systems,* NAT'L SCI. FOUND. (Aug. 10, 2020), https://www.xsede.org/-/nsf-funds-five-new-xsede-allocated-systems.

[30] *Cloudbank*, *supra* note 25.

> much like national laboratories (e.g., Sandia National Laboratories, Oak Ridge National Laboratory) that own sophisticated supercomputing facilities or academic supercomputing facilities.

Researcher Eligibility. While some have argued the NRC should be open for commercial access, for the purposes of this Article, we adhered to the spirit of the legislation forming the NAIRR Task Force and only reviewed the use of an NRC for academic and nonprofit AI research. We recommend that the NRC eligibility start with academics who hold "Principal Investigator" (PI) status (i.e., most faculty) at U.S. colleges and universities, as well as "Affiliated Government Agencies" willing to contribute previously unreleased, high-value datasets to the NRC in return for subsidized compute resources. PI status should be interpreted expansively to encompass all fields of AI application. Students working with PIs should presumptively gain access to the NRC. Scaling the NRC to meet the demand of all students in the United States may be challenging, but we also recommend the creation of educational programs as part of the new resource to help train the next generation of AI researchers.

Mechanism. In order to keep the award processing costs down, we recommend a base level of compute access to meet the majority of researcher computing needs. Base-level access avoids high overhead for grant administration and may meet the compute demands for the supermajority of researchers. For researchers with exceptional needs, we recommend a streamlined grant process for additional compute access.

E. Data Access Model

Focus on Government Data. We focus our recommendations for data provision/access to government data because: (1) there are already a wide range of platforms for sharing private data,[31] and (2) distribution by the NRC of private

---

[31] *See, e.g., National Data Service*, NAT'L DATA SERV., http://www.nationaldataservice.org (last visited Feb. 18, 2022); *The Open Science Data Cloud*, OPEN SCI. DATA CLOUD, https://www.opensciencedatacloud.org (last

datasets would raise a tangle of thorny IP issues. We recommend that researchers be allowed to compute on any datasets they themselves contribute, provided they certify they have the rights to that data, and the use of such data is for academic research purposes.

Tiered Access. We recommend a tiered access model: by default, researchers will gain access to government data that is already public; researchers can then apply through a streamlined process to gain access at higher security levels on a project-specific basis. It will be critical for the NRC to ultimately displace the current fragmented agency-by-agency relational approach. By providing secure virtual environments and harmonizing security standards (e.g., Federal Risk and Authorization Management Program (FedRAMP))[32], the NRC can collaborate with proposals for a National Secure Data Service[33] to provide a model for accelerating AI research, while protecting data privacy and prioritizing data security.

Agency Incentives. To incentivize federal agencies to share data with the NRC and improve the state of public sector technology, we recommend the NRC permit federal agency staff to use the NRC's compute resources. In keeping with the practices of existing data-sharing programs, such as the Coleridge Initiative,[34] we also recommend that the NRC provide training and support to work with agencies to modernize and harmonize their data standards.

Strategic Investment for Data Sources. In the short term, we recommend that the NRC focus its efforts on making available non-sensitive, low-to-moderate-risk government datasets, rather than sensitive government data (e.g., data about individuals) or data from the private sector, due to data privacy and intellectual property concerns. Researchers can still use NRC compute resources on private data but should

---

visited Feb. 18, 2022); *Harvard Dataverse*, HARVARD UNIV., https://dataverse.harvard.edu (last visited Feb. 18, 2022); *FigShare*, FIGSHARE, https://figshare.com (last visited Feb. 18, 2022).

[32] *FedRAMP*, FED. RISK & AUTHORIZATION MGMT. PROGRAM, https://www.fedramp.gov (last visited Feb. 18, 2022).

[33] *See Fact Sheet*, *supra* note 25.

[34] *See Administrative Data Research Facility*, COLERIDGE INITIATIVE, https://coleridgeinitiative.org/adrf/ (last visited July 26, 2021).

rely on existing mechanisms to acquire data for their own private buckets on the NRC. For example, images taken from Earth observation satellites, such as Landsat imagery, provide a promising low-risk, high-reward government dataset, as making such satellite imagery freely available to researchers has generated an estimated $3-4 billion in annual economic benefits, particularly when combined with high-performance computing.[35] Agencies such as the National Oceanic and Atmospheric Administration, the U.S. Geological Survey, the Census Bureau, the Administrative Office of the U.S. Courts, and the Bureau of Labor Statistics, for instance, also have rich datasets that can more readily be deployed. In the long run, access to high-risk datasets, such as those owned by the Internal Revenue Service (IRS) and the Department of Veterans Affairs (VA), will depend on the tiered access model.

F. Organizational Form

Where to institutionally locate the NRC poses a trade-off between ease of coordination to obtain compute and ease of data access. For instance, locating the NRC within a single agency would make coordination with compute providers easier, but would make data access across agencies more difficult, absent further statutory authority. Efforts to make data access to government data easier, most notably the Foundations for Evidence-Based Policymaking Act of 2018, have proven to be among the most daunting challenges of government modernization.[36] Building on those insights, we ultimately recommend that the NRC be instituted as a Federally Funded Research and Development Center (FFRDC) in the short run, and

---

[35] *See* LANDSAT DATA ACCESS, U.S. GEOLOGICAL SURV., https://www.usgs.gov/core-science-systems/nli/landsat/landsat-data-access (last visited July 23, 2021); FED. GEOGRAPHIC DATA COMM., THE VALUE PROPOSITION FOR LANDSAT APPLICATIONS (2014); CRISTA L. STRAUB ET AL., U.S. GEOLOGICAL SURV., ECONOMIC VALUATION OF LANDSAT IMAGERY (2019).

[36] *See* BIPARTISAN POL'Y CTR., BARRIERS TO USING GOVERNMENT DATA: EXTENDED ANALYSIS OF THE U.S. COMMISSION ON EVIDENCE-BASED POLICYMAKING'S SURVEY OF FEDERAL AGENCIES AND OFFICES 18-20 (2018); *see also* U.S. DEP'T OF HEALTH & HUM. SERV., THE STATE OF DATA SHARING AT THE U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES 4 (2018) (describing how data at the agency is "largely kept in silos with a lack of organizational awareness of what data are collected across the Department and how to request access.").

a public-private partnership (PPP) in the long run.

FFRDC. FFRDCs at Affiliated Government Agencies would reduce the significant costs of securing data from those host agencies. This approach will also cohere with the greater reliance on commercial cloud credits in the short run, making compute and data coordination less central. In the long run, however, streamlined coordination between data and compute may be more difficult with FFRDCs hosted at specific agencies when (1) the NRC moves away from commercial cloud credits and toward its own high-performance computing cluster, and (2) a greater number of interagency datasets become available.

PPP. In the long run, we recommend the creation of a PPP model, governed by officers from Affiliated Government Agencies, academic researchers, and representatives from the technology sector, which can house both compute and data resources.

G. Additional Considerations

Data Privacy. As an initial matter, an NRC where sensitive or individually identifiable administrative data from multiple agencies are used to build and train AI models will face challenges from the Privacy Act of 1974.[37] The Act is intended to put a check on interagency data-sharing and disclosure of sensitive data without consent.

- In order to avoid conflicts with nonconsensual interagency data-sharing, we recommend that the NRC should not be instituted as its own federal agency, nor should federal agency staff be allowed access to interagency data.
- To avoid conflicts with the Act's "no disclosure without consent" requirement, any data released to the NRC must not be individually identifiable. Despite these constraints, the majority of AI research will likely fall under the Act's statistical research exception, contingent

[37] Privacy Act of 1974, 5 U.S.C. § 552a.

on proposals aligning with an agency's core purpose.

- Given concerns about potential privacy risks, federal agencies may desire to share data, contingent on the use of technical privacy measures (e.g., differential privacy). While useful in many instances, technical approaches are no panacea and should not substitute for data access policies.
- The NRC should explore the design of virtual "data safe rooms" that enable researchers to access data in a secure, monitored, and cloud-based environment.
- Additional legislative interventions could also facilitate data-sharing with the NRC (e.g., requiring IT modernization to include data-sharing plans with the NRC).

Ethics. Rapid innovation in AI research raises a host of potential ethical challenges. Given the scope of the NRC, it will not be feasible to review every single research proposal for potential ethical violations, particularly since ethical standards are still in flux. The NRC should adopt a twofold approach.

- First, for default PI access to base-level data and compute, the NRC should establish an ex-post review process for allegations of ethical research violations. Access may be revoked when research is shown to violate ethical standards manifestly and seriously. We emphasize that the high standard for a violation should be informed by the academic speech implications and potential political consequences of government involvement in administering the NRC and determining academic research directions.
- Second, for applications requesting access to restricted datasets or resources beyond default compute, which will necessarily undergo some review, researchers should be required to provide an ethics impact statement. One of the advantages of beginning with PIs is that university faculty are accountable under existing IRBs for human subject research, as well as to the tenets of peer review.

- We urge non-NRC parties (e.g., universities) to explore a range of measures to address ethical concerns in AI compute (e.g., an ethics review process[38] or embedding ethicists in projects).[39]

Security. We recommend that the NRC take the lead in setting security classifications and protocols, in part to counteract a balkanized security system across federal agencies that would stymie the ability to host datasets. The NRC should use dedicated security staff to work with Affiliated Government Agencies and university representatives to harmonize and modernize agency security standards.

Intellectual Property (IP). While the evidence on optimal IP incentives for innovation is mixed, we recommend that the NRC adopt the same approach to allocating patent rights, copyrights, and data rights to NRC users that apply to federal funding agreements. The NRC should additionally consider conditions for requiring NRC researchers to disclose or share their research outputs under an open-access license.

Human Resources. Given its ambition, significant human resources—from systems engineers to data officers, and from grants administrators to privacy, ethics, and cybersecurity staff—will be necessary to make the NRC a success.

As we spell out, the NRC is an idea worth taking seriously. It is worth being clear, however, what it would and would not solve. The NRC *would* enable much greater access to—and in that sense, would democratize—forms of AI and AI research that have increased in computational demands, but it would *not* categorically prevent or shift the centralization of power within the tech industry. The NRC *would* shift the attention of current AI efforts into more public and socially driven dimensions by providing access to previously restricted government datasets, addressing longstanding efforts to improve access to high-value public sector data, but it would *not* create a system to prevent all unethical uses of AI. The NRC *would* facilitate

---

[38] Michael S. Bernstein et al., *ESR: Ethics and Society Review of Artificial Intelligence Research*, ARXIV (July 9, 2021), https://arxiv.org/pdf/2106.11521.pdf.
[39] Courtenay R. Bruce et al., *An Embedded Model for Ethics Consultation: Characteristics, Outcomes, and Challenges*, 5 AJOB EMPIRICAL BIOETHICS 8 (2014).

audits of large-scale models, datasets, and AI systems for privacy violations and bias, but it would *not* be tantamount to a regulatory requirement for fairness assessments and accountability. It is neither a tool of antitrust nor a certification body for ethical algorithms, which are areas worth taking seriously in independent policy proposals.[40] These broader considerations, however, do play into key areas of design and have very much informed our recommendations below on the design of the NRC. While it alone cannot solve all that ails AI, the NRC promises to take a major affirmative step forward.

Our article proceeds as follows. We begin with the fundamental question—*why* build the NRC? (Section 1)—and spell out what we view as a cogent theory of impact. We then cover *who* should have access to the NRC (Section 2), *what* comprises the NRC (Section 2), *how* access to restricted data may (or may not) be granted (Section 3), and *where* the NRC should be located (Section 4). We spend extensive time on the data access portion (Sections 3, 5, and 6), due to the complexities of government data-sharing under the Privacy Act of 1974.[41] As we note in those sections, the data portion of the NRC is complementary to long-standing efforts to enable greater research access to administrative data under, for instance, the Foundations for Evidence-Based Policymaking Act of 2018[42] and the National Secure Data Service Act proposal.[43] Such sharing must be carried out securely and in a privacy-protecting fashion. We also consider questions of ethical standards (Section 7), cybersecurity (Section 8), and intellectual property (Section 9) that inform the design of the NRC.

---

[40] *See, e.g.,* Facial Recognition and Biometric Technology Moratorium Act, S. 4084, 116th Cong. (2020); Bhaskar Chakravorti, *Biden's 'Antitrust Revolution' Overlooks AI—at Americans' Peril*, WIRED (July 27, 2021), https://www.wired.com/story/opinion-bidens-antitrust-revolution-overlooks-ai-at-americans-peril/.

[41] 5 U.S.C. § 552a.

[42] Foundations for Evidence-Based Policymaking Act of 2017, Pub. L. No. 115-435, 132 Stat. 5529 (2019).

[43] *See Fact Sheet*, *supra* note 25.

I. THE THEORY FOR A NATIONAL RESEARCH CLOUD

This section articulates a theory of impact for the NRC. In conventional policy analytic terms,[44] what problem (or market failure) does the NRC address? From one perspective, AI innovation is vibrant in the United States, with major advances occurring in language, vision, and structured data and applications developing across all sectors. Yet from another perspective, current commercialization of past innovation masks systematic underinvestment in basic, noncommercial AI research that could ensure the long-term health of technological innovation in this country.

Our case for the NRC is grounded in both efficiency and distributive rationales. First, the NRC may yield positive externalities, particularly over time, by supporting investments in basic research that may be commercialized decades later. Second, it may help to level the playing field by broadening researcher access to both compute and data, ensuring that AI research is feasible for not just the most elite academic institutions or large technology firms. Given the scale of economic transformation AI is posited to initiate over the next few decades, the stakes are potentially significant. While the largest private interests like platform technology companies and certain elite academic institutions continue to design, develop, and deploy AI systems that can be readily commercialized, a different story is playing out for the public sector and many academic institutions, which lack access to core inputs of AI research. The rising costs associated with carrying out research and development are exacerbating the disconnect between current winners and losers in the AI space.

This section proceeds in three parts. First, we survey the current landscape of AI research. Second, we articulate shifting trends in AI research and the academic-industry balance. Third, we spell out the risks of federal inaction and the benefits to an investment strategy that couples data and compute resources.

A. The AI Research Landscape

The field of AI research, as we consider it in this Article, is broadly construed. It includes not only academics who identify themselves as

---

[44] *See* STEPHEN G. BREYER, REGULATION AND ITS REFORM (1982); CLIFFORD WINSTON, GOVERNMENT FAILURE VERSUS MARKET FAILURE (2006).

researchers in artificial intelligence or machine learning, but also the broader community of researchers who use applied AI in their work, as well as those who examine its impacts on society and the environment.

Many believe, consistent with the legislation calling for the NAIRR Task Force, that AI will have a dramatic impact on society. Nine of the world's ten current largest companies by market capitalization are technology companies that place AI at the core of their business models.[45] Recent figures from the AI Index demonstrate the growing amount of investment AI companies have drawn. The most recent 2021 iteration of the Index details how global private investment in AI has grown by 40 percent since 2019 to a total of $67.9 billion, with the United States alone accounting for over $23.6 billion.[46] While multiple private sector predictions of the economic impact of AI emphasize the potential for AI to drive significant economic growth through a strong increase in labor productivity, others worry about the pace of structural change in the labor market and economic dislocation for workers automated out of their jobs or impacted by the gig economy.[47]

Such impacts are expected across domains. AI holds substantial promise to transform healthcare and scientific research: AI-related progress in the field of protein folding is poised to dramatically expedite vaccine development and pharmaceutical drug development.[48] The integration of AI-related systems into agriculture may improve crop yields through targeted use of pesticides and soil monitoring.[49] And national security experts have identified AI as a key driver of novel defense capabilities,[50] including cyberwarfare and intelligence collection.

---

[45] *Largest Companies by Market Cap*, COS. MARKET CAP, https://companiesmarketcap.com (last visited Feb. 18, 2022).
[46] STAN. UNIV. INST. FOR HUM.-CENTERED A.I., *supra* note 15, at 93.
[47] *See, e.g.*, MARY L. GRAY & SIDDARTH SURI, GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS (2019); Craig Webster & Stanislav Ivanov, *Robotics, Artificial Intelligence, and the Evolving Nature of Work, in* DIGITAL TRANSFORMATION IN BUSINESS AND SOCIETY 127, 132-35 (Babu George & Justin Paul eds., 2020); Weiyu Wang & Keng Siau, *Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda*, 30 J. DATABASE MGMT. 61 (2019).
[48] *AlphaFold: A Solution to a 50-Year-Old Grand Challenge in Biology*, DEEPMIND (Nov. 30, 2020), https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.
[49] Tanha Talaviya et al., *Implementation of Artificial Intelligence in Agriculture for Optimisation of Irrigation and Application of Pesticides and Herbicides*, 4 A.I. AGRICULTURE. 58 (2020).
[50] Greg Allen & Taniel Chan, *Artificial Intelligence and National Security,* BELFER CTR. FOR SCI., HARV. KENNEDY SCH. (July 2017), https://www.belfercenter.org/publication/artificial-intelligence-and-national-

Many countries have recognized the significance of AI as a driver of progress in economic, scientific, and national security, releasing national plans coordinating investment for continued progress in AI.[51] China's national plan announced billions of dollars in funding aimed at making the country the global leader in AI by 2030.[52] The Japanese government partnered with Fujitsu to build the world's fastest supercomputer (Fugaku).[53] Compute Canada has similarly provided research computing access to academics across the country. The U.K.'s national high-end computing resource, HECToR, was launched in 2007 at a cost of $118 million and has been used by nearly 2,500 researchers from more than 250 separate organizations who have produced over 800 academic publications.[54]

The U.S. government initially presented a more decentralized approach, providing support for AI development through National Science Foundation grants and defense spending, but refrained from releasing a unified national plan to coordinate resources across government, private industry, and universities.[55] The creation of a National AI Initiative Office,[56] the updating of the National Strategic Computing Initiative,[57] and the release of the National Security Commission on Artificial Intelligence's (NSCAI) final report[58] introduced a more comprehensive and coordinated approach. Within the United States, the closest model to the NRC may be the COVID-19 HPC consortium, which quickly provisioned compute of approximately fifty thousand GPUs and 6.8 million cores for close to one hundred projects across forty-three academic, industry, and federal government

---

security.

[51] STAN. U. INST. FOR HUM.-CENTERED A.I., *supra* note 15.

[52] JEFFREY DING, UNIV. OXFORD FUTURE OF HUMAN. INST., DECIPHERING CHINA'S AI DREAM (2018).

[53] Fugaku is being used extensively for AI research initiatives. *See* Atsushi Nukariya et al., *HPC and AI Initiatives for Supercomputer Fugaku and Future Prospects*, FUJITSU TECH. REV. 1 (Nov. 11, 2020), https://www.fujitsu.com/global/documents/about/resources/publications/technicalreview/2020-03/article09.pdf.

[54] ENG'G & PHYSICAL SCIENCES RSCH. COUNCIL, THE IMPACT OF HECToR (2014).

[55] JOSHUA NEW, *Why the United States Needs a National Artificial Intelligence Strategy and What it Should Look Like*, CTR. FOR DATA INNOVATION (Dec. 4, 2018), https://www2.datainnovation.org/2018-national-ai-strategy.pdf.

[56] Maggie Miller, *White House Establishes National Artificial Intelligence Office*, THE HILL (Jan. 12, 2021, 5:31 PM), https://thehill.com/policy/cybersecurity/533922-white-house-establishes-national-artificial-intelligence-office.

[57] *See* FAST TRACK ACTION COMM. ON STRATEGIC COMPUTING, NATIONAL STRATEGIC COMPUTING INITIATIVE UPDATE: PIONEERING THE FUTURE OF COMPUTING (2019).

[58] NAT'L SEC. COMM'N ON A.I., *supra* note 14.

consortium members united by the common goal of combating the COVID-19 pandemic.[59]

Historically, partnerships between government, universities, and industry have anchored the U.S. innovation ecosystem. The federal government played critical roles in subsidizing basic research, enabling universities to undertake high-risk research that can take decades to commercialize. This approach catalyzed radar technology,[60] the internet,[61] and GPS devices.[62] This history informed the NSCAI's recommendation for substantial new investments in AI R&D by establishing a national AI research infrastructure that democratizes access to the resources that fuel AI. Many policymakers believe that substantial investment will be needed over the next several years to support these efforts, while returns on such investments could potentially transform America's economy, society, and national security.[63]

To be sure, some may challenge the theory of impact. First, some studies dispute the premise that AI will be economically transformative. Some economists argue that many of the optimistic assessments fail to consider how constrained the uptake of AI innovation may be due to AI's inability to change essential, yet hard-to-improve tasks.[64] Others similarly critique the evidence for a fourth industrial revolution.[65] Second, some suggest that the provisioning of the NRC may strengthen the position of large platform technology companies (which of course provokes debates over antitrust in the technology sector),[66] as the NRC

---

[59] *The COVID-19 High Performance Computing Consortium*, COVID-19 HPC CONSORTIUM, https://covid19-hpc-consortium.org (last visited Feb. 18, 2022).
[60] *See* Aaron L. Friedberg, *Science, the Cold War, and the American State*, 20 DIPLOMATIC HIST. 107, 112 (1996) (book review); Sean Pool & Jennifer Erickson, *The High Return on Investment for Publicly Funded Research*, CTR. FOR AM. PROGRESS (Dec. 10, 2012), https://www.americanprogress.org/issues/economy/reports/2012/12/10/47481/the-high-return-on-investment-for-publicly-funded-research/.
[61] PETER L. SINGER, THE INFO, TECH. & INNOVATION FOUND., FEDERALLY SUPPORTED INNOVATIONS: 22 EXAMPLES OF MAJOR TECHNOLOGY ADVANCES THAT STEM FROM FEDERAL RESEARCH SUPPORT 14-15 (2014).
[62] NAT'L RSCH. COUNCIL, FUNDING A REVOLUTION: GOVERNMENT SUPPORT FOR COMPUTING RESEARCH 136-55 (1999).
[63] NAT'L SEC. COMM'N ON A.I., *supra* note 14, at 185.
[64] Philippe Aghion et al., *Artificial Intelligence and Economic Growth*, *in* THE ECONOMICS OF ARTIFICIAL INTELLIGENCE: AN AGENDA 237 (2019).
[65] Ian Moll, *The Myth of the Fourth Industrial Revolution*, 68 THEORIA 1 (2021); *see also* Tim Unwin, *5 Problems with 4th Industrial Revolution*, ICTWORKS (Mar. 23, 2019), https://www.ictworks.org/problems-fourth-industrial-revolution/.
[66] *See, e.g.*, Geoffrey A. Manne & Joshua D. Wright, *Google and the Limits of*

may be hard to launch without some involvement of hardware or cloud providers in the procurement process. Third, some would argue that the NRC would generate large negative externalities in the form of energy footprints. For instance, one study found that the amount of energy needed to train GPT-3, a leading natural language processing (NLP) model, required the greenhouse emissions equivalent of 552.1 tons of carbon dioxide,[67] approximately thirty-five times the yearly emissions of an average American.[68] Expanding access to compute without appropriate controls may contribute to wasteful computing.[69] Finally, some critics argue that any advances in AI are inherently too risky for further investment,[70] given widely documented risks of bias,[71] unintended consequences,[72] and harm.[73]

Antitrust: The Case Against the Antitrust Case Against Google, 34 HARV. J. L. & PUB. POL'Y 1 (2011); Lina M. Khan, *Amazon's Antitrust Paradox*, 126 YALE L.J. 710 (2016).

[67] David Patterson et al., *Carbon Emissions and Large Neural Network Training*, ARXIV (Apr. 23, 2021), https://arxiv.org/pdf/2104.10350.pdf. To be clear, however, the study found that training other sophisticated but smaller NLP models such as Meena and T5 required approximately 96 and 48 tons of carbon dioxide, respectively. *Id.* Another study found that the training state-of-the-art NLP models produced approximately 626,000 pounds (313 tons) of carbon dioxide, five times the lifetime emissions of the average car in the United States. Emma Strubell et al., *Energy and Policy Considerations for Deep Learning in NLP*, ARXIV (2019), https://arxiv.org/pdf/1906.02243.pdf.

[68] *Calculate Your Carbon Footprint*, NATURE CONSERVANCY, https://www.nature.org/en-us/get-involved/how-to-help/carbon-footprint-calculator/ (last visited Feb. 18, 2022).

[69] Economic studies in other fields also show that increasing access, supply, or quality of certain goods without appropriate pricing mechanisms or regulatory interventions can lead to overuse and waste. *See, e.g.*, Chengri Ding & Shunfeng Song, *Traffic Paradoxes and Economic Solutions*, 1 J. URB. MGMT. 63 (2011) (roads and traffic congestion); Ari Mwachofi & Assaf F. Al-Assaf, *Health Care Market Deviations from the Ideal Market*, 11 SULTAN QABOOS U. MED. J. 328 (2011) (doctors and quality of care).

[70] *See* Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 610 (2021).

[71] *See* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 1 (2018); Inioluwa Deborah Raji & Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, 2019 PROC. AAAI/ACM CONF. ON A.I., ETHICS & SOC'Y 429.

[72] *See* VIRGINIA EUBANKS, AUTOMATING INEQUALITY (2018); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016).

[73] *See* Christopher Whyte, *Deepfake News: AI-Enabled Disinformation as a Multi-Level Public Policy Challenge*, 5 J. CYBER POL'Y 199 (2020); Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018, 7:04 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G; James Vincent, *Google 'Fixed' its Racist*

We are cognizant of these critiques and take them seriously. This Article proceeds on the operative premise animating the NRC legislation: that it will be important for the country to maintain leadership in AI—including rigorous interrogation of its uses, limits, and promises—and that this requires supporting access to compute and data. Public investment in AI research for noncommercial purposes may help to address some of the issues of social harm we see presently in commercial contexts,[74] as well as contribute to shifting the broader focus of the field toward technology developed in the public interest by the public sector and civil society, including academia. The preceding considerations, however, have shaped our views in key respects, such as the sequential investment strategy, given the uncertainty of AI's potential; the serious consideration of publicly owned infrastructure; the provisions for ethical review of compute and data access; and, most importantly, the enablement of independent academic inquiry into the potential harms of AI systems. The NRC is not an endorsement of blind and naïve AI adoption across the board; it is a mechanism to ensure that a greater range of voices will have access to the basic elements of AI research.

B. Shifting Sources of AI Research

We now articulate how and why AI research has migrated away from basic, long-term research into commercial, short-term applications.

First, many current advances fueled by large-scale models are costly to train, relative to the size of typical academic budgets. For example, the estimated cost of training Alphabet subsidiary DeepMind's AlphaGo Zero algorithm, capable of beating the human world champion of the game Go, was more than $25 million.[75] For reference, the total annual 2019 budget for Carnegie Mellon University's Robotics Institute, one of the premier academic research institutions in the nation, was $90

Algorithm by Removing Gorillas from its Image-Labeling Tech, THE VERGE (Jan. 12, 2018, 10:35 AM), https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.

[74] KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 211 (2021) ("AI systems are built to see and intervene in the world in ways that primarily benefit the states, institutions, and corporations that they serve. In this sense, AI systems are expressions of power that emerge from wider economic and political forces, created to increase profits and centralize control for those who wield them.").

[75] Elizabeth Gibney, *Self-Taught AI is Best Yet at Strategy Game Go*, NATURE (Oct. 18, 2017), https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858.

million.[76] A white paper from the Bipartisan Policy Center[77] and the Center for a New American Security noted that the FY2020 budget for non-defense AI R&D announced by the White House was $973 million. In contrast, the combined spending on R&D in 2018 by five of the major technology platform companies was $80 billion. In sum, research universities cannot keep pace with private sector resources for compute. This is not to say that large-scale compute is necessary for all academic AI research, or that academic research is in competition with industry research, but it does illustrate why certain sectors of AI research are no longer accessible to the academic researcher.

Second, the academic-industry divide masks significant disparities between academic institutions. Using the QS World University Rankings since 2012, Fortune 500 technology companies and the top fifty universities have published five times more papers annually per AI conference than universities ranked between 200 and 500.[78] Private firms also collaborate six times more with top fifty universities than with those ranked between 301 and 500.[79] This internal compute divide across universities poses significant challenges for who is at the table.

Third, basic AI research has lost human capital.[80] When this is combined with decreased access to compute and data in academics, the prospect of conducting basic research at universities becomes less attractive. Top talent in AI now commands private sector salaries far in excess of academic salaries.[81] The departure of AI faculty from American universities has led to what some analysts have dubbed the AI Brain

---

[76] Bill Schackner, *Carnegie Mellon's Prestigious Computer Science School has a New Leader*, PITTSBURGH POST-GAZETTE (Aug. 8, 2019, 12:00 PM), https://www.post-gazette.com/news/education/2019/08/08/Carnegie-Mellon-University-computer-science-Martial-Hebert-dean-artificial-intgelligence-google-robotics/stories/201908080096.

[77] BIPARTISAN POL'Y CTR., CEMENTING AMERICAN ARTIFICIAL INTELLIGENCE LEADERSHIP: AI RESEARCH & DEVELOPMENT (2020).

[78] Nur Ahmed & Muntasir Wahed, *The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research*, ARXIV (Oct. 22, 2020), https://arxiv.org/pdf/2010.15581.pdf.

[79] *Id.*

[80] Fei-Fei Li, *America's Global Leadership in Human-Centered AI Can't Come From Industry Alone*, THE HILL (July 6, 2021, 12:30 PM), https://thehill.com/opinion/technology/561638-americas-global-leadership-in-human-centered-ai-cant-come-from-industry?rl=1.

[81] Cade Metz, *A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit*, N.Y. TIMES (Apr. 19, 2018), https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html.

Drain: while AI PhDs in 2011 were roughly as likely to go into industry as academia, two-thirds of AI PhDs now go into industry and less than a quarter go into academia.[82] One study suggests that the departure of AI faculty also has a negative effect on startup formation by students.[83]

Fourth, as large-scale AI research migrates to industry, the focus of research inevitably shifts. While academic researchers in AI may lack access to the volume of data needed to train AI models,[84] large-platform companies have access to vast datasets, including those about or created by their customers. This data divide in turn distorts AI research toward applications that are focused on private profit, rather than public benefit.[85] Put more colorfully by Jeff Hammerbacher, "The best minds of my generation are thinking about how to make people click ads."[86] The NRC can play a key role in unlocking access to public sector data, which may help to reorient the focus of AI research away from private sector datasets.[87]

The hollowing out of academic AI capacity can be seen in OpenAI's analysis of the relationship between compute and fifteen relatively well-known "breakthroughs" in AI between 2012 and 2018.[88] Although the analysis was meant to emphasize the role of computing power, it also illustrates an emerging gap between private sector and academic contributions over time. Of the fifteen developments examined, eleven were achieved by private companies while only four came from academic institutions. Furthermore, this imbalance increases over time: though private sector research has continued accelerating since 2012, academic output has stagnated. The last of the major compute-intensive breakthroughs in OpenAI's analysis stemming from academia was Oxford's 2014 release of its VGG image-recognition program; NYU's work on Convolutional Neural Networks dates back to 2013. From 2015

---

[82] STAN. U. INST. FOR HUM.-CENTERED A.I., *supra* note 15, at 118.
[83] Michael Gofman & Zhao Jin, *Artificial Intelligence, Education, and Entrepreneurship*, SSRN (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3449440.
[84] Sadowski, *supra* note 12.
[85] For example, researchers have clamored for Facebook to share some of its proprietary data so they can better understand the effect of social media on politics and societal discourse. Simon Hegelich, *Facebook Needs to Share More with Researchers*, 579 NATURE 473 (2020).
[86] Ashlee Vance, *This Tech Bubble Is Different*, BLOOMBERG (Apr. 14, 2011, 5:00 PM), https://www.bloomberg.com/news/articles/2011-04-14/this-tech-bubble-is-different.
[87] O'Hara & Medalia, *supra* note 13.
[88] Dario Amodei & Danny Hernandez, *AI and Compute*, OPEN AI (May 16, 2018), https://openai.com/blog/ai-and-compute/.

to 2018, all eight breakthroughs included in OpenAI's analysis came out of private companies. Taken together, this leads observers to argue that academic researchers are increasingly unable to compete at the frontier of AI research.[89] While academic researchers have continued to make important contributions in AI, these are increasingly restricted to less compute-intensive problems. With fewer compute-intensive academic breakthroughs, AI innovations have focused on private interests (e.g., online advertising) as opposed to long-term, noncommercial benefits. To be sure, the private sector has, of course, been central to AI research, but the concern is about the long-term balance of the AI innovation ecosystem.

C. Scoping Federal Intervention in Data and Compute

How can we achieve a more balanced approach toward research and development? We first consider the risks of federal inaction and discuss some of the unique advantages of addressing data and compute together.

1. Risks of Federal Inaction

The risks of federal inaction are twofold. First, basic AI research that has, to date, paved the way for advances in AI and machine learning will slow. According to a recent study, approximately 82 percent of the algorithms used today originated from nonprofit groups and universities supported by government spending.[90] Even when industry research is successful, it is typically product-focused or incremental, harder to reproduce, and may not be published or open-sourced. An interesting case lies in recent breakthroughs in protein folding. In late 2020, the Alphabet subsidiary DeepMind announced that it had developed a program called AlphaFold, an AI-driven system capable of accurately predicting the structure of a vast number of proteins using only the sequence of nucleotides contained in its DNA. Whether out of concern for the privatization or to accelerate the adoption of related systems, a consortium of academics, led by scientists at the University of

---

[89] *See, e.g.*, Ahmed & Wahed, *supra* note 78; Ian Sample, *'We Can't Compete': Why Universities Are Losing Their Best AI Scientists*, THE GUARDIAN (Nov. 1, 2017, 6:30 AM), https://www.theguardian.com/science/2017/nov/01/cant-compete-universities-losing-best-ai-scientists.

[90] Thompson et al., *supra* note 18.

Washington, developed an open-source competitor called RoseTTaFold.[91] DeepMind did make AlphaFold available to a broad audience, but the concerns illustrate the risks of science posed by exclusively private AI research, reminiscent of the race to sequence the human genome, where public investment in the Human Genome Project preempted concerns about a private firm patenting the human genome.[92]

Second, federal inaction could widen significant inequalities in the AI landscape. Without increased access to computing, education, and training, large parts of the economy may be unable to adapt—whether in financial services, healthcare, education, or government. Diversifying the range of AI research may also promote progress and productivity. One study suggests that the diversity of AI research trajectories—that is, the specific questions, topics, and problems researchers choose to investigate—has become more constrained in recent years and that private-sector AI research is less diverse than academic research.[93] Smaller academic groups with lower private sector collaboration appear to bolster the diversity of AI research.[94] From the standpoint of underdeveloped avenues of research, such as ethics and accountability in AI, increasing the range of research topics and methods in the field raises the likelihood of finding breakthroughs that make additional progress in the long term possible.[95] Recent evidence suggests that between 2005 and 2017, just five metro areas in the U.S. accounted for 90 percent of the growth in innovation sector jobs.[96] According to Stanford economist Erik Brynjolfsson, the likely impact of geographic concentration is "there are a whole lot of people— hundreds of millions in the U.S. and billions worldwide—who could be innovating and who are not because they do

---

[91] Minkyung Baek, *RoseTTAFold: Accurate Protein Structure Prediction Accessible to All*, UNIV. WASH. INST. FOR PROTEIN DESIGN (July 15, 2021), https://www.ipd.uw.edu/2021/07/rosettafold-accurate-protein-structure-prediction-accessible-to-all/; Minkyung Baek et al., *Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network*, 373 SCI. 871 (2021).

[92] *How Diplomacy Helped to End the Race to Sequence the Human Genome*, 582 NATURE 460 (2020).

[93] Joel Klinger et al., *A Narrowing of AI Research?*, ARXIV (Nov. 17, 2020), https://arxiv.org/pdf/2009.10385.pdf.

[94] *Id.*

[95] Alex Tamkin et al., *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models*, ARXIV (Feb. 4, 2021), https://arxiv.org/pdf/2102.02503.pdf.

[96] Those 5 were Boston, San Francisco, San Jose, Seattle, and San Diego. *See* Robert D. Atkinson et al., *The Case for Growth Centers: How to Spread Tech Innovation Across America*, BROOKINGS (Dec. 2019), https://www.brookings.edu/research/growth-centers-how-to-spread-tech-innovation-across-america/.

not have access to basic computer science skills, or infrastructure, or capital, or even culture and incentives to do so."[97] AI technologies can be hard to diagnose and interpret and can be prone to substantial bias.[98] Broadening the set of voices that can interrogate such systems will be critical to an inclusive and equitable future.

In sum, federal investment in public AI infrastructure may promote a more equitable distribution of participation in and gains to AI innovation broadly, bolster U.S. competitiveness, and support fundamental research into noncommercial and public sector applications.

II. ELIGIBILITY, ALLOCATION, AND INFRASTRUCTURE FOR COMPUTING

This section discusses eligibility, resource allocation, and computing infrastructure for the NRC: *Who* should get access to *what* and *how*?

First, when determining who should get access, it is critical to bear in mind the broad goals of the NRC. As discussed in Section 1, there is a large resource gap in academia as compared to private industry. In the interest of supporting basic research and democratizing the field, this section will focus on identifying a target group for eligibility. As we articulate below, we refrain from considering expansion to a broader set of commercial, nonacademic parties because of the NRC's focus on long-term, fundamental scientific research. One of the narrowest approaches would be a specialty faculty model that would target researchers engaged in core AI work. But the difficulties with defining AI and the rapidly expanding domains in which AI is being applied make this model too constrained to realize the full impact of the NRC. Instead, we recommend tracking the most common criterion for federal research funding and advocate that eligibility should hinge on "Principal Investigator" (PI) status at U.S universities.[99] One of the trade-offs is that PIs may be less diverse than a broader segment of researchers,[100] so a longer-term

---

[97] Interview with Professor Erik Brynjolfsson, Dir., Stan. Digit. Econ. Lab (Feb., 22, 2021).

[98] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).

[99] "Principal Investigator" status may differ from university to university, but typically represents the core faculty that are eligible to oversee research projects at their home institutions.

[100] *See* Beth Jensen, *AI Index Diversity Report: An Unmoving Needle*, STAN. UNIV. INST. FOR HUM.-CENTERED A.I. (Mar. 3, 2021), https://hai.stanford.edu/news/ai-

expansion could consider moving beyond this group. While the NRC aims to train the next generation of AI researchers, we caution that an immediate expansion to all graduate and undergraduate students would pose considerable challenges in scaling. Therefore, we recommend that students primarily gain access by participation in faculty-sponsored AI research, instead of blanket student access, and that they gain training through the creation of educational programs.

Second, we discuss three models for allocating computing credit: development of a new grant process, delegating block compute grants to universities for internal allocation among faculty, or universal access. Each of these models trades off the ease of administration against tailoring for specific NRC goals. We recommend an approach used by other national research clouds—namely a hybrid approach of universal default access for the majority of researchers, with a grant process for excess computing beyond the default allocation. Such an approach would keep administrative costs low for the vast majority of researchers, while enabling tailoring through a competitive grant process for the highest-need users.

Third, we consider the "make-or-buy" decision for the NRC. One option would be for the NRC to provide research grants for the use of commercial cloud services that many researchers already rely on (the "buy" decision). Alternatively, the NRC could create and provision access to a publicly high-performance computing cluster (the "make" decision). It is well-established that, based solely on hardware costs, it is more cost-effective to own infrastructure when computing demand is close to continuous. On the other hand, existing commercial cloud providers have developed highly usable software stacks that AI researchers have widely adopted. Commercial cloud providers offer a way to quickly expand capacity. We thus recommend a dual investment strategy to (a) quickly launch the NRC by subsidizing and negotiating cloud computing for AI researchers with existing vendors, expanding on existing initiatives like the National Science Foundation's CloudBank project; and (b) invest in a pilot for public infrastructure to assess the ability to provide similar resources in the long run. Such publicly owned infrastructure would likely be built under contract or grant but could be operated much like national laboratories that own sophisticated supercomputing facilities, as is the case with other national research resources (e.g., Compute Canada and Japan's Fugaku).

---

index-diversity-report-unmoving-needle.

Our recommendations are informed by a series of case studies that are presented throughout this section, as well as through the remainder of the Section. Table 1 summarizes how existing models compare on the three key design decisions. At the outset, we note that few existing initiatives have attempted to provide compute power at the scale of the NRC. At the same time, we view the NRC as complementary to more traditional areas of scientific computing. [101]

	ELIGIBILITY			ALLOCATION				OWNERSHIP	
Existing Program	PI Only	Any Faculty	Students	Existing Grant Process	University Allocation	New Process	Default Access w/Tiers	Private	Public
CloudBank	X		X	X					X
Stanford HAI-AWS Cloud Program		X			X			X	
Stanford Sherlock Cluster	X						X	X	
Google Colab		X	X				X	X	
Compute Canada	X						X		X
Fugaku		X				X			X
XSEDE	X	X					X		X
DOE INCITE	X					X			X

***Table 1:** Key design differences between computing case studies. "Other faculty" indicates an eligibility set for faculty other than PI status (e.g., requiring Stanford affiliation for the Sherlock cluster), and "new process" is used to indicate the creation of any process other than those currently listed (e.g., Fugaku is currently soliciting proposals with research facilities).

A. Eligibility

The first task is identifying which researchers should be eligible for the NRC. Section 1 discussed the need to support AI innovation in universities. Therefore, this section will scope eligibility within academia by analyzing the access-resource trade-offs in alignment with the NRC

[101] For a perspective, for instance, on the importance of modeling and simulation in physics, see Karen E. Willcox et al., *The Imperative of Physics-Based Modeling and Inverse Theory in Computational Science*, 1 NATURE COMPUTATIONAL SCI. 166 (2021).

goals.

At the outset, we note that we do not analyze eligibility in depth beyond academic researchers. The legislation constituting the NRC task force specifically contemplates "access to computing resources for researchers across the country."[102] The NRC is defined as "a system that provides researchers and students across scientific fields and disciplines with access to compute resources."[103] The most natural interpretation of this language suggests a core focus on scientific and academic research.[104]

Introducing commercial access to the NRC, particularly for under-resourced firms such as small businesses and startups, may very well benefit the U.S. innovation ecosystem. But the challenges of enabling commercial access to the NRC are enormously complex. First, including software developers at startup companies as "researchers" within the meaning of the NDAA would raise a wide range of boundary questions that the NRC may be poorly equipped to adjudicate. According to the Small Business Administration (SBA), there are over 31 million small businesses in the United States.[105] Over 627,000 businesses open each year.[106] Should all such businesses be eligible to compute on the NRC?

---

[102] 15 U.S.C. § 9415.
[103] *Id.*
[104] Contemporaneous accounts corroborate this core focus. The National Security Commission on AI, for instance, describes the proposal as "provid[ing] verified researchers and students subsidized access to scalable compute resources" with a specific reference to the "compute divide" that has left "middle- and lower-tier universities [lacking] the resources necessary for cutting-edge AI research." NAT'L SEC. COMM'N ON A.I., *supra* note 14, at 197. Upon the announcement of the NRC legislation, Jeff Dean, SVP of Google Research and Google Health, noted, "A National AI Research Resource will help accelerate US progress in artificial intelligence and advanced technologies by providing academic researchers access to the cloud computing resources necessary for experiments at scale." Brandi Vincent, C*ongress Inches Closer to Creating a National Cloud for AI Research*, NEXTGOV (July 2, 2020), https://www.nextgov.com/emerging-tech/2020/07/congress-inches-closer-creating-national-cloud-ai-research/166624/. Others have suggested that "researchers" under NRC could include individuals at small businesses, start-up companies, non-profits, and certain technology firms. One co-sponsor of the legislation, for instance, suggested that NRC resources should be provided to "developers" and "entrepreneurs." Press Release, Sen. Rob Portman, Portman, Heinrich Introduce Bipartisan Legislation to Develop National Cloud Computer for AI Research (June 4, 2020), https://www.portman.senate.gov/newsroom/press-releases/portman-heinrich-introduce-bipartisan-legislation-develop-national-cloud.
[105] *Frequently Asked Questions About Small Businesses*, U.S. SMALL BUS. ADMIN. OFF. OF ADVOC. (Oct. 2020), https://cdn.advocacy.sba.gov/wp-content/uploads/2020/11/05122043/Small-Business-FAQ-2020.pdf.
[106] Louise Balle, *Information on Small Business Startups*, HOUS. CHRON., https://smallbusiness.chron.com/information-small-business-startups-2491.html

How would one avoid gaming (e.g., strategic subsidiaries/spinoffs) eligibility? And how would this advance the scientific mission of the NRC? Second, while potentially valuable, it is less clear how the inclusion of startups and small businesses meets the theory of impact of the NRC. As currently construed, the concern animating the NRC lies in the importance of long-term, noncommercial fundamental research that can ensure AI leadership for decades to come. Commercialization is not the element of the AI innovation ecosystem that faces the structural challenges articulated in Section 1. Finally, scaling the NRC to allow meaningful commercial access would pose serious practical challenges. Because the Task Force must also consider the feasibility of the NRC, we have not considered in depth a conception that would extend the term "researcher" to encompass large portions of the commercial private sector. Expansion to non-academic, nonprofit organizations may be a more reasonable consideration, as the objective of some entities (e.g., not-for-profit investigative journalism, civil society organizations) may be closer to the core of the NRC's mission of empowering long-term beneficial research that cannot currently occur.[107] In the long term, the NRC should consider the trade-offs to such an expansion.

Even if the NRC adopts a broader computing model down the road, we believe that focusing on academic researchers is an important starting point as it illuminates some of the main operational considerations for NRC access.

1. Specialty Faculty Model

One of the narrowest approaches to NRC eligibility would be to restrict it to faculty engaged in AI research. Under this approach, policymakers would direct computing resources exclusively toward faculty working on identifiable AI projects, which often need large amounts of compute power. A benefit of this approach is that researchers' familiarity with the infrastructure would likely mean that fewer funds would be devoted to cloud service training for novice users.

Yet self-identified AI academics are few and concentrated in a small number of universities, which are already more likely to gain access to large-scale computing. Limiting access to core AI faculty would hence

---

(last visited Feb. 18, 2022).

[107] Such entities could potentially collaborate with academic partners, and the NRC would of course also need to set rules about collaborator eligibility.

undermine the mission of democratizing AI research. In addition, the application of AI is expanding rapidly across domains. Interdisciplinary research deploying AI in new domains will be vital for maintaining American leadership in AI, as well as for animating basic research questions. Restricting eligibility to core AI faculty (however defined) could jeopardize the ability of researchers from all academic disciplines (e.g., in the physical sciences, social sciences, and humanities) to contribute to realizing AI's full potential.

2. General Faculty Model

A more natural starting point for NRC eligibility is with Principal Investigators (PIs) at U.S. colleges and universities, which are the most commonly deployed criterion for federal grants. Requirements for PI status are set by individual universities and include a broad range of researchers certified by their university as qualified to lead large research projects.[108] While PI certification may vary from institution to institution, an important baseline criterion of PI status is that the researcher is subject to their institution's training and certification processes, which in turn clarify a researcher's responsibilities regarding the management and execution of their research proposals. Existing programs for allocating computing power typically set eligibility based on PI status as it ensures the researcher has the infrastructure to carry out a large-scale research project. CloudBank, an NSF program that distributes funds for commercial cloud computing resources, awards grants to PIs, who may distribute funds to other researchers and students

[108] PI status provides a level of standardization across faculty compared to other metrics, such as tenure-track or designation as research faculty. For example, the University of Michigan appoints individuals focused on full-time research as "research faculty," which is not a tenure-track position. *See* https://orsp.umich.edu/principal-investigator-pi. In contrast, research faculty at Purdue are eligible for tenure track. *See https://www.purdue.edu/policies/human-resources/vif8.html* . Distinct from the categorization used by both universities, MIT designates full-time researchers as "academic staff" rather than faculty. *See https://policies.mit.edu/policies-procedures/50-research-appointments/53-academic-research-staff-appointments ;* https://research.mit.edu/research-policies-and-procedures/research-and-academic-appointments. All three types of researchers, however, qualify for principal investigator status at their respective universities. Some universities go further by providing temporary PI status to non-PI status individuals affiliated with the university for a single project (including all three universities mentioned previously).

on the project.[109] Compute Canada allows all faculty granted PI status by their university to automatically receive a preset amount of computing credits and apply for further credit as needed. The PI may then sponsor others to access the credit.[110]

We recognize that PI status does not include all university-affiliated researchers. In 2013, of the over two hundred thousand self-identified academic researchers, just under sixty thousand were employed in a role other than full-time faculty, a position that may not be eligible for PI status.[111] From 1973 to 2013, the percentage of full-time faculty among engineering doctorate holders decreased by 2 percent, while the percentage of "other" academic jobs (including research associates) increased by 12 percent.[112] But the reliance on PI status would not prevent PIs from allocating access to non-PI status researchers on a project, and administrative ability weighs strongly in favor of consistency with current grant eligibility criteria.

3. Students

Should graduate and undergraduate students be able to access the NRC? One of the principal challenges here lies in scale and administrability. One estimate is that there are nearly 20 million college students in the U.S.[113] Second, PI-oriented eligibility does not preclude university students from accessing resources to undertake AI research under the direction of PIs. The Compute Canada model, for example, restricts eligibility to faculty but allows faculty to sponsor collaborators, including any student researcher. An access model for the NRC that allows PIs to sponsor students provides further research and training opportunities for students. Third, a number of existing cloud services already provide limited access to computing credits for educational purposes. Google Colaboratory, for instance, provides free, but not

---

[109] *Community & Education Resource Requests*, CLOUDBANK, https://www.cloudbank.org/training/cloudbank-community - toc-eligibilit-36nfpcrS (last visited Feb. 21, 2022).
[110] *Apply for an Account*, COMPUTE CAN., https://www.computecanada.ca/research-portal/account-management/apply-for-an-account/ (last visited Feb. 21, 2022).
[111] NAT'L SCI. BD., NAT'L SCI. FOUND., SCIENCE & ENGINEERING INDICATORS 2016 72 (2016).
[112] *Id.*
[113] *College Enrollment in the United States from 1965 to 2019 and Projections up to 2029 for Public and Private Colleges,* STATISTA (Jan. 2021), https://www.statista.com/statistics/183995/us-college-enrollment-and-projections-in-public-and-private-institutions/.

reliably guaranteed, access to cloud services.[114] Amazon Web Services provides up to $35 of AWS credits for free to all university faculty and students. Despite existing resources, students may need more resources. The Google subsidiary and online community Kaggle, for example, provides thirty hours of GPU access per week for free and found that 15 percent of users exceeded the limit.[115]

While the exact scope of student computing power needs is unclear, we recommend funding an educational resource once researcher needs and resource limitations have been gauged. Currently, the NSF's CloudBank is piloting a Community & Education Resource to earmark a small set of credits for educational purposes.[116] This resource allows a university professor to request a small number of credits for student coursework or small-scale research.

Regardless of which eligibility model the NRC adopts, there will also be a significant need for support staff, training documentation, and educational materials so researchers can effectively make use of the computer and data resources (see Appendix D). The reason some students and researchers may not take advantage of all available cloud credits could, for instance, stem from the difficulty in using cloud platforms. If the NRC serves academics from a range of disciplines, this question of human capital will be especially relevant to serve different models of research. A robust training program for users of the NRC will ensure ease of use and encourage appropriate utilization of the cloud.

B. Resource Allocation Models

We now consider three resource allocation models: (1) a new grant process; (2) block grant allocation to universities; and (3) universal — but potentially tiered — access.

1. NRC Grant Process

Establishing a new grant process for compute access would have one main advantage. The program could be built specifically for the purpose of AI research, with reviewers who are familiar with AI concepts,

---

[114] *Colaboratory, Frequently Asked Questions*, GOOGLE, https://research.google.com/colaboratory/faq.html (last visited Feb. 21, 2022).
[115] *Weekly Maximum GPU Usage*, KAGGLE, https://www.kaggle.com/general/108481 (last visited, Feb. 21, 2022).
[116] *Community & Education Resource Requests*, *supra* note 109.

practices, and trends. Such a process might therefore enable improved allocation decisions and provide the NRC with greater control over its investments.

That said, establishing a peer-review process for all applications would be resource-intensive, requiring the establishment of a grant administration program akin to those at the National Science Foundation (NSF) or the National Institutes of Health (NIH). For instance, to implement peer review required for the merit review process, the NSF annually needs a community large enough to conduct nearly 240,000 reviews per year.[117] Since the contemplated reach is broad, we are mindful of adding a significant service burden for faculty conversant in AI for every application for compute access. Peer review for compute access would require significant overhead and delays in compute allocation.

2. University Access

To reduce administrative costs, an alternative scheme would be to allocate credits to universities based on the number of eligible researchers. The NRC could allocate resources to universities as block grants, and in turn, rely on the university to distribute computing access. For example, the NRC could purchase significant amounts of compute from cloud providers, create virtual credits that are convertible into appropriate cloud resources, and delegate allocation to universities. This approach would have the advantage of tapping into the universities' local expertise for reviewing and distributing resources. It would, however, lead to a highly decentralized process, providing little oversight to understand the distribution of usage and would give the NRC little control over resource allocation. While we do not recommend this route as the principal allocation scheme, we do believe that some allocation to university-based IT support teams may be warranted to support researchers in using the NRC. XSEDE's "Campus Champions" program, for instance, provides university employees access to the system to support the computational transition.[118]

---

[117] *Merit Review: Why You Should Volunteer to Serve as an NSF Reviewer*, NAT'L SCI. FOUND., https://www.nsf.gov/bfa/dias/policy/merit_review/reviewer.jsp#1 (last visited Feb. 21, 2022).

[118] *See XSEDE Campus Champions,* XSEDE, https://www.xsede.org/community-engagement/campus-champions (last visited Feb. 21, 2022).

3. Universal Access

The last potential model would provision universal access to base-level compute to all eligible PIs. The closest model is Compute Canada's national research cloud, which provides base-level compute access to all faculty in Canada. This would significantly reduce administrative overhead, both for an institution running the review process, and academics seeking NRC access. The primary downside is that base-level compute may be insufficient for specialized needs.

We recommend combining a universal baseline model with a grant process for compute needs beyond base-level access. The reduced complexity in administering a universal baseline access compute model makes it an attractive option for the NRC in allocating compute resources, especially with respect to the NRC's goal of opening access to compute resources.[119] XSEDE, for instance, uses a similar model of streamlined "Startup Allocations" (issued for one-year terms, typically within two weeks of application) and "Research Allocations" for more significant compute requests. Compute Canada provides access to 15 percent of PIs to increased compute capacity based on a merit competition. A critical question will, of course, be the level of baseline computing that will determine overall costs, physical space requirements, and the like. To benchmark this, we recommend an in-depth study of the anticipated computing needs, based on existing academic computing centers.[120]

---

[119] Compute Canada, for instance, provides access to increased compute capacity to 15 percent of PIs based on a merit competition. In 2021, Compute Canada completed its review of 650 research submissions in about five months with only eighty volunteer reviewers from Canadian academic institutions to assess the scientific merit of each proposal. *Resource Allocation Competitions*, COMPUTE CAN., https://www.computecanada.ca/research-portal/accessing-resources/resource-allocation-competitions/; *2021 Resource Allocations Competition Results*, COMPUTE CAN., https://www.computecanada.ca/research-portal/accessing-resources/resource-allocation-competitions/rac-2021-results/ (last visited Feb. 21, 2022). Compare this with CloudBank, which allocates compute resources by leveraging NSF's grant administration process: In 2019, NSF needed 30,000 volunteer reviewers to handle over forty thousand proposals, with each proposal requiring about ten months to process from start to finish. NAT'L. SCI. FOUND., MERIT REVIEW PROCESS: FISCAL YEAR 2019 DIGEST (2020); *NSF Proposal and Award Process*, NAT'L SCI. FOUND., https://www.nsf.gov/attachments/116169/public/nsf_proposal_and_award_process.pdf (last visited Feb. 21, 2022).

[120] Another boundary question will be the resource allocation to PIs that are affiliated both with universities and with private companies. As a default, NRC resources should go toward academic projects, and not subsidize work that is conducted in one's private researcher capacity.

The grant process for additional compute could take multiple forms; for example, while one could allow individual PIs to apply directly to the NRC for excess compute, the NRC could also allocate "blocks" of resources at the university level and allow universities to oversee their administration. In any case, due to the size of such requests, grant reviews should be conducted on a merit basis and administered by a combination of NRC staff and an external advisory board of university faculty. In 2021, Compute Canada, for instance, completed its review of 650 research submissions in about five months, with only eighty volunteer reviewers from Canadian academic institutions to assess the scientific merit of the proposal.[121] In order to avoid conflicts of interest, we strongly recommend against the participation of any faculty or private sector advisers who have conflicts of interest with any vendors that provide services to the NRC. Ideally, proposal reviews should be independent, blinded, and based on scientific merit to the extent possible.

CASE STUDY: CloudBank

In 2018, the National Science Foundation's (NSF) Directorate for Computer and Information Science and Engineering (CISE) created the Cloud Access Solicitation to provide funding for AI-related research endeavors. Initially created to meet the needs of the NSF funding recipients to access public clouds, CloudBank is an interesting case study for exploring resource allocation models. Accessible through a portal, CloudBank aids researchers in using cloud resources fully by facilitating the process of "managing costs, translating and upgrading computing environments to the cloud, and learning about cloud-based technologies."[122]

CloudBank is a collaboration project established via an NSF Cooperative Agreement with the San Diego Supercomputer Center (SDSC) and the Information Technology Services Division at UC San Diego, the University of Washington eScience Institute, and UC Berkeley's Division of Data Science and Information.[123] Each of these institutions handles

[121] *Resource Allocation Competitions*, *supra* note 119.
[122] *Simplifying Cloud Services*, SCI. NODE (Dec. 2, 2019), https://sciencenode.org/feature/An%20easier%20cloud.php.
[123] *Frequently Asked Questions (FAQ)*, CLOUDBANK (Dec. 2, 2019), https://www.cloudbank.org/faq.

an area, according to its comparative advantage.[124] For example, SDSC is responsible for building the online portal, and UC San Diego is in charge of managing the accounts of the users.[125]

CloudBank also aims to reduce the cost of cloud computing: it uses both the ongoing discounts with cloud providers from the University of California and the discounts that come with bulk cloud purchases from the cloud procurement consulting firm Strategic Blue, which regularly partners with the likes of AWS, Microsoft, and Google.[126] Furthermore, there is no overhead cost associated with the cloud allocations through CloudBank, since the terms of the NSF cooperative agreement prohibit indirect costs.[127] With these cost-saving mechanisms, researchers can afford more computing capacities from a variety of major cloud vendors.

By requesting the use of CloudBank during their application to the selected NSF projects,[128] researchers can gain access not only to various advanced hardware resources but also to a variety of services to make the process more supported and monitored.[129] CloudBank also gives research community members access to its education and training information.[130]

C. Computing Infrastructure

Cloud computing environments connect local computing devices such as desktop computers to large, typically geographically distributed servers containing physical hardware. This hardware, in turn, is responsible for storing data and performing computation over computer networks—all of which is mediated through a collection of software services. This model centralizes the usual operational management for those using the network and provides adjustable units of computation

---

[124] *Simplifying Cloud Services*, *supra* note 122.
[125] *Id.*
[126] *Id.*
[127] *Frequently Asked Questions (FAQ)*, *supra* note 123.
[128] *Frequently Asked Questions (FAQs) for Budgeting for Cloud Computing Resources via CloudBank in NSF Proposals*, NAT'L SCI. FOUND., https://www.nsf.gov/pubs/2020/nsf20108/nsf20108.jsp (last visited Feb. 21, 2022).
[129] *Simplifying Access to Cloud Resources for Researchers: CloudBank*, AMAZON WEB SERVS. (Nov. 16, 2020), https://aws.amazon.com/blogs/publicsector/simplifying-access-cloud-resources-researchers-cloudbank/.
[130] *Community & Education Resource Requests*, *supra* note 109.

and data storage to allow for fluctuations in demand. Users interact with the cloud by launching virtual connections to the server—cloud instances—and running containerized processes remotely. These operations are managed by the cloud and available for monitoring through dashboards. Cloud computing may be serviced through on-premises clusters, via external vendors, or some combination thereof, and accessed over networks with varying security and connectivity, from internet-accessible to air-gapped regions.

The infrastructure of the NRC could be developed with two general approaches: (1) the NRC could use commercial cloud platforms as its infrastructure backbone; or (2) the federal government could engage a contractor to build a high-performance computing (HPC) public facility specifically for the NRC. This section addresses some advantages and disadvantages of both. We provide an estimated cost comparison of these two approaches in Appendix A. The two approaches discussed here are not mutually exclusive, and we ultimately recommend a hybrid investment strategy. In the short run, the NRC should scale up cloud credit programs (similar to NSF's CloudBank program) to provide both streamlined base-level access and merit review for applications going beyond base-level access. In the long run, the NRC should invest in a pilot to develop public computing infrastructure. Even with public infrastructure, it will be critical to meet "burst demand" (to expand resources when compute demand peaks). The success of the initial investments should guide the prospective model as to whether to rely on publicly or privately owned infrastructure in the longer term. We note that, in order to scale successfully to either resource, it will require building institutional capacity at academic institutions.

1. Commercial Cloud

The greatest advantage of using commercial cloud services for the NRC is that significant infrastructure already exists.[131] Under this model,

---

[131] Larry Dignan, *AWS Cloud Computing Ops, Data Centers, 1.3 Million Servers Creating Efficiency Flywheel*, ZDNET (June 17, 2016), https://www.zdnet.com/article/aws-cloud-computing-ops-data-centers-1-3-million-servers-creating-efficiency-flywheel/; Rich Miller, *Ballmer: Microsoft Has 1 Million Servers*, DATA CTR. KNOWLEDGE (July 15, 2013), https://www.datacenterknowledge.com/archives/2013/07/15/ballmer-microsoft-has-1-million-servers; Daniel Oberhaus, *Amazon, Google, Microsoft: Here's Who Has the Greenest Cloud*, WIRED (Dec. 18, 2019), https://www.wired.com/story/amazon-google-microsoft-green-clouds-and-hyperscale-data-centers/; Russell Brandom,

the NRC would simply subsidize credits for using commercial cloud services (similar to NSF's CloudBank program). Thus, rather than spending years building new computing resources, policymakers could launch the NRC soon after they determine the program's administrative details. We note, however, that there may still be significant GPU shortages in the short run; with the contemplated scale of the NRC, significant infrastructure would need to be built. Since many researchers already use commercial cloud services for their AI research, the transition into the NRC program could be relatively seamless. Furthermore, commercial cloud platforms offer the NRC greater flexibility to change the size and scope of the program. Commercial cloud platforms charge for the amount of compute actually used.[132] Thus, the size of the NRC could expand or retract in line with shifting demand. In contrast, a dedicated HPC system would have a set amount of hardware that costs the same, no matter how effectively it's being used.

Working directly with commercial cloud providers also offers several advantages for the NRC. The commercial cloud services market is highly competitive and features numerous providers capable of meeting the NRC's needs. The NRC would have the option of using one provider or multiple providers. If opting to use just one provider, the government's bargaining power may be at its strongest in helping to drive down prices for the NRC. Alternatively, using multiple providers gives the NRC greater flexibility in available services and hardware. Either way, policymakers would have the opportunity to negotiate contracts and prices with commercial cloud providers every few years, which will be critical to cost containment.[133] The NRC would also not be locked into using the same provider or set of providers for the duration of the program. Rather, NRC staff could reevaluate which commercial cloud provider's infrastructure would best meet the NRC's needs at the start of each new contract.

Mapping out Amazon's Invisible Server Empire, THE VERGE (May 10, 2019), https://www.theverge.com/2019/5/10/18563485/amazon-web-services-internet-location-map-data-center.

[132] *See, e.g., AWS Pricing*, AMAZON WEB SERVS., https://aws.amazon.com/pricing/ (last visited Feb. 21, 2022); *Overview of Cloud Billing Concepts*, GOOGLE CLOUD, https://cloud.google.com/billing/docs/concepts; *Azure Pricing*, AZURE, https://azure.microsoft.com/en-us/pricing/#product-pricing (last visited Feb. 21, 2022).

[133] Large research universities already negotiate enterprise agreements with cloud providers. *See, e.g.*, *https://uit.stanford.edu/announcement/2014-09-03-000000*; *https://research.computing.yale.edu/services/cloud-environments*.

CASE STUDY: XSEDE

The Extreme Science and Engineering Discovery Environment (XSEDE) is an NSF-funded organization that integrates and coordinates the sharing of advanced digital services such as supercomputers and high-end visualization and data analysis resources.[134] XSEDE is a collaborative partnership of 19 institutions, or "Service Providers," many of which are nonprofits or supercomputing centers at universities and provide computing facilities for XSEDE researchers.[135] XSEDE supports work from a wide variety of fields, including the physical sciences, life sciences, engineering, social sciences, the humanities, and the arts.[136] XSEDE allocations are available to any researcher or educator at a U.S. academic, nonprofit research, or educational institution, not including students.[137] However, researchers can share their allocations by establishing user accounts with other collaborators, including students.[138]

Researchers have two different paths to requesting allocations: "Startup Allocation" and "Research Allocation." Startup Allocations apportion XSEDE resources for small-scale computational activities.[139] Startup Allocations are one of the fastest ways to gain access to and start using XSEDE resources, as requests are typically reviewed and awarded within two weeks.[140] Startup Allocation requests also require minimal documentation: the project's abstract and the researchers' curriculum vitae (CV).[141] Startup Allocations typically last for one year, but requests supported by merit-reviewed grants can ask for allocations that last up to three years. Researchers can also submit renewal requests if their work needs ongoing low-level resources.[142]

---

[134] *What We Do*, XSEDE, https://www.xsede.org/about/what-we-do (last visited Sept. 19, 2021).
[135] *XSEDE Overall Organization*, XSEDE WIKI, https://confluence.xsede.org/display/XT/XSEDE+Overall+Organization (last visited Sept. 19, 2021).
[136] *XSEDE Allocations Info & Policies*, XSEDE, https://portal.xsede.org/allocations/policies (last visited Sept. 19, 2021).
[137] *Id.*
[138] *Id.*
[139] *Startup Allocations*, XSEDE, https://portal.xsede.org/allocations/startup (last visited Sept. 19, 2021).
[140] *Id.*
[141] *Id.*
[142] *Id.*

For research needs that go beyond the computational limits under a Startup Allocation, researchers must submit a Research Allocation request.[143] XSEDE strongly encourages its users to request a Startup Allocation prior to requesting a Research Allocation, in order to obtain benchmark results and more accurately document their research needs in the Research Allocation.[144] Research Allocation requests must include a host of documents, such as a resource-use plan, a progress report, code performance calculations, CVs, and references.[145] Requests are accepted and reviewed quarterly by the XSEDE Resource Allocations Committee (XRAC), which assesses the proposals' appropriateness of methodology, appropriateness of research plan, efficient use of resources, and intellectual merit.[146]

XSEDE abides by a "one-project rule," whereby each researcher only has one XSEDE allocation for their research activities.[147] For instance, if a researcher has several grants that require computational support, those lines of work should be combined into a single allocation request. This minimizes the effort required by the researcher to submit requests and reduces the overhead of reviewing those requests.

XSEDE also uses a "Campus Champion Program" to streamline access to resources.[148] The Campus Champion Program is a group of over seven hundred "Campus Champions" who are employees or affiliates at over three hundred U.S. colleges, universities, and research-focused institutions.[149] These Campus Champions facilitate and support the use of XSEDE-allocated resources by researchers, educators, and students on their campuses. For instance, the Campus Champions host awareness sessions and training workshops for their institutions' researchers while also capturing information on problems and challenges that need to be addressed by XSEDE resource owners.[150]

---

[143] *Research Allocations*, XSEDE, https://portal.xsede.org/allocations/research (last visited Sept. 19, 2021).
[144] *Id.*
[145] *Id.*
[146] *Id.*
[147] *XSEDE Allocations Info & Policies*, *supra* note 136.
[148] *XSEDE Campus Champions*, *supra* note 118.
[149] *Id.*
[150] *Id.*

> Finally, XSEDE welcomes collaboration opportunities with other members of the research and scientific community.[151] For example, XSEDE assists other organizations in acquiring and operating computing resources and helps to allocate and manage access to those resources. Recently, XSEDE worked with academics and private industry to form the COVID-19 High Performance Computing Consortium, which provides researchers with powerful computing resources to better understand COVID-19 and develop treatments to address infections.[152]

Commercial cloud platforms also provide other advantages to the NRC. The labor of managing, maintaining, and upgrading the hardware behind the NRC would be handled by private parties that already have expertise in running cloud services at scale and have invested billions of dollars into doing it. This arrangement allows researchers access to a greater variety of hardware that is constantly being expanded and upgraded.[153] With a strong economic incentive to keep improving cloud offerings, commercial cloud services offer an assortment of instance types—i.e., the various permutations and combinations of GPU/CPU, memory, storage, and networking specifications that constitute a compute instance—with different hardware at a range of price points. Thus, researchers would have the flexibility to choose both what hardware would best fit the needs of their projects and how best to allocate their limited cloud credits. Researchers could also have access to cutting-edge technology specially designed for AI research, such as chips optimized for training and inference, developed and exclusively used by commercial cloud providers.

Using commercial cloud services for the NRC comes with significant trade-offs, however. While the initial costs of subsidizing

---

[151] *XSEDE as a Collaborator on Proposals*, XSEDE, https://www.xsede.org/about/collaborating-with-xsede (last visited Sept. 19, 2021).
[152] *COVID-19 HPC Consortium*, XSEDE, https://www.xsede.org/covid19-hpc-consortium (last visited Sept. 19, 2021).
[153] Amazon, for example, introduced its P4, P3, and P2 instances in 2020, 1997, and 1996, respectively. Frederic Lardinois, *AWS Launches Its Next-Gen GPU Instances with 8 Nvidia A100 Tensor Core GPUs*, TECHCRUNCH (Nov. 2, 2020), https://social.techcrunch.com/2020/11/02/aws-launches-its-next-gen-gpu-instances/; Ian C. Schafer, *Amazon Elastic Compute Cloud P3 Launched alongside NVIDIA GPU Cloud*, SD TIMES (Oct. 26, 2017), https://sdtimes.com/ai/amazon-elastic-compute-cloud-p3-launched-alongside-nvidia-gpu-cloud/; Jeff Barr, *New P2 Instance Type for Amazon EC2 – Up to 16 GPUs*, AMAZON WEB SERVS. (Sept. 29, 2016), https://aws.amazon.com/blogs/aws/new-p2-instance-type-for-amazon-ec2-up-to-16-gpus/. The introduction years of the P4 and P3 instances line up with the release of NVIDIA's newest general purpose data center GPUs.

cloud credits might be less than building public infrastructure, many studies show that relying on commercial cloud services would likely be much more expensive in the long term.[154] For example, a study of Purdue University's Community Cluster Program shows that the amortized cost of its on-premises cluster over five years is 2.73 times cheaper than using AWS, 3.24 times cheaper than using Azure, and 5.54 times cheaper than using Google Cloud.[155] A similar study at Indiana University estimates that the total investment into its locally-owned supercomputer, Big Red II, is about $10.1 million, while the total cost of a three-year reservation on AWS about $24.9 million.[156] Cost comparisons in other studies are even more dramatic. For instance, a study of the Advanced Research Computing clusters at Virginia Tech shows that the five-year cost for its on-premises cloud is about $15.5 million, while the five-year cost for reserved AWS instances using the same workloads would be about $136.3 million.[157]

What explains these cost disparities? Estimates comparing commercial cloud services to a dedicated HPC cluster show that commercial cloud services are more expensive per compute cycle.[158] At least in part, this is due to the fact that commercial services are optimized for commercial applications. Compute Canada, for example, found that

---

[154] *See, e.g.*, Sarah Wang & Martin Casado, *The Cost of Cloud, a Trillion Dollar Paradox*, ANDREESSEN HOROWITZ (May 27, 2021), https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization/.

[155] Preston Smith et al., *Community Clusters or the Cloud: Continuing Cost Assessment of On-Premises and Cloud HPC in Higher Education*, 2019 PROC. PRAC. & EXPERIENCE ADVANCED RES. COMPUTING ON RISE OF THE MACHS. 1 (2019). The amortized cost includes the annual compute cost, subsidized hardware cost, and power costs, but does not include personnel costs, as such costs are fixed and would be recurred regardless of whether a cluster existed physically on-prem or on the cloud. *Id.*

[156] Craig A. Stewart et al., *Return on Investment for Three Cyberinfrastructure Facilities: A Local Campus Supercomputer; the NSF-Funded Jetstream Cloud System; and XSEDE*, 11 INT'L CONF. ON UTILITY & CLOUD COMPUTING 223 (2018).

[157] Srijith Rajamohan & Robert E. Settlage, *Informing the On/Off-prem Cloud Discussion in Higher Education*, 2020 PRAC. & EXPERIENCE ADVANCED RES. COMPUTING 64 (2020). The cost sources include hardware, software services, software administration, electricity, and facilities but do not include computational scientists support, scientific software licenses, and data transfer costs. The study is also limited to Virginia Tech's particular cloud workload.

[158] Jennifer Villa & Dave Troiano, *Choosing Your Deep Learning Infrastructure: The Cloud vs. On-Prem Debate*, DETERMINED A.I. (July 30, 2020), https://determined.ai/blog/cloud-v-onprem/; *Is HPC Going to Cost Me a Fortune?*, INSIDEHPC, https://insidehpc.com/hpc-basic-training/is-hpc-going-to-cost-me-a-fortune/ (last visited Mar. 2, 2022).

building its own infrastructure was cheaper than using commercial services because it did not have the same core use needs as commercial customers, a trade-off that gained its system more computing power at the expense of availability.[159] Although the analysis was published in 2016, Compute Canada's own benchmarking of costs concluded:

> "Currently, it is far more cost effective for the Compute Canada federation to procure and operate in-house cyberinfrastructure than to outsource to commercial cloud providers. Cloud-based costs ranged from 4x to 10x more than the cost of owning and operating our own clusters. Some components were dramatically more expensive, notably persistent storage which was 40x the cost of Compute Canada's storage."[160]

Ultimately, the cost difference between commercial cloud services and HPC systems depends on how often and how efficiently the HPC system is used. We provide a cost calculation that updates Compute Canada's below, arriving at cost differentials of comparable magnitude. Commercial cloud instances with comparable hardware under constant usage, even with substantial discounts, would be significantly more expensive over time for the NRC than a dedicated HPC system. Bringing the cost of commercial cloud services under that of an HPC system would require policymakers to either negotiate exceptionally high discounts with commercial cloud providers or make major sacrifices in hardware speed or overall scale of the NRC. A similar cost calculation is also what led Stanford University to simultaneously invest in both on-premises hardware and a commercial cloud-based solution for its Population Health Sciences initiative (see box case study in Section 3). The most common practice across NSF centers, such as the XSEDE initiative (see box case study below), is also to build infrastructure instead of relying on commercial cloud credits, due to these cost considerations.

Finally, relying on the commercial cloud may raise questions about industry consolidation. There are two main answers to this question. One is that building a dedicated, publicly owned HPC clusters would require purchasing sophisticated hardware from existing industry players, which also exist in concentrated industries. In other words, it is

---

[159] Interview with Suzanne Talon, Regional Director, COMPUTE CAN. (Jan. 14, 2021).
[160] COMPUTE CAN., CLOUD COMPUTING FOR RESEARCHERS (Dec. 2016).

difficult to imagine no involvement of private industry under either option. Another major constraint lies in time: a fully mature, public infrastructure NRC could not be stood up overnight. Moreover, a publicly owned cloud would still likely require a major technology company to build the infrastructure under contract, as is the case for National Labs, or using a grant, as is the case for XSEDE.

2. Public Infrastructure

Building a new HPC cluster would be a bespoke solution, tailored to fit the NRC's specific compute needs. This approach would be a relatively well-explored territory for the federal government.[161] The U.S. Department of Energy (DOE) and the U.S. Department of Defense (DOD) already regularly contract with a handful of companies to build HPC clusters every few years.[162] The DOE itself already uses two of the three fastest HPC clusters in the world and recently funded the development of two new supercomputers that, when completed, will be the world's fastest by a significant margin.[163] The National Science Foundation commonly issues grants for the construction of high-performance computing infrastructure.[164] Given this familiarity, policymakers would

---

[161] *US Plans $1.8 Billion Spend on DOE Exascale Supercomputing*, HPCWIRE (Apr. 11, 2018), https://www.hpcwire.com/2018/04/11/us-plans-1-8-billion-spend-on-doe-exascale-supercomputing/; *Federal Government*, ADVANCED HPC, https://www.advancedhpc.com/pages/federal-government; *United States Continues To Lead World In Supercomputing*, ENERGY.GOV, https://www.energy.gov/articles/united-states-continues-lead-world-supercomputing (last visited Mar. 2, 2022); *High Performance Computing*, ENERGY.GOV, https://www.energy.gov/science/initiatives/high-performance-computing (last visited Mar. 2, 2022).

[162] *See, e.g., DOE Announces Five New Energy Projects at LLNL*, LLNL (Nov. 13, 2020), https://www.llnl.gov/news/doe-announces-five-new-energy-projects-llnl; *New HPCMP System at the AFRL DSRC DoD Supercomputing Resource Center to Provide over Nine PetaFLOPS of Computing Power to Address Physics, AI, and ML Applications for DoD Users*, DOD HPC, https://www.hpc.mil/images/hpcdocs/newsroom/21-19_TI-21_web_announcement_AFRL_DSRC.pdf; Press Release, DOD HPC, Public Announcement (DD-LA-(AR) 1279) (May 5, 2021), https://www.hpc.mil/images/hpcdocs/newsroom/awards_and_press/HC101321D0002_PUBLIC_ANNOUNCEMENT_20210505.pdf.

[163] Devin Coldewey, *$600M Cray Supercomputer Will Tower Above the Rest — to Build Better Nukes*, TECHCRUNCH (Aug. 13, 2019), https://social.techcrunch.com/2019/08/13/600m-cray-supercomputer-will-tower-above-the-rest-to-build-better-nukes/; Press Release, Oak Ridge Nat'l Lab'y, CORAL-2 RFP (Apr. 9, 2018), https://procurement.ornl.gov/rfp/CORAL2/.

[164] *See, e.g.*, *NSF Funds Five New XSEDE-Allocated Systems*, XSEDE (Aug. 10, 2020), https://www.xsede.org/-/nsf-funds-five-new-xsede-allocated-systems.

have reasonable estimates for how much a new HPC cluster for the NRC would cost and would already have relationships with the companies that would submit bids for the contract.

The hardware costs for such compute scale are, of course, substantial.[165] For example, the IBM supercomputer used at Oak Ridge National Laboratory (ORNL)—known as "Summit"—cost $200 million.[166] At the time of its completion in 2018, Summit was the fastest supercomputer in the world and, as of 2020, is still the second-fastest.[167] Frontier, the new Cray supercomputer being built at ORNL in 2021, cost $500 million. When completed, it is anticipated to be the fastest supercomputer in the world at "up to 50 times" faster than Summit.[168] Nonetheless, these large up-front costs could come with the benefit of computing infrastructure specifically designed for AI research and the NRC's needs. Such a system would be more efficient in cost per cycle over the long term than subsidizing commercial cloud services. The NRC could also expand and upgrade multiple clusters over time to meet the changing needs and scope of the program.

In addition, a dedicated cluster for the NRC has the advantage of giving the federal government greater control over computational resources (e.g., reducing uncertainty over the products and platforms, such as the sudden deprecation of required APIs). This level of control over the hardware also allows policymakers greater flexibility with NRC operations. Taking the public infrastructure approach (i.e., "making" not "buying") comes with several significant trade-offs to weigh against the policy goals of the NRC. First, building a new HPC cluster would take about two years, in addition to the time it takes to solicit and evaluate

---

[165] Timothy Prickett Morgan, *Bending the Supercomputing Cost Curve Down*, THE NEXT PLATFORM (Dec. 2, 2019), http://www.nextplatform.com/2019/12/02/bending-the-supercomputing-cost-curve-down/; Ben Dickson, *The GPT-3 Economy*, TECHTALKS (Sept. 21, 2020), https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/.

[166] Elijah Wolfson, *The US Passed China with a Supercomputer Capable of as Many Calculations Per Second as 6.3 Billion Humans*, QUARTZ (June 9, 2018), https://qz.com/1301510/the-us-has-the-worlds-fastest-supercomputer-again-the-200-petaflop-summit/.

[167] *November 2020*, TOP500 (Nov. 2020), https://www.top500.org/lists/top500/2020/11/.

[168] *U.S. Department of Energy and Cray to Deliver Record-Setting Frontier Supercomputer at ORNL*, OAK RIDGE NAT'L LAB'Y (May 7, 2019), https://www.ornl.gov/news/us-department-energy-and-cray-deliver-record-setting-frontier-supercomputer-ornl.

proposals from potential contractors.[169] If the NRC hopes to quickly stimulate and help democratize AI research in the U.S., such a timeline for the program would not be ideal, given how quickly AI discoveries advance. Of course, contracting with cloud vendors or issuing grants for the construction of supercomputers would also require a process. Yet, building a cluster could raise more challenging contracting issues, such as budget overruns and project delays.[170] Contractors' experience with building this type of hardware may help mitigate some of these concerns, as well as their self-interest in being considered for future government contracts. But the risks are nonetheless still present.

Second, the usability and the feature set of the software stack for public infrastructure is by no means proven. One of the most common hurdles to researcher adoption of cloud computing lies in the usability of systems,[171] and public infrastructure has less of a track record of easing that onboarding path at the contemplated scale. This is why we recommend a pilot to assess whether a national HPC center can be administered in a way to ensure the ease of cloud transition and software stack that researchers have become accustomed to with private providers.

Third, policymakers would also need to account for the costs of maintaining and administering the system.[172] They would need to find facilities to house and manage the hardware and to account for the high energy costs of running an HPC cluster, as well as disaster prevention and recovery cost.[173] These costs are significant. In 2021, the Oak Ridge

---

[169] Coury Turczyn, *Building an Exascale-Class Data Center*, OAK RIDGE LEADERSHIP COMPUTING FACILITY (Dec. 14, 2020), https://www.olcf.ornl.gov/2020/12/11/building-an-exascale-class-data-center/.
[170] Don Clark, *Intel Slips, and a High-Profile Supercomputer Is Delayed*, N.Y. TIMES (Aug. 27, 2020), https://www.nytimes.com/2020/08/27/technology/intel-aurora-supercomputer.html; Mila Jasper, *10 of 15 of DOD's Major IT Projects Are Behind Schedule, GAO Found*, NEXTGOV (Jan. 4, 2021), https://www.nextgov.com/it-modernization/2021/01/10-15-dods-major-it-projects-are-behind-schedule-gao-found/171155/.
[171] *See* Nattakarn Phaphoom et al., *A Survey Study on Major Technical Barriers Affecting the Decision to Adopt Cloud Services*, 103 J. SYS. & SOFTWARE 167, 171-72 (2015) (describing data portability, integration with existing systems, migration complexity, and availability as major barriers to cloud adoption); Abdulrahman Alharthi et al., *An Overview of Cloud Services Adoption Challenges in Higher Education Institutions,* 2 WORKSHOP ON EMERGING SOFTWARE AS A SERV. & ANALYTICS 102, 107-08 (2015) (acknowledging the low rate of cloud computing adoption in higher education and emphasizing that bolstering both the perceived ease of use and the actual usefulness of cloud computing can increase the adoption rate).
[172] *See* DEP'T OF ENERGY, FY 2021 CONG. BUDGET REQUEST VOLUME 4: SCIENCE (2020).
[173] JOE WEINMAN, CLOUDONOMICS: THE BUSINESS VALUE OF CLOUD COMPUTING (2012).

Leadership Computing Facility requested $225 million to operate all of its systems.[174] The Argonne Leadership Computing Facility, in turn, requested $155 million.[175] Furthermore, the lifecycle of DOE HPC systems has traditionally been about seven years, after which new systems are built and old ones decommissioned.[176] While it is uncertain what the lifespan of newer systems will be, this seven-year figure would lead us to argue that the NRC should expect to either upgrade its systems or build new ones with some degree of regularity.

Last, giving the federal government greater control over the computing resources would not immediately make the NRC safe from attacks.[177] As with using commercial cloud infrastructure, security will primarily be contingent on the NRC's implemented data access model.[178] We discuss security issues in depth in Section 8.

CASE STUDY: Fugaku

In 2014, Japan's Ministry of Education, Culture, Sports, Science, and Technology launched a public-private partnership between the government-funded Riken Institute, the Research Organization for Information Science and Technology (RIST), and Fujitsu to create the supercomputer successor to the K computer that supports a wide range

---

[174] OLCF supports and manages ORNL's supercomputing resources, including Summit and eventually Frontier. This figure accounts for "operations and user support at the LCF facilities–including power, space, leases, and staff. Id. at 37-38.

[175] ACLF supports and manages Argonne National Laboratory's computing resources, including the Theta system and, later this year, the new Aurora computer, another DOE exascale HPC system. *Id.*

[176] *See* Turczyn, *supra* note 169. OLCF operated its Titan HPC system for 7 years. ACLF also operated its Mira HPC system for 7 years. Jared Sagoff & Jim Collins, *Argonne's Mira Supercomputer to Retire After Years of Enabling Groundbreaking Science*, HPCWIRE (Dec. 20, 2019), https://www.hpcwire.com/2019/12/20/argonnes-mira-supercomputer-to-retire-after-years-of-enabling-groundbreaking-science/. If still operational, these systems would rank about the 19th and 29th fastest in the world, respectively. *Cf. November 2020*, *supra* note 167, *with June 2019*, TOP500 (June 2019), https://www.top500.org/lists/top500/list/2019/06/.

[177] *See, e.g.*, Kim Zetter, *Top Federal Lab Hacked in Spear-Phishing Attack*, WIRED (Apr. 20, 2011), https://www.wired.com/2011/04/oak-ridge-lab-hack/; Natasha Bertrand & Eric Wolff, *Nuclear Weapons Agency Breached Amid Massive Cyber Onslaught*, POLITICO (Dec. 17, 2020), https://www.politico.com/news/2020/12/17/nuclear-agency-hacked-officials-inform-congress-447855; Ryan Lucas, *List of Federal Agencies Affected By a Major Cyberattack Continues to Grow*, NPR (Dec. 18, 2020), https://www.npr.org/2020/12/18/948133260/list-of-federal-agencies-affected-by-a-major-cyberattack-continues-to-grow.

[178] We discuss data access models in Section 3.

of scientific and societal applications.[179] The result was Fugaku, which was named the world's fastest supercomputer in 2020.[180]

The technical aim of Fugaku was to be one hundred times faster than the previous K computer, with a performance of 442 petaFLOPS in the TOP500's FP64 high-performance LINPACK benchmark.[181] It currently runs 2.9 times faster than the next fastest system (IBM Summit)[182] and is composed of slightly over 150,000 connected CPUs, with each CPU using ARM-licensed computer chips.[183] Despite having around 1.9 times more parts than its K computer predecessor, Fugaku was finished in three fewer months.[184] The six-year budget for Fugaku was around $1 billion.[185]

RIST solicited proposals for usage through the "Program for Promoting Research on the Supercomputer Fugaku." Under the program, Fugaku has already been used to study the effect of masks and respiratory droplets in order to inform Japanese policy during the COVID-19 pandemic.[186] For FY 2021, 74 public and industrial projects were selected for full-scale access to Fugaku.[187] Currently, RIST is still requesting proposals that fall under specific categories of usage, and any interested researcher may apply.[188]

---

[179] *See Ongoing Projects*, RIKEN CTR. FOR COMPUTATIONAL SCI., https://www.r-ccs.riken.jp/en/fugaku/research/covid-19/projects/ (last visited Mar. 2, 2022).
[180] *Fugaku Retains Title as World's Fastest Supercomputer*, HPCWIRE (Nov. 17, 2020), https://www.hpcwire.com/off-the-wire/fugaku-retains-title-as-worlds-fastest-supercomputer/.
[181] *See November 2020*, *supra* note 167.
[182] *Id.*
[183] *Behind the Scenes of Fugaku as the World's Fastest Supercomputer*, FUJITSU (Feb. 2, 2021), https://blog.global.fujitsu.com/fgb/2021-02-02/behind-the-scenes-of-fugaku-as-the-worlds-fastest-supercomputer-1manufacturing/.
[184] *Id.*
[185] Don Clark, *Japanese Supercomputer Is Crowned World's Speediest*, N.Y. TIMES (June 22, 2020), https://www.nytimes.com/2020/06/22/technology/japanese-supercomputer-fugaku-tops-american-chinese-machines.html.
[186] Justin McCurry, *Non-Woven Masks Better to Stop Covid-19, Says Japanese Supercomputer*, THE GUARDIAN (Aug. 26, 2020), https://www.theguardian.com/world/2020/aug/26/non-woven-masks-better-to-stop-covid-19-says-japanese-supercomputer.
[187] *Fujitsu and RIKEN Complete Joint Development of Japan's Fugaku, the World's Fastest Supercomputer*, FUJITSU (Mar. 9, 2021), https://www.fujitsu.com/global/about/resources/news/press-releases/2021/0309-02.html.
[188] *Id.*

3. Cost Comparison

To conclude this section, we provide a rough cost comparison between a leading commercial cloud service and a dedicated government HPC system (IBM Summit) (see Appendix A for details). We refer the reader to substantial work that has been published on the economics of cloud computing for a fuller analysis, much of which emphasizes the variance in computing demand.[189]

Building standalone public infrastructure is projected to be less expensive than implementing the NRC through a vendor contracting arrangement over five years. At a 10 percent discount on standard rates over five years, and under constant usage, AWS's more powerful cloud-computing option (known as P3 instances) could cost 7.5 times as much as Summit's total estimated costs, using comparable hardware. We use a 10 percent discount that was negotiated by a major research university with a commercial cloud provider. In contrast, the government would need to negotiate an 88 percent discount for AWS to be cost-competitive with a dedicated HPC cluster in the long run. Even in a scenario where NRC usage fluctuates dramatically, commercial cloud computing could cost 2.8 times Summit's estimated cost. While variability in usage factors heavily into these estimates, the use of schedulers can contribute to a leveling-out of demand.[190]

These cost estimates have important limitations. First, government may be able to negotiate the cost down. We have used as a benchmark one major university's enterprise agreement with AWS, which provides a 10 percent discount, relative to market rates. But unless the negotiated discount is for orders of larger magnitude, the commercial cloud will remain significantly more expensive. Second, these cost estimates primarily focus on computing.[191] As Compute Canada's analysis showed, the cost difference in storage was even greater. Third, the use of commercial rates is likely *more* favorable to cloud vendors, as

---

[189] *See, e.g.*, ROLF HARMS & MICHAEL YAMARTINO, THE ECONOMICS OF THE CLOUD (2010); Rajamohan & Settlage, *supra* note 157; Byung Chul Tak et al., *To Move or Not to Move: The Economics of Cloud Computing*, 3 USENIX CONF. ON HOT TOPICS IN CLOUD COMPUTING 1 (2011); Edward Walker, Walter Brisken & Jonathan Romney, *To Lease or Not to Lease from Storage Clouds*, 43 COMPUT. 44 (2010).

[190] *See, e.g.*, Di Zhang et al., *RLScheduler: An Automated HPC Batch Job Scheduler Using Reinforcement Learning*, ARXIV (Sept. 2, 2020), https://arxiv.org/pdf/1910.08925.pdf.

[191] For instance, we have not been able to identify good estimates of electricity and cooling costs for DOE supercomputers.

government security standards typically increase rates due to regulatory requirements. For instance, a "data sovereignty" requirement for data and hardware to reside within the United States, or private cloud requirements for certain agency datasets, may increase the cost of commercial cloud computing significantly. Fourth, this simple cost comparison is static, and does not reflect changes in hardware costs and pricing structures that are likely to occur over a five-year period under rapidly changing market conditions. But, if the NRC in fact scales, systems would be procured incrementally over time, upgrading available resources and providing options at different price points, similar to current commercial options. Last, as noted above, these cost estimates take into account maintenance as budgeted for the Summit, but may not take into account all such non-hardware costs, which is why we recommend a pilot to explore the ability to open up government computing facilities to NRC users.

In short, we offer this simple comparison to highlight some of the salient cost considerations to the make-or-buy decision, which arrives at a very similar conclusion to the analysis done by Compute Canada.

CASE STUDY: Compute Canada

Compute Canada formed in 2006 as a partnership between Canada's regional academic HPC organizations to share infrastructure across Canada.[192] The organization's stated mission is to "enable excellence in research and innovation for the benefit of Canada by effectively, efficiently, and sustainably deploying a state-of-the-art advanced research computing network supported by world-class expertise."[193]

Compute Canada's infrastructure includes five HPC systems that are hosted at research universities across Canada.[194] From 2015 to 2019, Compute Canada used about C$125 million in funding to build four of

[192] HUGH COUCHMAN ET AL., COMPUTE CANADA — CALCUL CANADA: A PROPOSAL TO THE CANADA FOUNDATION FOR INNOVATION 58 (2006).
[193] *About*, COMPUTE CAN., https://www.computecanada.ca/about/ (last visited Mar. 2, 2022).
[194] *National Systems*, COMPUTE CAN., https://www.computecanada.ca/techrenewal/national-systems/ (last visited Mar. 2, 2022).

these systems.[195] It also investigated using commercial cloud resources instead of building these new systems.[196] However, it ultimately concluded that relying on commercial cloud providers would be significantly more expensive and could not provide the desired latency for large-scale, data-intensive research.[197] In 2018, Compute Canada requested C$61 million to fund its operations, budgeting C$41 million for operating its HPC systems and C$20 million for support, training, and outreach.[198] Demand for Compute Canada's HPC resources far exceeds the infrastructure's current capacity and is expected to keep growing.[199] In 2018, Compute Canada estimated it would need about C$90 million per year over five years to invest in expanding infrastructure to the point where it could meet projected demands.[200]

About 16,000 researchers from all scientific disciplines use Compute Canada's infrastructure to support their work.[201] Compute Canada distributes its resources in two ways. First, Principal Investigators and sponsored users may request a scheduler-unprioritized resource allocation for their research group.[202] Compute Canada finds that many research groups can meet their compute needs this way.[203] Alternatively, researchers who need more or prioritized resources may submit a project proposal to the annual "Research Allocation Competitions."[204] Submitted proposals go through a scientific peer review and a technical staff review to rate their merits.[205] Scientific review examines the scientific excellence and feasibility of the specific research project, the appropriateness of the resources requested to achieve the project's objectives, and the likelihood that the resources requested will be

---

[195] *Compute Canada Technology Briefing,* COMPUTE CAN. (Nov. 2017), https://www.computecanada.ca/wp-content/uploads/2015/02/Technology-Briefing-November-2017.pdf.
[196] *Cloud Computing for Researchers*, COMPUTE CAN. (Dec. 2016), https://www.computecanada.ca/wp-content/uploads/2015/02/CloudStrategy2016-2019-forresearchersEXTERNAL-1.pdf.
[197] *Id.*
[198] COMPUTE CAN., BUDGET SUBMISSION 2018 5 (2018).
[199] Compute Canada projected it had only met about 55 percent of total demand for CPU compute hours in 2018. *Id.*
[200] *Id.*
[201] COMPUTE CAN., ANNUAL REPORT 2019-2020 4 (2020).
[202] *Rapid Access Service*, COMPUTE CAN., https://www.computecanada.ca/research-portal/accessing-resources/rapid-access-service/ (last visited Mar. 2, 2022).
[203] *Id.*
[204] *Resource Allocation Competitions*, *supra* note 119.
[205] *Id.*

efficiently used.[206] This review is conducted on a volunteer basis by 80 discipline-specific experts from Canadian academic institutions.[207]

Technical review is conducted by Compute Canada staff itself, who verify the accuracy of the computational resources needed for each project, based on the technical requirements outlined in the application, and makes recommendations about which resources should be allocated to meet the project's needs.[208] In 2021, Compute Canada received 651 applications to the Research Allocation Competition and fully reviewed all applications in the span of five months.[209]

III. SECURING DATA ACCESS

After compute resources, the next critical design decision for the NRC is how to both store datasets and provide its users access to them: the "data access" goal of the NRC. Indeed, as articulated in the original NRC call to action, government agencies should "redouble their efforts to make more and better-quality data available for public research at no cost," as it will "fuel" unique breakthroughs in research.[210] Investigating some of the most socially meaningful problems hinges on large but inaccessible datasets in the public sector. From climate data housed by the National Oceanic and Atmospheric Administration (NOAA), health data from the country's largest integrated healthcare system in the Department of Veterans Affairs (VA), or employment data in the Department of Labor (DOL), such data could fuel both fundamental research using AI and refocus efforts away from consumer-focused projects (e.g., optimizing advertising) to more socially pressing topics (e.g., climate change).

As noted in the congressional charge, facilitating broad data access is a crucial pillar of the NRC. Importantly, as we discuss below, we limit the scope of our recommendations to facilitating access to public sector government data, which, as a condition of accessing government administrative data, NRC researchers should only use for academic research purposes. NRC users should also be able to compute

---

[206] *Id.*
[207] *Id.*
[208] *Id.*
[209] *2021 Resource Allocations Competition Results*, *supra* note 119.
[210] *National Research Cloud Call to Action*, STAN. U. INST. FOR HUM.-CENTERED A.I. (2020), https://hai.stanford.edu/national-research-cloud-joint-letter.

on any private dataset available to them. There are available mechanisms for sharing such datasets, but we identify the NRC's major challenge as providing access to previously unavailable government data.

Government data is intentionally decentralized. By design of the Privacy Act of 1974, there is no centralized repository for U.S. government data or a core method for linking data across government agencies.[211] The result is a sprawling, decentralized data infrastructure with widely varying levels of funding, expertise, application of standards, and access and sharing of policies. Thus, the NRC will have to develop a unified data strategy that can work with a wide range of agencies, unevenly adopted security standards, and within existing data privacy legislation.

Previous efforts have sought to improve access to and sharing of federal data, both between agencies and with external researchers, but there are still significant barriers to enabling AI research access of the kind that the NRC demands.[212] By linking data governance policies with access to compute, building on existing successful models, and working with agencies to create interoperable systems that satisfy security and privacy concerns, the NRC can enable increased access to data that will aid AI researchers in answering pressing scientific and social questions and increase AI innovation.[213]

We will first explain why the NRC should focus its efforts on facilitating federal government data sharing rather than private-sector data sharing. We then examine how and why the status quo for federal data sharing fails to realize the massive potential of government data. While the concept of centralizing disparate data sources to unlock research insights is not new,[214] there are unique challenges for doing so within the context of the NRC. We will also discuss the key elements of our proposed model: (1) the use of FedRAMP as a system for categorizing datasets based on their sensitivity, and for modifying

---

[211] We discuss the Privacy Act and privacy considerations in more detail in Section Five.

[212] O'Hara & Medalia, *supra* note 13, at 140-41; *see also* PRESIDENT'S MGMT. AGENDA, FEDERAL DATA STRATEGY 2020 ACTION PLAN (2020), https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf.

[213] Improved data access would, as we describe below, also promote evidence-based policymaking and improve trust in science (as data access makes replication efforts much easier).

[214] *See, e.g.*, NICK HART & NANCY POTOK, MODERNIZING U.S. DATA INFRASTRUCTURE: DESIGN CONSIDERATIONS FOR IMPLEMENTING A NATIONAL SECURE DATA SERVICE TO IMPROVE STATISTICS AND EVIDENCE BUILDING (2020).

access to them through tiered credentials for NRC users; (2) the promotion of interagency standardization and harmonization efforts to modernize data-sharing practices; and (3) strategic considerations regarding how to sequence efforts in streamlining access to particular datasets.

The case studies included throughout this Section were chosen as exemplars of successful data-sharing initiatives[215] and to illustrate the range of available design decisions. While each case study provides a unique glimpse into different approaches, some common themes emerge. First, many of the data-sharing entities we studied not only have a single point of entry for researchers to request access, but also allow government agencies to retain some control over access requirements to their data. As we discuss below, this conception of the NRC as a data intermediary would provide real benefits in streamlining data access while still maintaining trust among agencies that wish to protect their data. Second, some initiatives use funding and personnel training as carrots to incentivize agencies to engage in data sharing. The NRC can learn from these initiatives in formulating its own set of incentives for agencies.

A. Private Data Sharing

Should the NRC affirmatively facilitate private dataset sharing? While there are definite benefits to providing researchers with access to private data,[216] the NRC will have its largest impact by focusing its efforts first on mechanisms to access and share government data.

As an initial matter, a variety of mechanisms for general data sharing already exist.[217] Private sector stakeholders, moreover, can and

---

[215] These initiatives are successful in that they are sustainable and have been used by researchers to access multi-agency government data. The only exception is the National Secure Data Service (NSDS), which has not yet been implemented. We discuss the NSDS alongside the Census Bureau and the Evidence-Based Policy-Making Act of 2018 below. Importantly, our focus in these case studies is not to evaluate their efforts or measure their exact levels of success but to identify and understand some of the differences and similarities in the range of data-sharing efforts.

[216] For instance, private sector data may facilitate research regarding social media use, internet behavior, or fill in gaps for federal statistics research through big data analysis. *See* ROBERT M. GROVES & BRIAN A. HARRIS-KOJETIN, INNOVATIONS IN FEDERAL STATISTICS: COMBINING DATA SOURCES WHILE PROTECTING PRIVACY 7 (2017).

[217] *See, e.g., National Data Service,* THE NAT'L DATA SERV., http://www.nationaldataservice.org (last visited Feb. 19, 2022); *The Open Science Data Cloud,* OPEN SCI. DATA CLOUD, https://www.opensciencedatacloud.org (last

have often built their own in-house platforms to allow access to approved datasets while minimizing intellectual property concerns,[218] or to provide access to their application programming interfaces (APIs) to make open-source data more easily accessible.[219] By focusing on providing access to public sector data, notably administrative data that is traditionally inaccessible to most researchers,[220] the NRC would play a unique and pertinent role for researchers across disciplines without having to deal with complex private-sector data concerns or the need to incentivize participation by non-government actors.

Complex intellectual property concerns would arise from the NRC permitting, facilitating, or even requiring, private sector stakeholders and independent researchers to share their private data freely alongside public sector data. First, this would involve complex questions regarding what licenses should be available or mandated for NRC users in order to encourage data sharing, despite apprehensions of how such sharing may affect future profitability and commercialization. While mandating an open-source (e.g., Creative Commons) license would benefit researchers most by providing the broadest access to data and would benefit NRC administrators by removing some possible IP infringement concerns, private sector stakeholders may feel deterred from uploading as a result. Conversely, if users have a choice to adopt a license that allows them to preserve their IP rights, private sector stakeholders may feel more comfortable sharing their data, but this would shift some liability to users—or to the NRC itself—by relying on users to abide by the license. This would involve an emphasis on

---

visited Feb. 19, 2022); *Harvard Dataverse*, HARV. DATAVERSE, https://dataverse.harvard.edu (last visited Feb. 19, 2022); *FigShare*, https://figshare.com (last visited Feb. 19, 2022).

[218] Meta Data for Good provides access to a variety of libraries, via in-house platforms. *See, e.g., Meta Data For Good*, META (2020), https://dataforgood.fb.com/ (last visited Feb. 16, 2022); *What is the Meta Ad Library and How do I Search It?*, META (2021), https://www.facebook.com/help/259468828226154 (last visited Feb. 16, 2022); *Facebook Disaster Maps Methodology*, META (May 15, 2019), https://research.facebook.com/blog/2017/6/facebook-disaster-maps-methodology/.

[219] For example, Twitter has a Developer Portal that provides access to their API to allow researchers to use user data for noncommercial purposes. *See Twitter Developers*, TWITTER (2021), https://developer.twitter.com/en/portal/petition/academic/is-it-right-for-you (last visited Feb. 16, 2022); *Take Your Research Further with Twitter Data*, TWITTER (2021), https://developer.twitter.com/en/solutions/academic-research (last visited Feb. 16, 2022). Thus, uploading Twitter data to a separate Cloud may provide few incentives to researchers who can use the API route.

[220] *See* GROVES, *supra* note 216, at 31-42.

enforcement, ranging from explanations and user disclaimers to the industry standard of a full-blown notice-and-takedown system.

Data owners may want to prevent the uploading of copyrighted works by, for instance, having the NRC itself assess whether private data is already protected by copyright. Industry standards for conducting data diligence, using manual or automated tools, would either be very labor-intensive[221] or prohibitively expensive.[222] Even if these industry standards were met, researchers may find an NRC data-sharing platform duplicative.

None of the above would prevent researchers from using NRC *compute* resources on their own private datasets. Like current cloud providers, the NRC can stipulate in an End User Licensing Agreement (EULA) that researchers must agree they own the intellectual property rights on the data they are using.[223] This EULA can also assign liability to the end-user, rather than the NRC, for any use of data that is encumbered by existing IP provisions. Additionally, the discussion above pertains to whether researchers should be required to share their *private data*, not to whether researchers should be required to share the *outputs* of their research conducted on the NRC. The latter point is discussed in Section 9.

---

[221] *See* JENNIFER M. URBAN, JOE KARAGANIS & BRIANNA M. SCHOFIELD, NOTICE & TAKEDOWN IN EVERYDAY PRACTICE 39 (2017) (illustrating the difficulty that online service providers face in manually evaluating a large volume of data for potential infringement; for example, one online service provider explained that "out of fear of failing to remove infringing material, and motivated by the threat of statutory damages, its staff will take "six passes to try to find the [identified content]."); *see also* Letter from Thom Tillis, Marsha Blackburn, Christopher A. Coons, Dianne Feinstein et. al, to Sundar Pichai, Chief Executive Officer, Google Inc. (Sept. 3, 2019), https://www.ipwatchdog.com/wp-content/uploads/2019/09/9.3-Content-ID-Ltr.pdf ("We have heard from copyright holders who have been denied access to Content ID tools, and as a result, are at a significant disadvantage to prevent repeated uploading of content that they have previously identified as infringing. They are left with the choice of spending hours each week seeking out and sending notices about the same copyrighted works, or allowing their intellectual property to be misappropriated.").

[222] To illustrate the costs of implementing Content ID on a large-scale platform, Google announced in a report in 2016 that YouTube had invested more than $60 million in Content ID. *See* GOOGLE, HOW GOOGLE FIGHTS PIRACY 6 (2018), https://www.blog.google/documents/25/GO806_Google_FightsPiracy_eReader_final.pdf/.

[223] *See, e.g., AWS Customer Agreement*, AMAZON (Nov. 30, 2020), https://aws.amazon.com/agreement/.

B. The Current Patchwork System for Accessing Federal Data

The NRC could play a pivotal role in streamlining access to government data in a system that is currently decentralized.[224] In some cases, agencies may simply lack a standardized method for sharing data.[225] Due to perceived legal constraints, risks, or security concerns, agencies often have little practical incentive to share their data.[226] Successful examples of researchers gaining access to government data from individual agencies frequently rely on the researchers having personal relationships with administrators, and a willingness on the part of the administrator to push against these constraints in service of the research project.[227] While this relationship-based process has produced some successes,[228] the far more common outcome is that data is simply not shared or accessed by researchers.[229] Indeed, one government official indicated that overcoming the obstacles to making certain government data available for research was the greatest challenge in a lengthy career.

Agencies typically require the recipient of the data to abide by a data-use agreement (DUA). These DUAs prescribe such limitations on data usage as the duration of use, the purpose of use, and guarantees on the privacy and security of data.[230] However, DUAs suffer from a central problem: the process for negotiating DUAs is highly fragmented and inconsistent across government agencies, drastically increasing the

---

[224] For instance, across the twenty-nine distinct agencies in the Department of Health and Human Services (HHS), data "are largely kept in silos with a lack of organizational awareness of what data are collected across the Department and how to request access. Each agency operates within its own statutory authority and each dataset can be governed by a particular set of regulations." U.S. DEP'T OF HEALTH & HUM. SERV., THE STATE OF DATA SHARING AT THE U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES 4 (2018), https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf.

[225] *See, e.g., id.* at 8 ("HHS lacks consistent and standardized processes for one agency to request data from another agency.").

[226] O'Hara & Medalia, *supra* note 13, at 140-41.

[227] *See id.* at 142 ("Most [data-sharing] agreements rely heavily on interpersonal relationships and informal quid pro quo arrangements, handling data requests in a less centralized fashion.").

[228] Jeffrey Mervis, *How Two Economists Got Direct Access to IRS Tax Records*, SCIENCE (May 22, 2014), https://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records.

[229] *See* ROBERT M. GROVES & ADAM NEUFELD, ACCELERATING THE SHARING OF DATA ACROSS SECTORS TO ADVANCE THE COMMON GOOD 17 (2017).

[230] *See, e.g., Data Use Agreement*, DEP'T HEALTH & HUMAN SERV., https://www.hhs.gov/sites/default/files/ocio/eplc/EPLC%20Archive%20Documents/55-Data%20Use%20Agreement%20%28DUA%29/eplc_dua_practices_guide.pdf.

complexity of obtaining approvals for them.[231] Some agencies have a designated office or process to handle DUAs, but other agencies rely on extemporaneous processes and ad hoc, quid pro quo arrangements.[232] One such example is the Research Data Assistance Center, a centralized unit within the Centers for Medicare & Medicaid Services (CMS) dedicated to supporting data access requests.[233] In contrast, DUAs within the Department of Housing and Urban Development and the Department of Education are handled in decentralized business units, each with different routing channels and legal teams, which can confuse reviewers when multiple data requests between the same parties are routed simultaneously but separately.[234] Indeed, university DUA negotiators in one survey complained that the process was a game of "bureaucratic hot potato" and wondered, "Why isn't there just one template for everything?"[235] Ultimately, the lack of standardization means that DUAs often require extensive review and revision, creating substantial delays.

Agency-by-agency requirements also impede data sharing. These requirements can range from mandating that researchers only access data at an onsite facility, using government-authorized equipment, to capping the amount of computational cycles that can be used to analyze data, or restricting the amount of data available simultaneously.[236] These restrictions are particularly problematic, given that modern AI models can require massive amounts of data and computation to be most effective.

---

[231] O'Hara & Medalia, *supra* note 13, at 138, 141.
[232] NICK HART & KODY CARMODY, BARRIERS TO USING GOVERNMENT DATA: EXTENDED ANALYSIS OF THE U.S. COMMISSION ON EVIDENCE-BASED POLICYMAKING'S SURVEY OF FEDERAL AGENCIES AND OFFICES 18-20 (2018), BIPARTISAN POL'Y CTR., https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2018/10/Barriers-to-Using-Government-Data.pdf.
[233] *See Research Data Assistance Center (ResDAC)*, CTR. FOR MEDICARE & MEDICAID SERV. (Aug. 30, 2018), https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/ResearchGenInfo/ResearchDataAssistanceCenter.
[234] O'Hara & Medalia, *supra* note 13, at 141.
[235] Michelle M. Mello et al., *Waiting for Data: Barriers to Executing Data Use Agreements*, 367 SCIENCE 150 (Jan. 10, 2020), https://www.sciencemagazinedigital.org/sciencemagazine/10_january_2020/MobilePagedArticle.action?articleId=1552284#articleId1552284.
[236] Interview with Amy O'Hara, Executive Director, Georgetown Federal Statistical Research Data Center (Apr. 22, 2021); *see also Special Sworn Research Program,* BUREAU OF ECON. ANALYSIS, https://www.bea.gov/research/special-sworn-researcher-program (last visited Feb. 16, 2022); NAT'L CTR. FOR EDUC. STAT., RESTRICTED-USE DATA PROCEDURES MANUAL (2011), https://nces.ed.gov/pubs96/96860rev.pdf.

Broadly, the reasons for this dysfunction range from valid concerns about security and liability to the mundane and prosaic. Information technology systems within some agencies operate literally decades behind the technological frontier; a 2016 report from the Government Accountability Office (GAO) detailed examples of these legacy systems, discussing how several agencies were dependent on hardware and software that were no longer updateable, and required specialized staff to maintain.[237] A lack of incentives, a risk-averse culture, and an agency's statutory authority also play an important role in enabling or obstructing data sharing.[238]

We are by no means the first observers to note these problems. Advocates have been working for years to standardize and modernize government practices around data and technology.[239] For example, the Federal Data Strategy is the culmination of a multiyear effort to promulgate uniform data-sharing principles to address the fact that the United States "lacks a robust, integrated approach to using data to deliver on mission, serve the public, and steward resources."[240] However, substantial challenges remain, particularly since the bulk of the efforts focused on opening access to government data have not been undertaken with the specific needs of machine learning and AI in mind.

C. Tiered Data Access and Storage

The decentralized nature of government data has cascading implications across many aspects of the government data ecosystem. One key area that will affect the NRC is a lack of consistent storage and authentication access protocols across government agencies.

Because many government datasets contain sensitive data (e.g., high risk due to individual privacy concerns),[241] a crucial component of the NRC's data model will consist of a tiered storage taxonomy that distinguishes between datasets based on their sensitivity and correspondingly restricts access to different research groups. Interpreting tiered storage and access as two sides of the same coin, we

---

[237] *See* U.S. GOV'T ACCOUNTABILITY OFF., FEDERAL AGENCIES NEED TO ADDRESS AGING LEGACY SYSTEMS (2016), https://www.gao.gov/products/gao-16-696t.
[238] O'Hara & Medalia, *supra* note 13, at 140-41.
[239] *See, e.g., id.;* U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 237.
[240] PRESIDENT'S MGMT. AGENDA, *supra* note 212, at 11.
[241] GROVES & NEUFELD, *supra* note 229, at 12-13. For a precise definition of sensitive data, see *Glossary: Sensitive Information*, NAT'L INST. STANDARDS & TECH., https://csrc.nist.gov/glossary/term/sensitive_information.

reference existing models that are based on dataset risk levels and propose a framework for the NRC that aims to achieve the dual goals of streamlining the process of enabling research access to government data while maintaining privacy and security.

1. FedRAMP: A Tiered Framework for Data Storage on the Cloud

One type of tiered storage taxonomy already exists for third-party government cloud services in one of the federal government's major cybersecurity frameworks, the Federal Risk and Authorization Management Program (FedRAMP).[242] Enacted in 2011, the framework was designed to govern all federal agency cloud deployments, with certain exceptions detailed in Section 8 of this Section. FedRAMP offers two paths for cloud services providers to receive federal authorization. First, an individual agency may issue what is known as an authority-to-operate (ATO) to a cloud service provider after the provider's security authorization package has been reviewed by the agency's staff and the agency has identified any shortcomings that need to be addressed.[243] These types of ATOs are valid for each vendor across multiple agencies, as other agencies are permitted to reuse an initial agency's security package in granting ATOs. The second option available to cloud services providers is to obtain a provisional ATO from the FedRAMP Joint Authorization Board, which consists of representatives from the Department of Defense (DOD), the Department of Homeland Security (DHS), and the General Services Administration (GSA). These provisional ATOs offer assurances to agencies that DHS, DOD, and the GSA have reviewed security considerations, but before any specific agency is allowed to use a vendor's services, that agency must issue its own ATO.[244] In both the first and second cases, FedRAMP categorizes systems into low, moderate, or high impact levels (see Table 2).

---

[242] Shanna Nasiri, *FedRAMP Low, Moderate, High: Understanding Security Baseline Levels*, RECIPROCITY (Sept. 24, 2019), https://reciprocity.com/fedramp-low-moderate-high-understanding-security-baseline-levels/.
[243] Michael McLaughlin, *Reforming FedRAMP: A Guide to Improving the Federal Procurement and Risk Management of Cloud Services*, INFO. TECH. & INNOVATION FOUND. (June 15, 2020), https://itif.org/publications/2020/06/15/reforming-fedramp-guide-improving-federal-procurement-and-risk-management.
[244] *ATO Process*, CLOUD.GOV, https://cloud.gov/docs/compliance/ato-process/ (last visited Jun. 21, 2021).

Because FedRAMP requirements apply to all federal agencies when federal data is collected, maintained, processed, disseminated, or disposed of on the cloud, the NRC itself will need to be compliant with FedRAMP security standards irrespective of the organizational form it takes.[245] Every dataset brought on to the NRC would need to be reviewed under FedRAMP with appropriate access levels. If a cloud service has already been evaluated under FedRAMP because it was used in the past to house federal data, the service can inherit the same FedRAMP compliance level in the NRC without an additional evaluation.[246]

Besides classifying datasets, the other function of FedRAMP is to identify a comprehensive set of "controls," i.e., requirements and mechanisms that the cloud service providers must implement before the government dataset can be housed on them.[247] They are based on the National Institute of Standards and Technology (NIST) Special Publication 800-53, which provides standards and security requirements for information systems used by the federal government.[248]

These controls range widely and include requirements such as ensuring that the organization requesting certification "automatically disables inactive accounts," "establishes and administers privileged user accounts in accordance with a role-based access scheme that organizes system access and privileges into roles," "provides security awareness training on recognizing and reporting potential indicators of insider threat," or develops regular security plans in the event of a breach.[249]

---

[245] *Frequently Asked Questions*, FEDRAMP, https://www.fedramp.gov/faqs (last visited Feb. 16, 2022).
[246] *Do Once, Use Many - How Agencies Can Reuse a FedRAMP Authorization*, FEDRAMP (May 7, 2020), https://www.fedramp.gov/how-agencies-can-reuse-a-fedramp-authorization/.
[247] FEDRAMP, FEDRAMP LOW, MODERATE, AND HIGH SECURITY CONTROL BASELINES (2021), https://www.fedramp.gov/baselines/.
[248] *Security and Privacy Controls for Information Systems and Organizations*, NAT'L INST. STANDARDS & TECH. (Sept. 23, 2020), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf.
[249] *See, e.g., id.*; *NIST Risk Management Framework AC-2: Account Management*, NAT'L INST. STANDARDS & TECH., https://csrc.nist.gov/Projects/risk-management/sp800-53-controls/release-search#!/control?version=4.0&number=AC-2; *NIST Risk Management Framework AC-3: Access Enforcement*, NAT'L INST. STANDARDS & TECH., https://csrc.nist.gov/Projects/risk-management/sp800-53-controls/release-search#!/control?version=5.1&number=AC-3.

Requirements get more strenuous for FedRAMP "high impact" data (e.g., creating system-level air gaps to protect sensitive data).[250]

LEVEL	TYPE OF DATA	IMPACT OF DATA BREACH	NUMBER OF CONTROLS
Low-impact risk Low baseline Low-impact SaaS	Data intended for public use	Limited adverse effects; preserves the safety, finances, reputation, or mission of an agency	125
Moderate-impact risk (E.g., personally identifiable information)	Controlled unclassified data not available to the public	Can damage an agency's operations	325
High-impact risk (E.g., law enforcement, healthcare, emergency services)	Sensitive federal information	Catastrophic impacts such as shutting down an agency's operations, causing financial ruin, or threatening property or life	421

Table 2: FedRAMP levels are designated based on the degree of risk associated with the breach of an information system. The security baseline levels are based on confidentiality, availability, and integrity, as defined in Federal Information Processing Standard 199.[251]

There can be significant costs with obtaining these certifications and creating compliance plans, even if the underlying technical specifications can be addressed or already exist. A key issue for structuring the NRC is that the principal burdens of ensuring FedRAMP compliance should fall on NRC institutional staff, not originating agencies or individual academic researchers. As part of the FedRAMP certification process, NRC staff will have to consider how to give access to PIs in compliance with FedRAMP rules, but that process can and should avoid requiring originating agencies or individual universities to incur substantial expenses associated with hiring consultants and attorneys to certify FedRAMP compliance.[252]

---

[250] *See* Mark Bergen, *Google Engineers Refused to Build Security Tool to Win Military Contracts*, BLOOMBERG (June 21, 2018), https://www.bloomberg.com/news/articles/2018-06-21/google-engineers-refused-to-build-security-tool-to-win-military-contracts.

[251] *See* NAT'L INST. STANDARDS & TECH., STANDARDS FOR SECURITY CATEGORIZATION OF FEDERAL INFORMATION AND INFORMATION SYSTEMS (2004), https://nvlpubs.nist.gov/nistpubs/fips/nist.fips.199.pdf.

[252] *Partnering with FedRAMP*, FEDRAMP, https://www.fedramp.gov/cloud-service-providers/ (last visited Feb. 16, 2022). While it may cost cloud service providers between $365,000 and $865,000 and take six to twelve months to receive FedRAMP compliance, ADAM ISLES, SECURING YOUR CLOUD SOLUTIONS: RESEARCH AND ANALYSIS ON MEETING FEDRAMP/GOVERNMENT STANDARDS 21 (2017), such costs are borne by the cloud service providers themselves, not the providers' customers. Indeed,

While FedRAMP sets out common standards for cloud storage of government data within agencies,[253] it is an exception to an otherwise balkanized federal data-sharing standards landscape,[254] though it does not facilitate data exchange. The NRC needs to maintain compliance with not only FedRAMP requirements but also the requirements of any agency it is partnering with for data access.[255] Advocates interested in increasing government data availability have long fought to establish a universal FedRAMP equivalent across different agencies that provides shared standards for data sharing based on data sensitivity.[256] As we discuss in Section 8, establishing such universal, "centralized" security standards not only ensures internal uniformity but also removes barriers to data sharing.

The NRC's implementation of FedRAMP standards can also provide partnering agencies an important opportunity to reexamine their own standards and share best practices with one another.[257] This could involve raising or lowering requirements that are out of date,[258] given the current threat to the environment and research needs. The NRC can take

---

FedRAMP uses a "do once, use many" model: Once a cloud service provider obtains an authorization to operate (ATO), that ATO can be leveraged and reciprocated across multiple customers, eliminating duplicative efforts and inconsistencies that would come from requiring multiple re-authorizations. *See Do Once, Use Many, supra* note 246 at 11.

[253] Even within FedRAMP there are substantial amounts of variation in how different organizations ensure compliance with the relevant controls and standards, with many of the controls written broadly enough to give room for substantial interpretation. However, it does lay out a variety of considerations and requirements that are consistent across domains and allows a degree of predictability and reliance that is not present in other aspects of federal data governance.

[254] O'Hara & Medalia, *supra* note 13, at 141 ("Data sharing is taking place on a mandatory or voluntary basis, and data requests are managed through a designated staff/process or diffusely through an organization.").

[255] BIPARTISAN POL'Y CTR., *supra* note 232, at 17 ("The lack of standard procedures or guidelines for sharing data across federal agencies that fund research makes efforts to link and share data difficult or inefficient.").

[256] *See, e.g.,* Amy O'Hara*, US Federal Data Policy: An Update on The Federal Data Strategy and The Evidence Act,* 5 INT'L J. POPULATION DATA SCI. 5 (2020).

[257] While existing federal efforts and initiatives are already aimed at harmonizing data sharing best practices, *see, e.g.,* PRESIDENT'S MGMT. AGENDA, *supra* note 212, the NRC can accelerate these efforts. Indeed, the development of clear, consistent standards is crucial in facilitating data-sharing. DAVID CROTTY, IDA SIM & MICHAEL STEBBINS, OPEN ACCESS TO FEDERALLY FUNDED RESEARCH DATA 7 (2020).

[258] These requirements are inconsistent and out-of-date due to difficulties in defining risk as well as risk aversion on the parts of agencies. *See* O'Hara & Medalia, *supra* note 13, at 140-41; *see also* David S. Johnson et al., *The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility*, 657 ANNALS AM. ACAD. POL. & SOC. SCI. 247, 252-53 (2015).

inspiration from agencies' best practices, as well as from FedRAMP to develop a common NRC standard for determining data to be high, moderate, or low risk, as well as what consequences should flow from that assessment. In the later section on strategic considerations, we discuss how to enable this process by incentivizing agencies to participate in the NRC and selecting datasets that present a lower privacy and security risk.

In addition, given the diversity of data types and sources that could be stored on the platform, NRC policy should ensure that standards and protections exist for data storage in areas where FedRAMP has blind spots. FedRAMP is in part animated by risks from malicious actors like cybercriminals or adversarial foreign governments, but as we discuss in Section 6, privacy risks may arise even for the intended use case of analysis by NRC researchers. Of particular concern are instances where disparate datasets are combined, which may allow new inferences that make previously anonymous data individually identifiable, even when the data itself did not contain identifiable information.[259] Such combinations may also alter the original risk level of the data, creating an output that merits a higher risk classification. Furthermore, machine-learning models and representations may unintentionally reveal properties of the data used to train them,[260] and dissemination of these models could pose privacy risks.

This is not a challenge unique to the NRC; the U.S. Census Bureau and other government agencies engaged in data linkage have also had to develop means to address this issue.[261] One solution involves applying methods of additional noise to the data (differential privacy) in order to obfuscate individual data while preserving the data's utility for research. We discuss it and other privacy-enhancing technologies in greater detail in Section 7.[262] However, privacy-enhancing technologies are no panacea, and depending on the nature of the particular dataset, the goals of ensuring anonymization, while also enabling researchers to access fine-grained data can conflict.

---

[259] For a discussion of inference threats, see NAT'L ACAD. OF SCI., ENG'G & MED., FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION: NEXT STEPS 68 (Robert M. Groves & Brian A. Harris-Kojetin eds., 2017).
[260] Congzheng Song & Ananth Raghunathan, *Information Leakage in Embedding Models*, ARXIV (Mar. 31, 2020), https://arxiv.org/abs/2004.00053.
[261] *See, e.g.*, *Statistical Safeguards*, CENSUS BUREAU (July 1, 2021), https://www.census.gov/about/policies/privacy/statistical_safeguards.html.
[262] Alexandra Wood et al., *Differential Privacy: A Primer for a Non-Technical Audience*, 21 VAND. J. ENT. & TECH. L. 209 (2018).

The NRC can also draw from the "Five Safes" data security framework used by the UK Data Service,[263] the Federal Statistical Research Data Centers Network, and the Coleridge Initiative, a model centered on data, projects, people, access settings, and outputs.[264] The implementation of the 2019 Evidence Act is already using a similar Five Safes framework in making determinations around data linkage.[265] Through a combined framework, the NRC could place different anonymization requirements on datasets, depending on the circumstances of their access and the privacy agreements through which they were collected. Similarly, the NRC could control the dissemination scope of models, code, and data, depending on the sensitivity. Theoretical identifiability is less likely to be a concern when access and dissemination are restricted and the data is of a less sensitive nature or is not about individuals at all.[266]

2. Facilitating Researcher Access with a Tiered Access Model

How should researchers gain access to specific data resources? Currently, approval proceeds on an agency-by-agency basis.[267] Just as the value of the NRC for supporting AI research will depend in part on the extent to which it can bring together datasets from different agencies, it will also depend on the extent to which it can streamline the process for accessing data. One way to achieve this streamlining will be through a tiered access system for the NRC users, similar to FedRAMP's tiered system for storing federal data on the cloud, where higher tiers would enable access to higher-risk data, subject to the other requirements on compute and data use. We discuss this access system in more depth in Section 7.

Section 2 made the case that compute access should start with PIs at academic institutions. This authorization can also serve as the baseline, where all NRC-registered PIs can freely access and use low-risk

---

[263] *Regulating Access to Data*, UK DATA SERV., https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/five-safes (last visited Feb. 17, 2022).
[264] *Administrative Data Research Facility*, COLERIDGE INITIATIVE, https://coleridgeinitiative.org/adrf/ (last visited Feb. 17, 2022).
[265] *See* O'Hara, *supra* note 256.
[266] For additional discussion of the privacy implications of the NRC, see Section 5.
[267] *See* U.S. OFFICE OF MGMT. & BUDGET, BARRIERS TO USING ADMINISTRATIVE DATA FOR EVIDENCE-BUILDING 7 (2016).

datasets on the NRC. Additional tiers would impose more requirements, such as citizenship, security clearance, distribution restrictions, or compute and system restrictions. These access tiers will be similar to those used for determining FedRAMP classification for data storage, but while access and storage sensitivity may invoke similar considerations; they might not necessarily be the same.

CASE STUDY: Coleridge Initiative (Administrative Data Research Facility)

In partnership with the Census Bureau and funding from the Office of Management and Budget, the Coleridge Initiative, a nonprofit organization, launched the Administrative Data Research Facility (ADRF), a secure computing platform for governmental agencies to share and work with agency micro-data.[268] The ADRF is available on the Federal Risk and Authorization Management Program (FedRAMP) Marketplace and has a FedRAMP Moderate certification. Currently, the platform supports over one hundred datasets from fifty agencies.[269]

The ADRF provides access to agency-sponsored researchers and agency-affiliated researchers going through the ADRF training programs for free. Over the past three years, over five hundred employees from approximately one hundred agencies have gone through ADRF training programs.[270]

The ADRF provides a shared workspace for projects and the Data Explorer, a tool to view an overview and metadata (name, field description, and data type) of available datasets on the ADRF.[271] In order to access restricted data, users must meet review requirements set by the agency providing the data. In order to export data, users must go through a unique "Export Review" process.[272] The ADRF has a highly involved

[268] *Administrative Data Research Facility*, *supra* note 264.
[269] *Id.*
[270] *Applied Data Analytics Training*, COLERIDGE INITIATIVE, https://coleridgeinitiative.org/training-programs/.
[271] *ADRF User Guide: Data Explorer*, COLERIDGE INITIATIVE, https://web.archive.org/web/20210115211829/https://coleridgeinitiative.org/adrf/documentation/using-the-adrf/data-explorer/ (last visited Feb. 20, 2022).
[272] *ADRF User Guide: Exporting Results*, COLERIDGE INITIATIVE, https://web.archive.org/web/20210115212728/https://coleridgeinitiative.org/adrf/documentation/using-the-adrf/exporting-results/ (last visited Feb. 20, 2022).

default review process, requiring researchers to submit all code and output for the project for approval to the data steward and generating additional charges if requesting the export of more than ten files.[273] The agency providing the data can also amend the default review process if it wishes to do so.

Prior to transferring data files, the ADRF provides an application for data hashing to safely transmit data.[274] The ADRF also follows the "Five Safes" security model used by other government agencies, such as the UK Data Service.[275]

Data stewardship for the ADRF is defined in compliance with the Title III of the Evidence-Based Policymaking Act of 2018.[276] Once a restricted dataset is shared with the ADRF, one person within the agency will be assigned the data steward for all project requests. From there, procedures are developed with the agency, in terms of expectations for how the data will be protected, authorized users, and audit procedures for continued compliance.

Data stewards have access to an online portal in the ADRF. All project requests for specific data are routed to the data steward through this proposal. Once access has been granted, the data steward also has options to monitor the project for compliance.

Existing models for researcher access to sensitive datasets can help paint a picture of how the NRC might maintain and monitor a tiered access system. The NRC can emulate both the Coleridge Initiative and Stanford's Center for Population Health Sciences (PHS),[277] for instance, which serve as data intermediaries, in facilitating access to government

---

[273] *Id.*
[274] *ADRF User Guide: Data Hashing Application*, COLERIDGE INITIATIVE, https://web.archive.org/web/20210115201931/https://coleridgeinitiative.org/adrf/documentation/adrf-overview/data-hashing-application/(last visited Feb. 20, 2022).
[275] *ADRF User Guide: Security Model and Compliance*, COLERIDGE INITIATIVE, https://web.archive.org/web/20210115203555/https://coleridgeinitiative.org/adrf/documentation/adrf-overview/security-model-and-compliance/ (last visited Feb. 20, 2022).
[276] *Overview for Collaborators*, COLERIDGE INITIATIVE, https://coleridgeinitiative.org/collaborators/ (last visited Feb. 20, 2022).
[277] *Data*, STAN. MED. CTR. FOR POPULATION HEALTH SCI., https://med.stanford.edu/phs/data.html (last visited Feb. 17, 2022) [hereinafter *Stanford Medical Data*].

data. Indeed, these intermediaries have been documented as effective means to overcome barriers to data-sharing because they, at their core, negotiate and streamline relationships between data contributors and users.[278] For example, as a trusted intermediary, the NRC could centralize the DUA intake process by promulgating a universal standard form for agency DUAs.[279]

Furthermore, similar to the Coleridge Initiative example, a designated representative(s) within the agency could be assigned as the data steward for all project requests for a certain restricted dataset. Any project requiring access to data in higher tiers could commence only after its proposal was reviewed and approved by a relevant representative. Because NRC access begins with PIs, researchers would also have to obtain approval from their university Institutional Review Boards (IRBs), as needed. After project approval and NRC researcher clearance, data would be made available through the NRC's secure portal. Any violations of the terms of use or subject privacy could result in penalties ranging from a demotion of access tier to removal of NRC privileges or professional, civil, or criminal penalties, as relevant.

CASE STUDY: Stanford Center for Population Health Sciences

The Stanford Center for Population Health Sciences (PHS) provides a growing set of population health-related datasets and access methods to Stanford researchers and affiliates.[280] The PHS Data Ecosystem hosts high-value datasets, data linkages and filters, and analytical tools to aid researchers. The PHS partners with a wide range of public, nonprofit, and private entities to license population-level datasets for university researchers, ranging from low-risk, public datasets to restricted data containing Protected Health Information (PHI) and Personally-

[278] STANFORD CTR. ON PHILANTHROPY & CIV. SOC'Y, TRUSTED DATA INTERMEDIARIES 2-3 (2018).
[279] Others have also recognized the benefit of universal DUA templates. *See* Mello et al., *supra* note 235, at 150; *Guidance for Providing and Using Administrative Data for Statistical Purposes*, OFF. OF MGMT. & BUDGET (Feb. 14, 2014), https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf.
[280] *Stanford Medical Data*, *supra* note 277.

Identifiable Information (PII), such as Medicare, commercial claims such as Optum and Marketscan data,[281] and electronic medical records.

In addition to securing data storage and computational tools for researchers, PHS provides standardized and well-documented data access and management protocols, which increases data proprietor comfort with sharing data. PHS also has full-time staff who cultivate and maintain relationships with organizations holding data. This allows PHS to work with these groups to centralize data hosting and provide secure access to a wide array of researchers.

The PHS Data Portal is hosted on a third-party platform that enables data discovery, exploration, and clearly delineated, standardized steps for data access. The third-party platform, Redivis, utilizes a four-tier access system: (1) overview of data and basic documentation; (2) metadata access, including definitions, descriptions, and characteristics; (3) a 1 or 5 percent sample of the dataset; and (4) full data access.[282]

If data is classified as public, researchers can access it using specialized software, or simply download it directly.[283] For restricted data, the portal has forms integrated to easily apply for access.[284] After identifying the dataset, the researcher must apply for membership in the organization hosting the data.[285] An administrator of the organization owning the dataset can set member and study requirements that must be met, including training and institutional qualification, in order to access the data. Member applications can be set to auto-approval or require administrative approval. Once access has been granted to a data set, researchers can manipulate the data using specialized software. Usage restrictions are also specified individually on each dataset to control whether full, partial, or no output can be exported, and what review level

---

[281] *See Stanford PHS – Datasets*, REDIVIS, https://redivis.com/StanfordPHS/datasets?orgDatasets-tags=109.medicare (last visited Feb. 17, 2022).
[282] *Access Levels*, REDIVIS (JULY 2020), https://docs.redivis.com/reference/data-access/access-levels (last visited Feb. 17, 2022).
[283] *Step 1: Getting Access*, STAN. MED. PHS DOCUMENTATION, https://phsdocs.developerhub.io/start-here/getting-data-access (last visited Feb. 17, 2022).
[284] *Id.*
[285] *Id.*

is required for exporting. All applications for data and export are handled directly on the Data Explorer platform.

Currently, the PHS Data Portal is primarily for Stanford faculty, staff, students, or other affiliates.[286] Even with affiliate status, certain commercial datasets may require further data rider agreements for access. Non-Stanford collaborators must complete all of the same access requirements as Stanford affiliates, plus any requirements imposed by their own institution. Additionally, a "data rider" agreement on the original DUA is frequently necessary.[287]

To work with restricted data, the PHS provides two computing services for high-risk data: (1) Nero, with both an on-premises and Google Cloud Platform (GCP) platform versions; and (2) PHS-Windows Server cluster.[288] Both are managed by the Stanford Research Computing Center (SRCC). Both services are HIPAA compliant.[289] Unrestricted data can be used on any of Stanford's other computational environments (Sherlock, Oak) or simply downloaded to the researcher's local machine.

D. Promoting Interagency Harmonization and Adoption of Modern Data Access Standards

The federal data-sharing landscape suffers from divergent standards and practices, and individual agencies, left alone, have traditionally faced high hurdles to harmonizing and modernizing their data access standards.[290] As we have discussed, this state of affairs presents formidable barriers to AI R&D from a researcher's perspective but is also problematic both from an agency and societal perspective. As a report by the Administrative Conference of the United States finds from surveying the use of AI in the federal government, nearly half of agencies have experimented with AI to improve decision-making and operational

---

[286] *PHS Data-Use Workflow*, STAN. MED. PHS DOCUMENTATION, https://phsdocs.stanford.edu/start-here/phs-data-use-workflow (last visited Feb. 17, 2022).
[287] *Id.*
[288] *PHS Computing Environment*, STAN. MED. PHS DOCUMENTATION, https://phsdocs.stanford.edu/computing-environment (last visited Feb. 17, 2022).
[289] *Id.*
[290] *See* U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 237 (noting that from 2010-2015, many federal agencies increased their spending on operations and maintenance due to legacy systems).

capabilities, but they often lack the technical infrastructure and data capacity to use modern AI techniques and tools.[291] The lack of a modern, uniform standard for data sharing in AI research, therefore, makes it harder for agencies to realize gains in accuracy, efficiency, and accountability, which subsequently impacts citizens downstream, who are affected by agency decisions.[292]

The NRC can help overcome agency reluctance to share data by enabling access to agencies to compute their own data. This would solve at least two crucial problems for government agencies. First, access to the NRC's collective computing resources would overcome some difficulties that agencies have traditionally faced in setting up their own compute resources.[293] Second, facilitating agency access to modern data and compute resources would attract and build further in-house government AI expertise.[294] From a societal perspective, this could increase the government's capabilities in the responsible adoption of AI, help reduce the cost of core governance functions, and increase agency efficiency, effectiveness, and accountability.[295]

The NRC can also learn from and align with other initiatives to harmonize and modernize standards. The Evidence Act—which requires agencies to appoint chief data and evaluation officers—is one example. The legislation authorizing the creation of the NRC could provide a federal mandate to encourage the adoption of sharing best practices.[296] However, as we discuss in Section 5, a federal mandate alone, without any additional aid or incentives, may not be enough to incentivize the harmonization of data access and sharing standards.[297] The Task Force should therefore consider bundling the mandate with additional benefits, such as providing funding to assist agencies in expanding their technical or staff capabilities in furtherance of the NRC and the national AI strategy. The NRC is aligned with the existing bipartisan case for the

---

[291] DAVID FREEMAN ENGSTROM, ET AL., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 6, 71-72 (2020).
[292] *Id.* at 6-7.
[293] *Id.* at 71-72.
[294] *Id.* at 73.
[295] *Id.* at 6.
[296] *See* RESULTS FOR AMERICA, THE PROMISE OF THE FOUNDATIONS FOR EVIDENCE-BASED POLICYMAKING ACT AND PROPOSED NEXT STEPS (2019).
[297] For example, the Uniform Federal Crime Reporting Act of 1988 requires federal law enforcement agencies to share crime data with the FBI. *See* 34 U.S.C. §§41303(c)(2), (3), (4). Unfortunately, though, no federal agencies apparently currently share their data with the FBI under this law. NAT'L ACAD. OF SCI., ENG'G, & MED, *supra* note 259, at 41.

National Secure Data Service (NSDS) (described in the case study below), a service that would facilitate researcher access to data with enhanced privacy and transparency, recommended by the Commission on Evidence-Based Policymaking in 2018. Both the NRC and NSDS are complementary data-sharing initiatives that have the potential to considerably improve public service operational effectiveness. We elaborate further on the NSDS proposal in Section 5. Lastly, training programs are promising avenues to increase NRC adoption and agency support. For example, as described in the case study above, the Coleridge Initiative has hosted workshops to train over five hundred employees from approximately one hundred agencies on data use over the past three years.

CASE STUDY: The Evidence Act

In pursuit of greater, more secure access to and linkage of government administrative data, a bipartisan Commission on Evidence-Based Policymaking was set up by Congress in March 2016. The commission's final report[298] included twenty-two recommendations for the federal government to build infrastructure, privacy-protecting mechanisms,[299] and institutional capacity to provide secure access to public data for statistical and research purposes. One recommendation was to create a "National Secure Data Service" (NSDS) to facilitate access to data for the purpose of building evidence, while maintaining privacy and transparency. Through this service, the NSDS could help researchers by temporarily linking existing data and providing secure access, without itself creating a data clearinghouse.

The Foundations for Evidence-Based Policymaking Act of 2018[300] created some of the legislative footing for the commission's recommendations. In particular, it created new roles for chief data, evaluation, and statistical officials, and sought to increase access and

[298] KATHARINE G. ABRAHAM & RON HASKINS, COMM'N ON EVIDENCE-BASED POLICYMAKING, THE PROMISE OF EVIDENCE-BASED POLICYMAKING (2018).
[299] These privacy-preserving mechanisms are especially important in light of ongoing legal and political challenges in differential privacy application to Federal data. *See, e.g.,* DAN BOUK & DANAH BOYD, DEMOCRACY'S DATA INFRASTRUCTURE (2021).
[300] Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435, 132 Stat. 5529 (2018).

linkage of datasets previously within the scope of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[301]

Finally, the 2020 Federal Data Strategy and associated Action Plan[302] sought to put those legislative provisions into action. The strategy included plans to improve data governance, to make data more accessible, to improve government use of data, and to boost the use and quality of data inventories, metadata, and data sensitivity.

The central remaining step envisioned by the initial Evidence-Based Policymaking Commission is a National Secure Data Service (NSDS) modeled on the UK's Data Service.[303] The UK's Data Service provides access to a range of public surveys, longitudinal studies, UK census data, international aggregate data, business data, and qualitative data. Alongside access, it provides guidance and training for data use, develops best practices and standards for privacy, and has specialized staff who apply statistical control techniques to provide access to data that are too detailed, sensitive, or confidential to be made available under standard licenses.

E. Sequencing Investment into Data Assets

Given the significant hurdles in negotiating data access, the NRC will need to strategically sequence which agencies and datasets to focus on for researcher use. The federal government collects petabytes of data,[304] each with varying degrees of restrictions or openness. In considering which datasets to prioritize, the NRC can draw from the example of other data-sharing initiatives, as well as focus on data sets in the short term that do not pose complex challenges with regard to data privacy or sharing. One private-sector example is Google Earth Engine, which aggregated petabytes (approximately 1 million gigabytes) of satellite images and geospatial datasets and then linked that access to Google's cloud-computing services to allow scientists to answer a variety

---

[301] Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, 116 Stat. 2962 (2002).
[302] *Overview*, FED. DATA STRATEGY (2020), https://strategy.data.gov/overview/ (last visited Feb. 19, 2022).
[303] *UK Data Service*, UK DATA SERV., https://www.ukdataservice.ac.uk/ (last visited Jun. 21, 2021).
[304] For example, the Social Security Administration alone has over 14 petabytes of data, stored in roughly 200 databases. *See* ENGSTROM, *supra* note 291, at 72.

of crucial research questions.[305] This process of aggregating complex data and hosting it in a friendly computing infrastructure to facilitate research, demonstrates the compelling value of coupling compute and data. As another example, ADR UK identifies specific areas of research that are of pressing policy interest, such as "world of work,"[306] and prioritizes data access for researchers working on those topics. The UK Data Service offers datasets derived from survey, administrative, and transaction sources, including productivity data from the Annual Respondents Database,[307] innovation data from the UK Innovation Survey,[308] geospatial data from the Labour Force Survey,[309] Understanding Society,[310] and sensitive data about childhood development.[311]

When prioritizing datasets and agencies for NRC partnership, we recommend the following criteria:

- *Data that is valuable to AI researchers but is not currently available in a convenient form.* For example, in a July 2019 request for comments, the Office of Management and Budget (OMB) asked members of the public to provide input on characteristics of models that make them well-suited to AI R&D, what data is currently restricted, and how liberation of such data would accelerate high-quality AI R&D.[312] In one

---

[305] *Google Earth Engine*, GOOGLE EARTH ENGINE, https://earthengine.google.com (last visited Aug. 15, 2021).
[306] *World of Work*, ADR UK, https://www.adruk.org/our-work/world-of-work/ (last visited Feb. 19, 2022).
[307] *Annual Respondents Database, 1973-2008: Secure Access*, UK DATA SERV. (2020), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6644.
[308] *UK Innovation Survey*, UK DATA SERV. (2021), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6699.
[309] *Quarterly Labour Force Survey, 1992-2021: Secure Access*, UK DATA SERV. (2021), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6727.
[310] *Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009: Secure Access*, UK DATA SERV. (2021), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6676.
[311] These datasets have helped researchers tackle some specific, public good questions. *See, e.g.*, Francisco Perales, *Why Does the Work Women Do Pay Less Than the Work Men Do?*, UK DATA SERV. (Dec. 8, 2011), https://beta.ukdataservice.ac.uk/impact/case-studies/case-study?id=62; Eva-Maria Bonin, *Do Parenting Programmes Reduce Conduct Disorder and Its Cost to Society?*, UK DATA SERV. (Apr. 4, 2012), https://beta.ukdataservice.ac.uk/impact/case-studies/case-study?id=93.
[312] *Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing; Request for Information*, OFF. MGMT & BUDGET, 84 Fed. Reg. 32962 (July 10, 2019).

response, the Data Coalition argued that controlled release of private but structured indexed data in data.gov would be valuable for research.[313] The Data Coalition also urged agencies to consider releasing raw, unstructured datasets, such as agency call center logs, consumer inquiries, and complaints, as well as regulatory inspection and investigative reports.[314] Another example of data that is currently challenging to access, but is a matter of public record, is electronic court records housed in a system by the Administrative Office of the U.S. Courts.[315]

- *Data housed within agencies that have statutory authority to share data and/or that have previous data-sharing experience*. The Census Bureau, for instance, has greater existing statutory interagency linkage than other agencies and has preexisting substantial in-house data analysis expertise.[316] The U.S. Bureau of Labor Statistics has an existing process for sharing restricted datasets (in the categories of employment and unemployment, compensation and working conditions, and prices and living conditions) with researchers.[317]

- *Data with limited privacy implications*. For example, agencies whose data concerns natural phenomena, rather than individuals, may be easier to manage from a privacy perspective—e.g., NASA, the US Geological Service, and the National Oceanic and Atmospheric Administration. Datasets like those housed in NASA's Planetary Data System,[318] but that are not easily available to researchers, may serve as a valuable

---

[313] Nick Hart, *Data Coalition Comments on AI Data and Model R&D RFI*, DATA COALITION (Aug. 9, 2019), http://www.datacoalition.org/wp-content/uploads/2019/09/Comment.RFI_.OMB_.2019-14618.DataCoalition.pdf.
[314] *Id.*
[315] *See* Adam R. Pah et al., *How to Build a More Open Justice System*, 369 SCIENCE 134 (2020), https://www.science.org/doi/full/10.1126/science.aba6914; *see also* Seamus Hughes, *The Federal Courts Are Running an Online Scam*, POLITICO (Mar. 20, 2019), https://www.politico.com/magazine/story/2019/03/20/pacer-court-records-225821/.
[316] *Legal Authority and Policies for Data Linkage at Census*, CENSUS BUREAU (Apr. 4, 2018), https://www.census.gov/about/adrm/linkage/about/authority.html.
[317] *BLS Restricted Data Access*, U.S. BUREAU OF LAB. STAT., https://www.bls.gov/rda/restricted-data.htm (last updated May 20, 2021).
[318] *Welcome to the PDS*, NASA, https://pds.nasa.gov (last visited Feb. 19, 2022).

starting point for the NRC. Increasing the availability and interoperability of datasets from these agencies would advance the core mission of the NRC and could be done without jeopardizing individual privacy.

IV. ORGANIZATIONAL DESIGN

What institutional form should the NRC take? Two overarching considerations are: (1) ease of access to data; and (2) ease of coordination with compute resources.[319] As we discussed in Section 3 and will detail in depth in Section 5, the federal data-sharing landscape among agencies is highly fragmented, with many agencies reluctant to or legally constrained from sharing their data. The NRC will need to coordinate between the entities supplying compute infrastructure and researchers themselves. As the NRC's goal is to provide researchers with access to government data *and* high-performance computing power, one without the other will fall short of achieving the NRC's mission.

Drawing on extensive work in support of the Evidence Act, we recommend the use of Federally Funded Research and Development Centers (FFRDCs) and private-public partnerships (PPPs) as possible organizational forms for the NRC. We recommend the creation of an FFRDC at affiliated government agencies in the short term, as we believe this path allows for the easiest facilitation of both the compute infrastructure and access to government data. In the longer term, the establishment of a PPP could facilitate greater data sharing and access between the public and private sectors. Importantly, other options include creating an entirely new federal agency or bureau within an existing agency. While these options might simplify coordination with compute resources, both pose challenges, with respect to data accessibility and interagency data sharing.

A. Federally Funded Research and Development Center

FFRDCs are quasi-governmental nonprofit corporations sponsored by a federal agency but operated by contractors, including

---

[319] While we believe that these are the primary axes for consideration, some secondary considerations include organizational clout, talent retention, and bureaucratic overhead.

universities, other nonprofit organizations, and private-sector firms.[320] The FFRDC model confers the benefits of a close agency relationship, alongside independent administration, in facilitating access to data. Due to their intimate subcontracting relationships with their parent agency, all FFRDCs benefit from data access that goes "beyond that which is common to the normal contractual relationship, to Government and supplier data, including sensitive and proprietary data."[321]

A recent report by Hart and Potok on the National Secure Data Service (NSDS) (see case study in Section 3) also supports the FFRDC model as an optimal way to facilitate access to and linkage of government administrative data.[322] The report considered FFRDCs, alongside such other institutional forms, as creating an entirely new agency, housing the NSDS in an existing agency, and developing a university-led, data-sharing service, but the report ultimately recommended the FFRDC model for several reasons. An FFRDC can scale quickly, because it can access government data and high-quality talent more easily than other options.[323] An FFRDC can also leverage existing government expertise. The NSF, for instance, already sponsors five separate FFRDCs and has extensive experience cultivating and maintaining networks of researchers.[324]

However, the FFRDC model comes with a few limitations. First, an FFRDC's role is restricted to research and development for its sponsoring agency that "is closely associated with the performance of inherently governmental functions."[325] Thus, it would be important to

---

[320] CONGRESSIONAL RSCH. SERV., FEDERALLY FUNDED RESEARCH AND DEVELOPMENT CENTERS (FFRDCS): BACKGROUND AND ISSUES FOR CONGRESS 1 (2020).

[321] *Id. See also, About IDA*, INST. DEFENSE ANALYSES, https://www.ida.org/about-ida (last visited Feb. 19, 2022) (emphasizing that IDA, the private sector subcontractor that operates the Science & Technology Policy Institute and several other FFRDCs, "enjoys unusual access to classified government information and sensitive corporate proprietary information."); U.S. GOV'T ACCOUNTABILITY OFFICE, FEDERALLY FUNDED RESEARCH AND DEVELOPMENT CENTERS: IMPROVED OVERSIGHT AND EVALUATION NEEDED FOR DOD'S DATA ACCESS PILOT PROGRAM 6 (2020) (discussing how the Department of Defense was able to establish a three-year pilot program that allowed its FFRDC researchers to forgo having to obtain nondisclosure agreements with each data owner in order to streamline the data-access process).

[322] NICK HART & NANCY POTOK, MODERNIZING U.S. DATA INFRASTRUCTURE: DESIGN CONSIDERATIONS FOR IMPLEMENTING A NATIONAL SECURE DATA SERVICE TO IMPROVE STATISTICS AND EVIDENCE BUILDING (2020).

[323] *Id.* at 26.

[324] *Id.* at 26–27, 29–30.

[325] *Id.* at 6. Note that while the FFRDC must operate to serve its sponsors, in establishing an FFRDC, the sponsor must ensure that it operates with substantial independence; the FFRDC must be "operated, managed, or administered by an

ensure alignment during the contracting phase with the NRC's core functions.

Second, the success of an FFRDC model for the NRC will depend on the ability of the sponsoring agency to gain cooperation across the federal government to provide the data needed for research. One way to do this would be for multiple agencies to co-sponsor the FFRDC, reducing contracting friction for datasets.[326] Another option would be to create *multiple* FFRDCs housed in different agencies, incentivizing each of those agencies to share their data with the respective FFRDC. An analogous example could include the National Labs as a network, where each National Lab would be an *instantiation* of the NRC within its own relevant agency.[327]

Third, multiple FFRDCs would require separate processes for compute resources. In the short term, the NRC may alleviate this problem by contracting for commercial cloud credits, which is likely already the short-term solution for the NRC to provide compute access. As discussed earlier, private sector cloud providers already have extensive experience in providing compute resources to the government[328] and to academic institutions.[329] Familiarity with these private cloud providers may reduce the friction in allocating compute

---

autonomous organization or as an identifiably separate operating unit of a parent organization." *See* FAR § 35.017(a)(2) (2020).

[326] One example of this is the Science & Technology Policy Institute, which we discuss in a case study below.

[327] U.S. DEP'T OF ENERGY, THE STATE OF THE DOE NATIONAL LABORATORIES 11-13 (2020).

[328] *See, e.g., More Federal Agencies Head to the Cloud with Azure Government*, APPLIED INFO. SCI. (Feb. 23, 2018), https://www.ais.com/more-federal-agencies-head-to-the-cloud-with-azure-government/; *see also AWS GovCloud*, AMAZON, https://aws.amazon.com/govcloud-us/ (last visited Feb. 19, 2022). Microsoft was also previously awarded a $10 billion contract from the Pentagon. *See* Kate Conger, *Microsoft Wins Pentagon's $10 Billion JEDI Contract, Thwarting Amazon*, N.Y. TIMES (Sept. 4, 2020), https://www.nytimes.com/2019/10/25/technology/dod-jedi-contract.html. However, this contract was recently canceled "due to evolving requirements, increased cloud conservancy and industry advances." Ellie Kaufman & Zachary Cohen, *Pentagon Cancels $10 Billion Cloud Contract Given to Microsoft Over Amazon*, CNN (July 6, 2021), https://www.cnn.com/2021/07/06/tech/defense-department-cancels-jedi-contract-amazon-microsoft/index.html. The Pentagon will now instead seek new bids for an updated Joint Warfighting Cloud Capability (JWCC) contract from Amazon and Microsoft. *Id.*

[329] *See, e.g.,* Bram Bout, *Helping Universities Build What's Next with Google Cloud Platform*, GOOGLE (Oct. 25, 2016), https://blog.google/outreach-initiatives/education/helping-universities-build-whats-next-google-cloud-platform; *Cloud Computing for Education*, AMAZON, https://aws.amazon.com/education/ (last visited Apr. 10, 2022).

resources among researchers at multiple FFRDCs.

In the longer term, the FFRDC model may not be the most efficient. From a cost and sustainability perspective, FFRDCs have traditionally suffered from significant overruns, as they "operate under an inadequate, inconsistent patchwork of federal cost, accounting and auditing controls, whose deficiencies have contributed to the wasteful or inappropriate use of millions of federal dollars."[330] Another concern is that, historically, FFRDC infrastructure has not been routinely updated. A 2017 Department of Energy report highlighted that FFRDC infrastructure was inadequate to meet the mission.[331] NASA's Inspector General also highlighted that more than 50 percent of the Jet Propulsion Laboratory (a NASA FFRDC) equipment was at least fifty years old.[332] If an FFRDC version of the NRC experiences these same challenges, we recommend that the NRC, in the long run, switch to a public-private partnership model.

CASE STUDY: Science & Technology Policy Institute (STPI)

STPI is an FFRDC chartered by Congress in 1991 to provide rigorous objective advice and analysis to the Office of Science and Technology Policy and other executive branch agencies.[333] STPI is managed by the Institute for Defense Analyses (IDA), a nonprofit organization that also manages two other FFRDCs: the Systems and Analyses Center and the Center for Communications and Computing.[334] IDA has no other lines of business outside the FFRDC framework.[335]

STPI's primary federal sponsor is the National Science Foundation, but research at STPI is also co-sponsored by other federal agencies, including the National Institute of Health (NIH), Department of Energy (DOE), Department of Transportation (DOT), Department of Defense (DOD),

---

330 CONG. RSCH. SERV., *supra* note 320, at 11–12 (2020).
331 U.S. DEP'T OF ENERGY, ANNUAL REPORT ON THE STATE OF THE DOE NATIONAL LABORATORIES 87 (2017).
332 CONG. RSCH. SERV., *supra* note 320, at 19.
333 CONG. RSCH. SERV., OFFICE OF SCI. AND TECH. POL'Y (OSTP): HISTORY AND OVERVIEW 9 (2020). STPI's duties are also specified in 42 U.S.C. § 6686.
334 *What are FFRDCs?*, INST. DEFENSE ANALYSES, https://www.ida.org/ida-ffrdcs (last visited Feb. 19, 2022).
335 *Id.*

and Department of Health and Human Services (HHS).[336] Due to the "unique relationship" between an FFRDC and its sponsors, STPI "enjoys unusual access to highly classified and sensitive government and corporate proprietary information."[337]

NSF appropriations provide the majority of funding for STPI, including $4.7 million in FY 2020,[338] but a limited amount of funding is also provided from other federal agencies.[339] STPI has approximately forty full-time employees and has access to the expertise of IDA's approximately eight hundred other employees.[340] As an FFRDC, STPI may also contract for expertise, as required for a particular project.[341] The statute specifying STPI's duties also directs it to consult widely with representatives from private industry, academia, and nonprofit institutions, and to incorporate those views in STPI's work to the maximum extent practicable.[342]

STPI is also required to submit an annual report to the president on its activities, in accordance with requirements prescribed by the president,[343] which provides additional accountability for the FFRDC. According to STPI's 2020 report, STPI worked across multiple federal agencies, supporting them on forty-eight separate technology policy analyses throughout 2020.[344]

B. A Public-Private Partnership (PPP)

A Public-Private Partnership (PPP) would create a partnership between federal agencies and private-sector organizations to jointly house and manage data-sharing efforts and run compute infrastructure. Because different agencies and private-sector members may have

---

336 *Sponsors*, INST. DEFENSE ANALYSES, https://www.ida.org/en/about-ida/sponsors (last visited Feb. 19, 2022).
337 *Id.*
338 CONG. RSCH. SERV., *supra* note 333, at 9-10.
339 For instance, from 2008–2012, these other federal agencies contributed a total of $9.8 million of funding to STPI while NSF contributed about $24 million. U.S. GOV'T ACCOUNTABILITY OFF., FEDERALLY FUNDED RESEARCH CENTERS: AGENCY REVIEWS OF EMPLOYEE COMPENSATION AND CENTER PERFORMANCE 43-44 (2014).
340 CONG. RSCH. SERV., *supra* note 333, at 9-10.
341 *Id.*
342 42 U.S.C. § 6686(d).
343 42 U.S.C. § 6686(e).
344 SCI. & TECH. POL'Y INST., REPORT TO THE PRESIDENT FISCAL YEAR 2020 (2020).

different contracting preferences, intellectual property goals, and security allowances for data access, creating a data-sharing partnership within this patchwork framework could be challenging in the immediate future. Nonetheless, PPPs can provide a number of long-term benefits, as they have been used successfully as data clearinghouses to produce, analyze, and share data between the public and private sector.[345] Indeed, recognizing the benefits of the PPP model, the European Union has launched a new initiative called the Public Private Partnerships for Big Data that will offer a secure environment for cross-sector collaboration and experimentation using both commercial and public data.[346] In general, PPPs for data-sharing can increase the quality and quantity of R&D, increase the value and efficiency of sharing public sector data, and reduce the long-run cost necessary to manage and maintain the data-sharing infrastructure.[347]

CASE STUDY: ALBERTA DATA PARTNERSHIPS (ADP)

Founded in 1997, the ADP PPP is designed to provide long-term management of comprehensive digital data sets for the Alberta market.[348] The PPP is structured as a joint venture between ADP, a nonprofit, and Altalis Ltd. whereby the ADP is the "custodian" of government data and Altalis is the "operator."[349] More specifically, geospatial data is owned by the provincial government, but exclusive licensing arrangements are granted to ADP to allow for sales.[350] Meanwhile, Altalis, under the direction and oversight of ADP, builds software to securely load and distribute these provincial spatial datasets to users. Altalis also provides training to end-users and is responsible for

---

345 *See, e.g., Open Government*, MILLENNIUM CHALLENGE CORP., https://www.mcc.gov/initiatives/initiative/open (last visited Feb. 19, 2022); NAT'L GEOSPATIAL ADVISORY COMM., ADVANCING THE NATIONAL SPATIAL DATA INFRASTRUCTURE THROUGH PUBLIC-PRIVATE PARTNERSHIPS AND OTHER INNOVATIVE PARTNERSHIPS (2020); NAT'L AERONAUTICS & SPACE ADMIN., PUBLIC-PRIVATE PARTNERSHIPS FOR SPACE CAPABILITY DEVELOPMENT 33-36 (2014).
346 *Big Data Value Public-Private Partnership*, EUROPEAN COMM'N (Mar. 9, 2021), https://digital-strategy.ec.europa.eu/en/library/big-data-value-public-private-partnership.
347 RAND, PUBLIC-PRIVATE PARTNERSHIPS FOR DATA-SHARING: A DYNAMIC ENVIRONMENT 33, 99 (2000).
348 *See Homepage - Alberta Data Partnerships*, ALTA.DATA P'SHIPS, http://abdatapartnerships.ca (last visited Aug. 15, 2021).
349 ALTA. DATA P'SHIPS, A P3 SUCCESS STORY 1 (2017).
350 *Id.*

cleaning, updating, and standardizing datasets.[351]

In choosing its "operating partner" (i.e., Altalis) for the joint venture, the ADP board initially issued a "Request for Information" that solicited proposals from private-sector companies whose core business was the improvement, maintenance, management, and distribution of spatial data.[352] The ADP board ultimately chose Altalis, not only because it had the superior offering and existing capabilities, but also because Altalis was willing to take on all of the investment required, at its own risk, to build and operate the ADP system in accordance with ADP specifications.[353]

Today, all Altalis and ADP costs are covered by the operations of the joint venture.[354] The joint venture earns revenues, for instance, through directed project funding and data access fees from stakeholders, which include municipalities, regulatory agencies, energy, forestry, and mining organizations.[355] Any profits from the joint venture are split roughly 80/20 between Altalis and ADP, respectively, and ADP subsequently uses its profit shares to reinvest in data and system improvements.[356]

The ADP PPP claims to have generated efficiencies for data sharing. For instance, the ADP estimates that a traditional government-only approach to maintaining and distributing datasets would have ranged between $65 million and $120 million cumulatively since ADP's inception, and ADP claims to have provided its users with $6.8 million in cost savings.[357]

A PPP model could reduce the friction of coordination between data and compute. One example of using a PPP for compute resources is the COVID-19 High-Performance Computing Consortium, spearheaded by the Office of Science and Technology Policy, DOE, NSF, and IBM.[358] Drawing on the experience of XSEDE, the consortium has forty-three

---

[351] *Id.* at 19, 35.
[352] *Id.* at 15.
[353] *Id.*
[354] *Id.* at 1.
[355] NAT'L GEOSPATIAL ADVISORY COMM., PUBLIC-PRIVATE PARTNERSHIP USE CASE: ALBERTA DATA PARTNERSHIPS 1 (2020).
[356] ALTA. DATA P'SHIPS, *supra* note 349, at 15.
[357] *Id.* at 16.
[358] *The COVID-19 High Performance Computing Consortium*, COVID-19 HPC CONSORTIUM, https://covid19-hpc-consortium.org (last visited Feb. 20, 2022).

members from the public and private sectors that volunteer free compute resources to researchers with COVID-19-related research proposals.[359] The voluntary nature of compute provisioning, in this instance, provides benefits to both the researchers, who gain immediate access to compute, and the consortium members, who contribute to innovation and reap public relations benefits.

We also acknowledge that the evidence around the efficacy of PPPs is contested.[360] Indeed, there is no one-size-fits-all PPP model; PPPs differ vastly, according to the responsibilities allocated between the private sector and the public sector. The success of a PPP can depend on its structure.[361] According to a RAND Report of thirty case studies of successful public-private data clearinghouses, these clearinghouses have widely different organizations, access requirements, and strategies for managing data quality.[362] Such decision points are crucial. For example, some scholars emphasize the need for a trusted environment for the private and public sectors to handle privacy and ethics violations in sensitive industries.[363] Similarly, in the siloed federal data-sharing context, a PPP must consider how to divide functions in tackling these additional considerations in privacy, ethics, security, and intellectual property.

C. The NRC as a Government Agency

The NRC could also be constructed as a new government agency or bureau. The main advantages of this model would be the development of a distinct public-sector institution, devoted to AI compute and data. The NRC could be to cloud and data what the U.S. Digital Service is to government information technology. Such an agency would have to be

---

[359] *Id.*
[360] *See, e.g.,* DAVID HALL, WHY PUBLIC-PRIVATE PARTNERSHIPS DON'T WORK (2015); *Disadvantages and Pitfalls of the PPP Option*, APMG INT'L, https://ppp-certification.com/ppp-certification-guide/54-disadvantages-and-pitfalls-ppp-option (last visited Apr. 10, 2022).
[361] GRAEME A. HODGE, CARSTEN GREVE & ANTHONY E. BOARDMAN, INTERNATIONAL HANDBOOK ON PUBLIC–PRIVATE PARTNERSHIPS, 187–90 (2012).
[362] For example, on one end of a spectrum, the California Teale Data Center creates, owns, maintains, and archives its own datasets for private sector use. In contrast, the Pennsylvania Spatial Data Access houses metadata, requires users to ask the actual data sources for access. RAND, *supra* note 29, at 102–03. We encourage the Task Force to examine this comprehensive report to assess the various organizational options for a PPP data clearinghouse model.
[363] Angela Ballantyne & Cameron Stewart, *Big Data and Public-Private Partnerships in Healthcare and Research*, 11 ASIAN BIOETHICS R. 315, 315 (2019).

established by statute or executive mandate. Enabling legislation could create dedicated, professional staff to build and develop the NRC, vest the NRC with authority to mandate interagency data sharing, and create a long-term plan that is informed by the National AI Strategy.

There are, however, significant disadvantages to creating a new agency or bureau. First, the NRC could lay claim to no government datasets at all, and could subsequently encounter significant headwinds with having to negotiate with each originating agency for data, not to mention the constraints under the Privacy Act, discussed in Section 5. That said, enabling legislation could exempt the agency from the Privacy Act's data linkage prohibitions and transfer litigation risk for data leakages to the new agency. Second, a new agency may face greater challenges in recruiting top-flight talent.[364] According to the 2020 Survey on the Future of Government Service, a majority of respondents at federal agencies agreed that they often lose good candidates because of the time it takes to hire, and less than half agreed that their agencies have enough employees to do a quality job.[365] Moreover, many respondents highlighted inadequate career growth opportunities, inability to compete with private-sector salaries, and lack of a proactive recruiting strategy as major factors contributing to an inadequately skilled workforce in federal agencies.[366] FFRDCs, in contrast, can be negotiated with existing organizations, making the startup costs potentially lower. Third, while national laboratories have expertise contracting with entities to construct high-performance computing facilities, it is unclear how a new federal agency/office would approach such a task. It is one thing for an entity like the U.S. Digital Service to help develop IT platforms for U.S. agencies; it is another to simultaneously build a very large supercomputing facility and solve longstanding challenges with data access. Finally, it will be important to isolate the research mission of the NRC from political influence. To the extent that a new agency might provide less isolation from changes in presidential administrations and politically appointed administrators, this is an important consideration.

---

[364] *See* GOV'T ACCOUNTABILITY OFF., HUMAN CAPITAL: IMPROVING FEDERAL RECRUITING AND HIRING EFFORTS (2019); *see also Catch and Retain: Improving Recruiting and Retention at Government Agencies*, SALESFORCE, https://www.salesforce.com/solutions/industries/government/resources/government-recruitment-software/ (last visited Feb. 20, 2022).

[365] P'SHIP FOR PUB. SERV., SURVEY ON THE FUTURE OF GOVERNMENT SERVICE 2 (2020).

[366] *Id.*

While these disadvantages are considerable, ambitious legislative action could, in fact, make a new government agency a viable option.

V. DATA PRIVACY COMPLIANCE

The vision motivating the NRC is to support academic research in AI by opening access to both compute and data resources. Federal data can fuel basic AI research discoveries and reorient efforts from commercial domains toward public and social ones. As stated in the NRC's original call, "Researchers could work with agencies to develop and test new methods of preserving data confidentiality and privacy, while government data will provide the fuel for breakthroughs from healthcare to education to sustainability."[367]

But is an NRC seeded with public sector data, particularly administrative data from U.S. government agencies, even possible given the legal constraints? Research proposals that sweep broadly across agencies for personally identifiable or otherwise sensitive data[368] will rightly trigger concerns about potential privacy risk. The Privacy Act of 1974, the chief federal law governing data collected by government agencies, fundamentally challenges the notion of an NRC as a one-stop shop for federal data. Its research exceptions leave some uncertainty about open-ended research endeavors that go beyond statistical research or policy evaluation supporting an agency's core mission. Even if agencies deemed such research possible, researchers would be subject to access constraints, and the data itself may potentially require technical privacy treatments.

We make the following recommendations regarding data privacy and the NRC. First, agencies may be able to share anonymized administrative data with the NRC within the boundaries of the Privacy Act for the purposes of AI research, based on the Act's statistical research exemptions. Second, the NRC will require a staff of privacy professionals

---

[367] *National Research Cloud Call to Action*, *supra* note 210.

[368] Sensitive information, as defined by the National Institute of Standards and Technology, is information where the loss, misuse, or unauthorized access or modification could adversely affect the national interest or the conduct of federal programs, or the privacy to which individuals are entitled under 5 U.S.C. § 552a (the Privacy Act); that has not been specifically authorized under criteria established by an Executive Order or an Act of Congress to be kept classified in the interest of national defense or foreign policy. *See Glossary: Sensitive Information*, NAT'L INST. STANDARDS & TECH., https://csrc.nist.gov/glossary/term/sensitive_information (last visited Feb. 20, 2022).

that include roles tasked with legal compliance, oversight, and technical expertise. These professionals should build relationships with peers across agencies to facilitate data access. Third, the NRC should explore the design of virtual "data safe rooms" that enable researchers to access raw administrative microdata in a secure, monitored, and cloud-based environment. Fourth, we recommend the NRC Task Force engage the policy and statistical research communities and consider coordination with proposals for a National Secure Data Service, which has grappled extensively with these issues.

This section proceeds as follows. We first review the existing laws that apply to government agencies and the restrictions they impose on data access and sharing. We then describe current agency practices for sharing data with researchers and agencies under the Privacy Act. Last, we assess the implications of current legal constraints on NRC data sharing and the most important cognate proposal to promote data sharing under the Evidence Act.

We note at the outset that this section largely takes existing statutory constraints as a given. At a macro level, however, the challenges in data sharing also suggest that an ambitious legislative intervention could overcome many existing constraints, such as by statutorily (a) exempting the NRC from the Privacy Act's prohibition on data linkage; (b) granting the NRC the power to assume agency liabilities for data breaches; (c) mandating that agencies transfer any data that has been shared under a data use agreement or Freedom of Information Act (FOIA) request to the NRC, and; (d) requiring IT modernization plans to include provisions for data-sharing plans with the NRC.[369]

A. The Privacy Act

Data privacy issues are at the core of debates about sharing data, and the NRC will be no exception. Most data privacy debates in the U.S. today focus on the consumer data sector where data protection laws in the U.S. are limited to nonexistent. In contrast, many U.S. government agencies are subject to a robust privacy law, the Privacy Act of 1974, which was passed in response to concerns about government abuses of power.[370] For nearly fifty years, this legislation has been effective in its

---

[369] We thank Mark Krass for these insights.

[370] Agencies covered by the Act include "any Executive department, military department, Government corporation, Government controlled corporation, or other establishment in the executive branch of the [federal] Government (including the

primary goal of preventing the U.S. government from centralizing and broadly linking data about individuals across agencies. However, this approach has come at a cost, which is that most government agencies are prevented from freely sharing and linking data across agency boundaries, which in turn hampers agencies' operational and research efforts.[371] According to one government privacy expert, even when authorized or mandated to share data in limited circumstances, federal agencies are often reluctant to do so due to a myriad of factors, most prominently a lack of adoption of consistent data security standards, as well as difficulties with measuring and assessing privacy risks.[372] To that end, many agencies see promise in adopting technical privacy measures, such as differential privacy, or the creation of synthetic datasets as proxies for actual data, as a necessary precursor for enabling data sharing for both research purposes and interagency goals.[373]

In the nearly fifty years since the Privacy Act's passage, there have been periodic efforts to address the government's approach to data management while preserving data privacy. Examples include the E-Government Act of 2002,[374] the Confidential Information Protection and Statistical Efficiency Act of 2002,[375] and most recently, the Foundations for Evidence-Based Policymaking Act[376] and the National Data Strategy.[377] Most of these efforts have been aimed at sharing government data for statistical analysis and policy evaluation, and the scope of provisions may need to be broadened to support AI research. We view these efforts to be complementary: the NRC should build on these efforts while bringing increased attention to the compute resources that enable

---

Executive Office of the President), or any independent regulatory agency." 5 U.S.C. § 552(f)(1).

[371] U.S. GENERAL ACCOUNTING OFF., RECORD LINKAGE AND PRIVACY: ISSUES IN CREATING NEW FEDERAL RESEARCH AND STATISTICAL INFORMATION 10 (2001).

[372] Interview with Marc Groman, Former Senior Advisor for Privacy, White House Office of Management and Budget (Feb. 18, 2021); *see also* BIPARTISAN POL'Y CTR., BARRIERS TO USING GOVERNMENT DATA: EXTENDED ANALYSIS OF THE U.S. COMMISSION ON EVIDENCE-BASED POLICYMAKING'S SURVEY OF FEDERAL AGENCIES AND OFFICES 10 (2018).

[373] *See* Joseph Near & David Darais, *Differentially Private Synthetic Data*, NAT'L INST. STANDARDS & TECH. (May 3, 2021), https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data; *see also* Steven M. Bellovin et al., *Privacy and Synthetic Datasets*, 22 STAN. L. REV. 1 (2019).

[374] E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899.

[375] Confidential Information Protection and Statistical Efficiency Act of 2002, 44 U.S.C. § 3501 (2012).

[376] Foundations for Evidence-Based Policymaking Act of 2017, Pub. L. No. 115-435, 132 Stat. 5529 (2019).

[377] PRESIDENT'S MGMT. AGENDA, FEDERAL DATA STRATEGY 2020 ACTION PLAN (2020).

AI development as well as advanced data analysis.

B. Statutory Constraints on Data Sharing

One vision of the NRC is for it to act as a data warehouse for all government data. But that vision collides with fundamental constraints from laws designed to hamper broad and unconstrained data sharing between U.S. government agencies. Lacking an overarching, comprehensive privacy regime, similar to the European Union's General Data Protection Regulation (GDPR), the U.S. landscape is fragmented between a mix of sector-specific consumer laws and certain government-specific laws, such as the Privacy Act of 1974,[378] and limited-scope federal guidance, such as the Fair Information Practice Principles.[379] In particular, the Privacy Act, which focuses broadly on data collection and usage by federal agencies, and restricts sharing between them, poses challenges to the ambitions of the NRC's goal to make otherwise restricted government datasets more widely available.

Existing efforts, buttressed by such bills as the E-Government Act of 2002 and the Foundations of Evidence-Based Policymaking Act, have attempted to increase access by researchers to government data assets. Yet, these approaches were animated by the primary purposes of policy evaluation, not basic AI research, nor do they consider any ambitions on the part of agencies themselves to pursue AI research and development.[380]

Application of these laws and regulations to the NRC, in part, hinges on three factors: (1) the institutional form of the NRC, as we discuss in Section 4; (2) whether NRC users can invoke the Privacy Act's existing statistical research exception; and (3) whether researchers are

---

[378] Privacy Act of 1974, 5 U.S.C. § 552a (2012).

[379] There are many versions of the Fair Information Practice Principles, and the U.S. government has not institutionalized a specific version, though the version used by the Department of Homeland Security is commonly referenced (available at: https://www.dhs.gov/publication/privacy-policy-guidance-memorandum-2008-01-fair-information-practice-principles). The Organisation for Economic Cooperation and Development produced an influential version of them in 1980 (revised in 2013), which remains an authoritative source. *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, OECD (2013), https://www.oecd.org/digital/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm.

[380] *See* DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020) (documenting present use of AI by government agencies).

accessing data from multiple federal agencies. Here we briefly discuss the legal obligations of federal agencies. Even if the NRC does not take the form of a new standalone federal agency, agencies contributing data will remain subject to these constraints.

1. The Privacy Act's Limitations and Exemptions

The Privacy Act was enacted in response to growing anxiety about digitization, as well as the Watergate scandal during the Nixon presidency. The Act was motivated by concerns about the government's ability to broadly collect data on citizens and centralize it into digital databases, an emergent practice at the time. It is the primary limiting regulation for government data sharing, and has consequences for the NRC and, more directly, for any government agency wishing to share data with the NRC.

a. Data Linkage

The Privacy Act applies to systems of records, which are defined as "a group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual."[381] Importantly, the Act places strict limits on "record matching," or linking between agencies, for the purposes of sharing information about individuals.[382] Matching programs are only allowed when there is a written agreement in place between two agencies defining the purpose, legal authority, and the justification for the program; such agreements can last for eighteen months, with the option of renewal.[383] These limits were put in place in order to prevent the emergence of a centralized system of records that could track U.S. citizens or permanent residents across multiple government domains, as well as to limit the uses of data for the purposes it was collected for. Indeed, while linkage across datasets may be important for AI research,[384] it could potentially

---

[381] 5 U.S.C. § 552a(a)(5).
[382] 5 U.S.C. §§ 552a(a)(8)(A)(i)(I), (II).
[383] *The Privacy Act of 1974*, ELEC. PRIVACY INFO. CTR., https://epic.org/privacy/1974act/ (last visited Aug. 15, 2021).
[384] *See Fact Sheet: National Secure Data Service Act Advances Responsible Data Sharing in Government,* DATA COAL. (May 13, 2021), https://www.datacoalition.org/fact-sheet-national-secure-data-service-act-advances-responsible-data-sharing-in-government/; U.S GOV'T ACCOUNTABILITY OFF., RECORD

enable abuse, surveillance, or the infringement of such rights such as free speech by enabling persecution across the many areas in which a U.S. citizen or resident interacts with the federal system.[385]

Because the restriction on data linkages applies to linkages *between agencies*, the restriction applies in two particular scenarios for the NRC. First, if the NRC is instituted as a federal agency, then agency data-sharing with the NRC would run against the data linkages limitation of the Privacy Act. Second, federal agency staff access to the NRC could raise questions about interagency data linkage under the Privacy Act. However, the recommendation in Section 3 is focused on granting agencies streamlined access to the computing resources on the NRC and their own agency data, not to any multi-agency data hosted on the NRC. If the NRC is not designed as a federal agency and does not grant agency members access to interagency data, the Privacy Act's restrictions on data linkages may not apply.

We note that this approach to data management is both unusual and out of step with the private sector, as well as AI research specifically. The ability for both industry[386] and researchers[387] to associate multiple data sources and data points with a specific (anonymized) individual is common practice outside of government. In fact, this limitation is not one that many governments[388] or U.S. states[389] place on their data

---

LINKAGE AND PRIVACY: ISSUES IN CREATING NEW FEDERAL RESEARCH AND STATISTICAL INFORMATION (2001).

[385] It is no small irony that private companies in the U.S. have fulfilled that mission today. In fact, the U.S. government now approaches private industry, either through legal process or through procurement, when it requires data about individuals that the government itself does not collect. Senator Ron Wyden has proposed legislation to prevent the government from making these purchases. *Wyden, Paul and Bipartisan Members of Congress Introduce the Fourth Amendment Is Not For Sale Act*, RON WYDEN U.S. SENATOR FOR OR. (Apr. 21, 2021), https://www.wyden.senate.gov/news/press-releases/wyden-paul-and-bipartisan-members-of-congress-introduce-the-fourth-amendment-is-not-for-sale-act-.

[386] *See, e.g.*, WORLD ECON. FORUM, THE NEXT GENERATION OF DATA-SHARING IN FINANCIAL SERVICES (2019).

[387] *See, e.g.*, Stacie Dusetzina et al., *Linking Data for Health Services Research: A Framework and Instructional Guide,* AGENCY FOR HEALTHCARE RSCH. & QUALITY (Sept. 1, 2014), https://www.ncbi.nlm.nih.gov/books/NBK253315/.

[388] *See, e.g.,* EUROPEAN COMM'N, A EUROPEAN STRATEGY FOR DATA (2020) (arguing for cross-border data aggregation and linkage of both private and public sector data); M Sanni Ali et al., *Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment*, 10 FRONTIERS IN PHARMACOLOGY 984 (2019); *Data Linkage*, AUSTRALIAN INST. OF HEALTH & WELFARE (Jan. 4, 2020), https://www.aihw.gov.au/our-services/data-linkage.

[389] *See, e.g.*, ELSA AUGUSTINE, VIKASH REDDY & JESSE ROTHSTEIN, LINKING ADMINISTRATIVE DATA: STRATEGIES AND METHODS (2018) (describing tips for

systems. However, the Privacy Act's restrictions on data linkage remain uncontested, even in the various reform efforts we discuss below. It is worth noting that the federal government's broad bar against data linkages does incur welfare costs. For example, during the COVID-19 pandemic, the inability to share and link public health data created difficulties in tracking the spread and severity of the virus.[390] While projects like Johns Hopkins' Coronavirus Research Center[391] and the COVID Tracking Project[392] attempted to aggregate available data, the lack of data integration slowed important operational and research responses.[393] Other countries, for instance, integrated immigration and travel records to triage cases and prevent hospital outbreaks.[394]

We acknowledge the potential for data linkage to tackle important societal problems without recommending wholesale, unencumbered data linkage. Broad or unrestricted data linkage raises legitimate concerns about both individual privacy and widespread government surveillance,[395] made concrete by the disclosures of government whistle-blower Edward Snowden,[396] among others. An initiative to link Federal Aviation Administration (FAA) data with other agency data for COVID-19 response, for instance, would meet resistance from the Privacy Act. The Task Force should appreciate these tensions and trade-offs. Indeed,

---

conducting data linkages in California); *see also* U.S. DEP'T OF HEALTH & HUM. SERVS., STATUS OF STATE EFFORTS TO INTEGRATE HEALTH AND HUMAN SERVICES SYSTEMS AND DATA (2016).

[390] Ben Moscovitch, *How President Biden Can Improve Health Data Sharing For COVID-19 And Beyond*, HEALTH AFFS. (Mar. 1, 2021), https://www.healthaffairs.org/do/10.1377/hblog20210223.611803/full/.

[391] *Home*, JOHNS HOPKINS CORONAVIRUS RESOURCE CTR., https://coronavirus.jhu.edu/.

[392] THE COVID TRACKING PROJECT, https://covidtracking.com/ (last visited Feb 20, 2022).

[393] Fred Bazzoli, *COVID-19 Emergency Shows Limitations of Nationwide Data Sharing Infrastructure*, HEALTHCARE IT NEWS (June 2, 2020), https://www.healthcareitnews.com/news/covid-19-emergency-shows-limitations-nationwide-data-sharing-infrastructure.

[394] *See, e.g.,* C. Jason Wang et al., *Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing*, JAMA (Mar. 3, 2020), https://jamanetwork.com/journals/jama/fullarticle/2762689/; Fang-Ming Chen et al., *Big Data Integration and Analytics to Prevent a Potential Hospital Outbreak of COVID-19 in Taiwan*, 54 J. MICROBIOLOGY, IMMUNOLOGY & INFECTION 129-30 (2020).

[395] *See, e.g., Q&A on the Pentagon's "Total Information Awareness" Program*, AM. C.L. UNION, https://www.aclu.org/other/qa-pentagons-total-information-awareness-program (last visited Feb. 20, 2022); *The Five Problems with CAPPS II: Why the Airline Passenger Profiling Proposal Should Be Abandoned*, AM. C.L. UNION, https://www.aclu.org/other/five-problems-capps-ii (last visited Feb. 20, 2022).

[396] *See, e.g.,* BARTON GELLMAN, DARK MIRROR: EDWARD SNOWDEN AND THE AMERICAN SURVEILLANCE STATE (2020); EDWARD SNOWDEN, PERMANENT RECORD (2019).

agencies view technical measures for privacy preservation as a necessary component of any government data strategy, as methods such as multiparty computation or homomorphic encryption (which we discuss in Section 8) may allow for some forms of data linkages between agencies, without violating the Privacy Act.

b. No Disclosure Without Consent

Another core restriction of the Privacy Act is the "No Disclosure Without Consent" rule, which prohibits disclosure of records to any agency or *person* without prior consent from the individual to whom the record pertains.[397] Because the NRC would disclose federal agency data to researchers (i.e., to "person[s]"), this rule—unlike the restriction on record linkage—is legally relevant and unavoidable.

The Privacy Act, however, contains a number of exceptions to this rule. Most pertinent to the NRC's data-sharing efforts are exemptions for: (1) "routine use"; (2) specified agencies; and (3) statistical research. Under the first exemption, the Privacy Act permits agencies to disclose personally identifiable administrative data when such disclosure is among one of the "routine uses" of the data.[398] A dataset's "routine use" is defined on an agency-to-agency basis and is simply a specification filed with the Federal Register on the agency's plan to use and share its data.[399] As a result, the more broadly an agency defines "routine use" of its data, the more broadly that agency can share its data with other agencies without disclosure.[400] While courts have limited how broadly an agency can describe "routine uses,"[401] a large number of use cases can still be

---

[397] 5 U.S.C. § 552(b).
[398] 5 U.S.C. § 552(b)(3).
[399] The Privacy Act also contains specific carve-outs for disclosures to the Census Bureau and to the National Archives and Records Administration. However, the carve-outs for these two agencies require that the disclosures be made for the purposes of a census survey and of recording historical value, respectively. Because the NRC's explicit purpose is to democratize AI innovation, it is unlikely that the NRC can take advantage of this existing exception to dataset disclosures under the Privacy Act.
[400] For example, the Federal Emergency Management Agency's list of routine uses includes broad disclosure "[t]o an agency or organization for the purpose of performing audit or oversight operations as authorized by law, but only such information as is necessary and relevant to such audit or oversight function." Privacy Act of 1974; Department of Homeland Security Federal Emergency Management Agency-008 Disaster Recovery Assistance Files System of Records, 78 Fed. Reg. 25282 (May 30, 2013).
[401] *See, e.g., Britt v. Naval Investigative Serv.*, 886 F.2d 544 (3d Cir. 1989).

covered by a short, general statement.[402] Further research should be conducted on the conditions for when data sharing for research purposes constitutes routine use.

c. Implications for Data Sharing with Researchers

Much will rest on the interpretation of the "statistical research" exception, as applied to AI research. Despite the Privacy Act's constraints on data sharing, researchers have conventionally been able to access data directly from agencies, based on the statistical research exception to the Privacy Act. This exception allows disclosure of records "to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable."[403] Doing so requires either access to an approved research dataset, or for the researcher to negotiate an MOU directly with the agency, a role we suggest the NRC may be able to fill as an intermediary, acting as a negotiating partner to facilitate access requests between multiple researchers and agencies (discussed in Section 3).

While the Privacy Act does not define "statistical research," subsequent laws and policies have elaborated on the definition. For example, the E-Government Act defines "statistical purpose" to include the development of technical procedures for the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups.[404] Meanwhile, a "nonstatistical purpose" includes the use of personally identifiable information for any administrative, regulatory, law enforcement, adjudicative, or other purpose that affects the rights, privileges, or benefits of any individual.[405] That is, while researchers may use personally identifiable data for the broad purpose of analyzing group characteristics, they cannot use such data for targeted purposes to aid agencies with, for instance, specific adjudicative or enforcement functions.

The precise meaning of "statistical purpose," however, remains

---

[402] *The Privacy Act of 1974*, *supra* note 383.
[403] 5 U.S.C. § 552(b)(5).
[404] 5 U.S.C. § 552a(a)(8)(B)(i), (ii).
[405] 44 U.S.C. § 3561(8), (12).

"obscure and the evaluation criteria may be difficult to locate."[406] Yet, "statistical purpose" may well encompass data sharing for certain AI applications. The Act explicitly designates the Bureau of Labor Statistics, Bureau of Economic Analysis, and the Census Bureau as statistical agencies that have heightened data-sharing powers for statistical purposes.[407] These agencies regularly use AI in conducting their statistical activities.[408] While definitions of AI are themselves contested, statistical research may encapsulate at least some forms of machine learning and AI, if such research analyzes group characteristics[409] and does not identify individuals.

To be sure, the NRC should not enable researchers or agencies to conduct an end run around the Privacy Act. To that end, the NRC will require staff devoted to privacy compliance and oversight to ensure compliance. Key questions regarding individual identifiability, sensitivity of the data, or the potential for linkage and reidentification will need to be assessed by such staff.

d. Implications for Agency Data Sharing with the NRC

Notwithstanding the above avenues, agencies may nonetheless be reluctant to share data with the NRC and its researchers. Instances abound where federal agencies face constraints to sharing data, even if it is entirely legal or even federally mandated. For example, the Uniform Federal Crime Reporting Act of 1988 requires federal law enforcement agencies to report crime data to the FBI.[410] Yet, no federal agencies appear to have shared their data with the FBI under this law.[411] Similarly,

---

406 U.S. DEP'T OF HEALTH & HUM. SERVS., THE STATE OF DATA SHARING AT THE U.S. DEP'T OF HEALTH AND HUMAN SERVICES 16 (2018).
407 44 U.S.C. § 3575(4).
408 *See* ENGSTROM, HO, SHARKEY & CUÉLLAR, *supra* note 380, at 16 (finding that the Bureau of Labor Statistics is one of the top ten agencies that uses artificial intelligence); *Machine Learning*, CENSUS BUREAU (Apr. 17, 2019), https://www.census.gov/topics/research/data-science/about-machine-learning.html (asserting that the Census Bureau "needs" machine learning capabilities); BUREAU OF ECON. ANALYSIS, 2020 STRATEGIC ACTION PLAN 7 (2020) (highlighting the importance of artificial intelligence and machine learning to BEA's strategy).
409 Group level data analyses also have inherent privacy risks and harms. *See, e.g.*, Linnet Taylor, *Safety in Numbers? Group Privacy and Big Data Analytics in the Developing World*, *in* GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES 13 (2017).
410 *See* 34 U.S.C. § 41303(c)(2), (3), (4).
411 NAT'L ACAD. OF SCI., INNOVATIONS IN FEDERAL STATISTICS 41 (2017).

the Census Bureau is enabled by legislation that authorizes it to obtain administrative data from any federal agency and requires it to try to obtain data from other agencies whenever possible.[412] However, the statute does not similarly require the program agencies to provide their data to the Census Bureau. That is, although the Census Bureau is required to ask other agencies for data, those agencies are not required to, and often do not, provide it.[413]

Failure to engage in data sharing, even in the face of a statutory authorization, can stem from risk aversion. According to a GAO report, agencies choose not to share data because they tend to be "overly cautious" in their interpretation of federal privacy requirements.[414] Because legal provisions authorizing or mandating data sharing are often ambiguous,[415] agencies may err on the side of caution and choose not to share their data for fear of the downside risk that recipient use of the data may violate privacy or security standards.[416] To make matters worse, because agencies need to devote significant resources to facilitate data sharing, they may simply choose not to prioritize data sharing at all. The lack of resources poses a significant problem—according to a Bipartisan Policy Center study on agency data sharing, about half of agencies cited inadequate funding or the inability to hire appropriate staff as their "most critical" barrier to data sharing.[417]

The NRC may overcome these hurdles by clarifying legal provisions, ensuring that the benefits to agencies of data sharing outweigh the risks and costs, and advocating for resources. For instance, O'Hara and Medalia describe how the Census Bureau was able to obtain food stamp and welfare data from state agencies. In the face of ambiguous statutes authorizing the U.S. Department of Agriculture (USDA) and the U.S. Department of Health and Human Services (HHS) to perform data linkages across federally sponsored programs, states originally arrived at different statutory interpretations. Some states agreed to share their data only after (1) the Office of General Counsel at

---

[412] *See* 13 U.S.C. § 6.
[413] NAT'L ACAD. OF SCI., *supra* note 45, at 40.
[414] According to the study, "an agency's legal counsel may advise against sharing data as a precautionary measure rather than because of an explicit prohibition." U.S. GOV'T ACCOUNTABILITY OFFICE, SUSTAINED AND COORDINATED EFFORTS COULD FACILITATE DATA SHARING WHILE PROTECTING PRIVACY 1 (2013).
[415] *See* O'Hara & Medalia, *supra* note 13, at 141.
[416] ROBERT M. GROVES & ADAM NEUFELD, ACCELERATING THE SHARING OF DATA ACROSS SECTORS TO ADVANCE THE COMMON GOOD 12 (2017).
[417] BIPARTISAN POL'Y CTR., *supra* note 372, at 18–20.

both the USDA and HHS issued a memo clarifying that data sharing with the Census Bureau for statistical purposes was legal and encouraged; and (2) the states were convinced that data sharing would enable evidence building that could help them administer their programs.[418]

Broader data sharing with the NRC that combines multiple agency or external data sources may be facilitated by the passage of additional laws requiring agencies to share their data, subject to specific limitations on how that data is used by the NRC. Even then, the effect of that requirement is hardly a foregone conclusion. More is needed by way of both clarifying the extent to which data sharing is permitted and providing benefits that incentivize agencies to share their data.

Finally, to ensure compliance with the Privacy Act, as well as to facilitate the NRC's role as a data intermediary, the NRC will require a staff of privacy professionals that include positions tasked with legal compliance, oversight, and technical methods expertise. These professionals should build relationships with peers across agencies to facilitate data access.

CASE STUDY: Administrative Data Research UK

Administrative Data Research UK (ADR UK) is a new body, set up in July 2018, to facilitate secure, wide access to linked administrative datasets from across government for the purpose of public research.[419]

ADR UK was set up as a central, coordinating point between four national partnerships—ADR England, ADR Northern Ireland, ADR Scotland, and ADR Wales—as well as the UK-wide national statistics agency, Office for National Statistics (ONS). ADR UK labels itself as a "UK-wide strategic hub:" a central point that promotes the use of administrative data for research, engages with government departments to facilitate secure access to data, and funds public good research that uses administrative data.[420]

Funding for ADR UK came from a research council (Economic and Social Research Council, ESRC) and was initially committed from July 2018 to

---

[418] *See* O'Hara & Medalia, *supra* note 13, at 141.
[419] *ADR UK - Administrative Data Research UK. Data-Driven Change*, ADR UK, https://www.adruk.org (last visited Apr. 10, 2022).
[420] *About ADR UK*, ADR UK, https://www.adruk.org/about-us/about-adr-uk/ (last visited Feb. 20, 2022).

March 2022. A total of £59 million was provided.[421]

ADR UK serves three core functions. First, the promotion of the value and availability of government administrative datasets for research. ADR UK acts as a general advocate for the use of administrative datasets from across the British government. It also acts as a specific driver of research for the public good. It has identified specific areas of research that are of pressing policy interest (e.g., "world of work"[422]), and is focusing on creating access to linked datasets for researchers who tackle those priority themes.

The second core function is serving as a coordination point to encourage government data sharing, standards, and linkage of administrative datasets. Especially for its research calls, ADR UK is able to highlight multiple datasets, often spanning different government departments' scope areas that can be linked and used in research. In doing so, ADR UK plays an important role in facilitating research.

Third, ADR UK has a strategic funding approach to further the use of administrative datasets in research that has three categories of funding:

Building new research datasets: ADR UK's Strategic HubFund initially solicited invitation-only bids for researchers who would build new research datasets of public significance in the course of their work.[423] These new, research-ready datasets are now accessible to a wide range of researchers.[424]

Research Fellowship Schemes: A major funding focus now is on funding research through competitive open-bid invitations under a Research Fellowship Scheme.[425] Specific researchers are identified through the competition. They are accredited for secure data access and

[421] *Id.*
[422] *See World of Work,* ADR UK, https://www.adruk.org/our-work/world-of-work/ (last visited Feb. 20, 2022).
[423] *Funding Opportunities*, ADR UK, https://www.adruk.org/news-publications/funding-opportunities/ (last visited Feb. 20, 2022).
[424] *Id.*
[425] *Id.*

placed right at the heart of government (with 10 Downing Street), with access to linked datasets to answer questions of public significance.[426]

Methods Development Grants: Separately, ADR UK invites research proposals that further methodological progress for the use of large-scale administrative datasets, such that the wider social science community can draw on developed methods in research.[427]

C. Privacy and Security

The UK's 2017 Digital Economy Act[428] created a legal gateway for research access to secure government data. Deidentified data held by a public authority in connection with the authority's functions could be disclosed for research, under the assurance that individual identities would not be specified.

Any data shared with researchers is anonymized: personal identifiers are removed, and checks are made to protect against re-identification.[429] A rigorous accreditation process—for both the researcher and proposed research— is undertaken to ensure public benefit. Data access primarily takes place via a secure physical facility, or a secure connection to that facility, provided by ADR UK's constituent partners.[430] There is close monitoring of researcher activity and outputs, and any output is checked before release.[431]

From a researcher's point of view, access to ADR UK datasets requires the following steps:[432]

- Researcher submits proposal for project to ADR UK.
- Project is approved by relevant panels.
- Researcher engages in training and may take assessment (e.g., access to linked data held by ONS required accreditation to

---

[426] *Funding Opportunity: A Unique Chance to Shape Data Science at the Heart of UK Government*, ADR UK (Apr. 8, 2021), https://www.adruk.org/news-publications/news-blogs/funding-opportunity-a-unique-chance-to-shape-data-science-at-the-heart-of-uk-government-384/.
[427] *Funding Opportunities*, *supra* note 423.
[428] Digital Economy Act 2017, c. 30 (UK).
[429] ADR UK, TRUST, SECURITY AND PUBLIC INTEREST: STRIKING THE BALANCE 28 (2020).
[430] *Id.* at 29.
[431] *Id.*
[432] *How Do We Work with Researchers?*, ADR UK, https://www.adruk.org/our-mission/working-with-researchers/ (last visited Feb. 28, 2022).

ONS' Secure Research Service,[433] and can access data either in person or, where additionally accredited, through remote connection).

- Required data is determined by ADR UK (through one of the four regional partners, or ONS), then ingested by the relevant data center.
- De-identified data is made available through a secure data service (either at the ONS, or one of the four regional partners).
- Researcher conducts analysis; activity and outputs are monitored.
- Outputs are checked for subject privacy. Research serving the public good is published.

D. Complementary Efforts to Improve the Federal Approach to Data Management

The barriers to data sharing created by the Privacy Act have long posed a challenge to researchers interested in using government data to evaluate or inform policy.[434] The policy and statistical research communities, both within and outside the federal government, have engaged in admirable reform efforts to facilitate data sharing for policy evaluation.[435]

The Foundations for Evidence-Based Policymaking Act (EBPA) of 2018, which enacted reforms to improve data access for evidence-based decision-making, is a key achievement of these efforts to date. However, several of the provisions in the Act that helped to address some of the barriers to data linking and sharing were not passed by Congress. These provisions—known collectively as the National Secure Data Service (NSDS)—remain a high priority for facilitating further progress for sharing data for research purposes. According to the nonprofit Data Foundation, one of the major supporters of the NSDS, its passage will

---

[433] *Accessing Secure Research Data as an Accredited Researcher*, OFF. FOR NAT'L STAT., https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme (last visited Feb. 28, 2022).

[434] *See* NICK HART & NANCY POTOK, MODERNIZING U.S. DATA INFRASTRUCTURE: DESIGN CONSIDERATIONS FOR IMPLEMENTING A NATIONAL SECURE DATA SERVICE TO IMPROVE STATISTICS AND EVIDENCE BUILDING 17, 6-7 (2020).

[435] *Id.*

"create the bridge across the government's decentralized data capabilities with a new entity that jointly maximizes data access responsibilities with confidentiality protections."[436]

The NSDS is envisioned as an independent legal entity within the federal government that would have the legal authority to acquire and use data. However, this authority is currently conceived of as emanating from the EBPA, which focuses on using statistical data for evidence-building purposes. A broader source of authority may be necessary for AI research purposes under the NRC, which may be distinct from agency obligations. One clear area of overlap is the proposal's call for the NSDS to facilitate its own computing resources, which could be harmonized with the compute needs of the NRC. Like Section 4's discussion of organizational options, NSDS supporters identify a fundamental need for both a reliable funding source as well as thoughtful placement of the NSDS either within an existing agency or as an independent agency or FFRDC. The areas of common ground between the NRC and NSDS, as well as the expertise and momentum behind the proposal, strongly suggest that the NRC engage and coordinate with these efforts.

Another complementary initiative is the Federal Data Strategy (FDS), launched in 2018 by the executive branch and led by the OMB. FDS is a government-wide effort to reform how the entire federal government manages its data. The plan calls out the need for "safe data linkage" through technical privacy techniques,[437] and incorporates a directive from the 2019 Executive Order on Maintaining American Leadership in Artificial Intelligence to "[e]nhance access to high- quality and fully traceable federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections, consistent with applicable laws and policies."[438] The FDS directs OMB to "identify barriers to access and quality limitations" and to "[p]rovide technical schema formats on inventories," with a focus on open data sources (i.e., non-sensitive or individually identifying data).[439] Datasets identified by this process could be key candidates for populating the NRC.

While both the NSDS and the FDS may promote data sharing,

---

[436] *Id.* at 15.
[437] PRESIDENT'S MGMT. AGENDA, *supra* note 212, at 9.
[438] *Id.* at 31.
[439] *See What is Open Data?*, OPEN DATA HANDBOOK, https://opendatahandbook.org/guide/en/what-is-open-data/ (last visited Feb. 28, 2022).

these efforts are presently focused primarily on furthering policy evaluation purposes. Fortunately, there is much overlap and complementarity between these initiatives and the NRC, illustrating the broad importance of more effective mechanisms to share federal data securely and in a privacy-protecting way.

VI. TECHNICAL PRIVACY AND VIRTUAL DATA SAFE ROOMS

We now discuss the role of technical privacy methods for the NRC. In the past several decades, researchers have devised a variety of computational methods that enable data analysis while preserving privacy. These methods hold considerable promise for enabling the sharing of government data for research purposes. We note at the outset that technical methods are merely one mechanism to strengthen privacy protections. While effective, such methods may be neither sufficient nor universally appropriate. The application of any particular method does not obviate the need to inquire into whether the data itself adheres to articulated privacy standards. The methods discussed here are not "replacements" for the recommendations discussed earlier and never themselves justify the collection of otherwise problematic data.

Use of data from the NRC introduces two threats to individual privacy. The first type involves accidental disclosure by agencies (agency disclosure): An agency uploads a dataset to the NRC which lacks sufficient privacy protection and contains identifying information about an individual. A researcher—either analyzing this dataset alone or in conjunction with other NRC datasets—discovers this information and re-identifies the individual.[440] The second type involves accidental disclosure by researchers (researcher disclosure). Here, a researcher releases products computed on restricted NRC data (e.g., trained machine learning models, publications). However, the released products lack sufficient privacy protection, and an outside consumer of the research product learns sensitive information about an individual or individuals in the original dataset used by the researcher.[441]

---

[440] *See Keeping Secrets: Anonymous Data Isn't Always Anonymous*, BERKELEY SCH. OF INFO. (Mar. 15, 2014), https://ischoolonline.berkeley.edu/blog/anonymous-data/; Arvind Narayanan & Vitaly Shmatikov, *How to Break Anonymity of the Netflix Prize Dataset*, ARXIV (Nov. 22, 2007), https://arxiv.org/pdf/cs/0610105.pdf.

[441] Matt Fredrikson, Somesh Jha & Thomas Ristenpart, *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures,* 22 PROCS. ACM SPECIAL INT. GRP. ON SEC., AUDIT & CONTROL 1322 (2015); Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, ARXIV (June 15, 2021),

We recommend that, due to the infancy and uncertainty surrounding uses of privacy-enhancing technologies, privacy should primarily be approached via access policies to data. While there will be circumstances that suggest, or even mandate, technical treatments, access policies, discussed in Section 3, are the primary line of defense: they ensure sensitive datasets are protected by controlling who can access the data. We recommend a tiered access policy, with more sensitive datasets placed in more restricted tiers. For instance, highly restricted access data may correspond to individual health data from the VA, while minimally restricted access data may correspond to ocean measurements from NOAA. Proposals requesting access to highly restricted data would face heightened standards of review, and researchers may be limited to accessing only one restricted access dataset at a time. This approach mirrors current regimes where researchers undergo special training to work with certain types of data.[442]

Technical treatments are a different line of defense: They significantly reduce the chances of deanonymizing a dataset. There are a range of technical methods that can enable analysis while ensuring privacy:

- Techniques like k-anonymity and ℓ-diversity attempt to offer group-based anonymization by reducing the granularity of individual records in tabular data.[443] While effective in simple settings and easy to implement, both methods are susceptible to attacks by adversaries who possess additional information about the individuals in the dataset.
- One of the most popular techniques is differential privacy,[444] which provides provable guarantees on privacy, even when an adversary possesses additional information about records in the dataset. However, differential privacy requires adding random amounts of statistical "noise" to data and can sometimes compromise the accuracy of data analyses.

---

https://arxiv.org/pdf/2012.07805.pdf.

[442] *See generally, HIPAA Training, Certification, and Compliance*, HIPAA TRAINING, https://www.hipaatraining.com/; *Research Data Management*, UK DATA SERV., https://ukdataservice.ac.uk/learning-hub/research-data-management/ (last visited Feb. 20, 2022).

[443] Ashwin Machanavajjhala et al., *L-Diversity: Privacy Beyond K-Anonymity*, 22 INT'L CONF. DATA ENG'G 24 (2006).

[444] CYNTHIA DWORK & AARON ROTH, THE ALGORITHMIC FOUNDATIONS OF DIFFERENTIAL PRIVACY (2014).

Although differential privacy has become a point of contention with respect to the Census Bureau's new disclosure avoidance system,[445] the technique remains a powerful defense against bad actors seeking to take advantage of public data for the purposes of re-identification.

- Researchers have also identified other promising methods. Recent work has demonstrated that machine learning can be used to generate "synthetic" datasets, which mirror real-world datasets in important ways but consist of entirely synthetic examples.[446] Other work has focused on the incorporation of methods from cryptography, including secure multiparty computation[447] and homomorphic encryption.[448]

Methods that obscure data introduce fundamental tensions with the way machine-learning researchers develop models. For example, when considering questions of algorithmic fairness, in some instances privacy protections can undercut the power to assess whether such a technical method as differential privacy results in demographic disparities, particularly for small subgroups.[449] Similarly, "error analysis"—the study of samples over which a machine-learning model performs poorly—is central to how researchers improve models. It requires understanding the attributes and characteristics of the data in order to better understand the deficiencies of an algorithm. Therefore, such methods as differential privacy, which make raw data more opaque, will invariably impede the process of error analysis. Synthetic data

---

[445] *See, e.g.,* Tara Bahrampour & Marissa J. Lang, *New System to Protect Census Data May Compromise Accuracy, Some Experts Way*, WASH. POST (June 1, 2021), https://www.washingtonpost.com/local/social-issues/2020-census-differential-privacy-ipums/2021/06/01/6c94b46e-c30d-11eb-93f5-ee9558eecf4b_story.html; Kelly Percival, *Court Rejects Alabama Challenge to Census Plans for Redistricting and Privacy*, BRENNAN CTR. (June 30, 2021), https://www.brennancenter.org/our-work/analysis-opinion/court-rejects-alabama-challenge-census-plans-redistricting-and-privacy.

[446] *See, e.g.,* LEONARD E. BURMAN ET AL., SAFELY EXPANDING RESEARCH ACCESS TO ADMINISTRATIVE TAX DATA: CREATING A SYNTHETIC PUBLIC USE FILE AND A VALIDATION SERVER (2018); *see also* THE SYNTHETIC DATA VAULT, https://sdv.dev (last visited Feb. 28, 2022).

[447] Valerie Chen, Valerio Pastro & Mariana Raykova, *Secure Computation for Machine Learning with SPDZ*, ARXIV (Jan. 2, 2019), https://arxiv.org/pdf/1901.00329.pdf.

[448] Louis J. M. Aslett et al., *A Review of Homomorphic Encryption and Software Tools for Encrypted Statistical Machine Learning*, ARXIV (Aug. 26, 2015), https://arxiv.org/pdf/1508.06574.pdf.

[449] *See* Hongyan Chang & Reza Shokri, *On the Privacy Risks of Algorithmic Fairness*, ARXIV (Apr. 7, 2021), https://arxiv.org/pdf/2011.03731.pdf.

typically captures relationships between variables only if those relationships have been intentionally included in the statistical model that generated the data,[450] and thus, may be poorly suited to certain AI models that discover unanticipated relationships among data. While homomorphic encryption may not require similar assumptions on data structure, existing methods are computationally expensive.

While promising, understanding and applying these methods is an evolving scientific process. The NRC is poised to contribute to their evolution by directly supporting research into their application.

A. Criteria and Process for Adoption

The NRC will contain a rich array of datasets, each presenting unique privacy implications over different types of data formats (e.g., individual tabular records, unstructured text, images). Including a dataset on the NRC raises a question of choice: Which technical privacy treatment should be applied (e.g., *k*-anonymity vs. differential privacy), and how should it be applied? This question often requires technical determinations about different algorithmic settings, but such technical choices can also have important substantive consequences.[451]

First, we recommend that these determinations are made with respect to the following factors:

- **Dataset sensitivity:** Different datasets will pose privacy risks that range in type and magnitude. Health records, for instance, are more sensitive than weather patterns. The privacy method chosen should reflect this sensitivity. As we discuss in Section 3, these privacy methods should correspond and be tiered to the appropriate FedRAMP classification for the dataset.

- **Dataset utility:** As discussed above, applying a privacy method can distort the original data, diminishing the accuracy

---

[450] Ruggles et al., *Differential Privacy and Census Data: Implications for Social and Economic Research*, 109 AM. ECON. ASS'N PAPERS & PROCS. 403, 406 (2019).

[451] In Computer Science literature, such algorithmic settings are often referred to as *hyperparameters*. For instance, *k* is a hyperparameter for *k*-anonymity. By setting *k* to different values (e.g., 5, 10, 100), practitioners can modulate the amount of anonymity afforded to records in the data. As we note however, the choice of hyperparameters controls both the privacy effected on a dataset as well as the fidelity of that data.

and utility of analysis. Because different methods affect different levels of distortion, the choice of method should be informed by the perceived utility of the data. High-utility datasets—where accurate analyses are highly important (e.g., medical diagnostic tools)—may necessitate methods that produce less distortion.

- **Equity:** Certain privacy measures can disproportionately impact underrepresented subgroups in the data.[452] In determining which method to apply, the presence of sensitive subgroups and their relation to the objectives of the dataset should be evaluated.

For any given dataset, we recommend that agencies providing the data collaborate with NRC staff to identify and recommend any privacy treatments. Originating agencies and NRC staff will possess domain and research expertise to make evaluations on the balance of privacy, utility, and equity, but agencies should consult with NRC staff and researchers on the most appropriate treatments. Given the cost of review, such privacy treatments should be much less widely considered for low-risk datasets.

B. Virtual Data Safe Rooms

For individual research proposals that would be greatly hampered by technical privacy measures, the NRC should explore the use of virtual "data-safe rooms" that enable researchers to access raw administrative microdata in a secure, monitored environment. Currently, the Census Bureau implements these safe rooms in physical locations and moderates access to raw interagency data through its network of Federal Statistical Research Data Centers (FSRDCs). However, the NRC should not adopt the FSRDC model wholesale. Indeed, the barriers to using FSRDCs are

---

[452] *See Differential Privacy for Census Data Explained*, NAT'L CONF. OF STATE LEGISLATURES (July 1, 2021), https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx; Hongyan Chang & Reza Shokri, *On the Privacy Risks of Algorithmic Fairness*, ARXIV (Apr. 7, 2021), https://arxiv.org/pdf/2011.03731.pdf. Some scholars even find that the incorporation of differential privacy into machine learning algorithms can have disparate impact on underrepresented groups. *See* Eugene Bagdasaryan & Vitaly Shmatikov, *Differential Privacy Has Disparate Impact on Model Accuracy*, ARXIV (Oct. 27, 2019), https://arxiv.org/pdf/1905.12101.pdf.

high, and "only the most persistent researchers are successful."[453] For instance, applying for access and gaining approval to use an FSRDC takes at least six months, requires obtaining "Special Sworn Status," which involves a Level Two security clearance, and is limited to applicants who are either U.S. citizens or have been U.S. residents for three years.[454] To further complicate matters, agencies have different review and approval processes for research projects that wish to access agency data using an FSRDC.[455] Finally, even after approval is granted, researchers can only access the data in person by going to secure locations, such as the FSRDC itself.[456]

To be clear, some of these restrictions are unique to the Census Bureau. U.S. law provides that any Census datasets that do not fully protect confidentiality may only be used by Census staff.[457] Researchers trying to access such data therefore must go through the rigorous process of becoming a sworn Census contractor. The extent to which these restrictions apply to the NRC will depend on whether the NRC institutionally houses itself in the Census Bureau, which we ultimately do not recommend.[458] Other problems, however, such as the lack of interagency uniformity in granting access to datasets is not a problem unique to Census, but a common problem throughout the federal government (see Section 3).

Another common problem—not necessarily tied to FSRDCs or the Census Bureau—is the use of a physical data room to access raw microdata. The NRC should explore a virtual safe room model, whereby researchers can *remotely* access such microdata. For instance, in the private sector, the nonpartisan and objective research organization, NORC, located at the University of Chicago, is a confidential, protected environment where authorized researchers can securely store, access, and analyze sensitive microdata remotely.[459] Some federal government

---

453 STEVEN RUGGLES, DIFFERENTIAL PRIVACY AND CENSUS DATA: IMPLICATIONS FOR SOCIAL AND ECONOMIC RESEARCH 17.
454 *Id.* at 18-19.
455 NAT'L ACAD. OF SCI., INNOVATIONS IN FEDERAL STATISTICS 86 (2017). The fragmented FSRDC review process is similar to the fragmented data access regime we discussed in Section 3.
456 *Special Sworn Researcher Program*, BUREAU OF ECON. ANALYSIS, https://www.bea.gov/research/special-sworn-researcher-program (last updated Nov. 5, 2021).
457 13 U.S.C. § 9.
458 The institutional form of the NRC is discussed in depth in Section 4.
459 NORC, NORC DATA ENCLAVE (June 2, 2016), https://www.norc.org/PDFs/BD-Brochures/2016/Data%20Enclave%20One%20Sheet.pdf.

agencies have also implemented their own virtual data safe rooms. The Center for Medicare and Medicaid Services' Virtual Research Data Center (VRDC), for instance, grants researchers direct access to approved data files through a Virtual Private Network.[460] In a 2019 Request for Information (RFI), the National Institutes of Health also solicited input for its own administrative data enclave and whether such an enclave should be physical or virtual.[461] As articulated in responses to the RFI from the American Society for Biochemistry and Molecular Biology and the Federation for of American Societies for Experimental Biology, a virtual enclave would greatly facilitate researcher access to data and can be designed and administered in a way to preserve privacy and security.[462]

A National Research *Cloud* cannot function effectively if access to certain datasets is ultimately tied to a National Research *Room*.

CASE STUDY: California Policy Lab

The California Policy Lab (CPL) is a University of California research institute that provides research and data support to help California state and local governments craft evidence-based public policy.[463] CPL offers a variety of services to governments, including data analysis services and secure infrastructure for hosting and linking the vast amounts of data collected by government entities.[464] These services help bridge the gap between academia and government by helping policymakers gain access to researchers and providing researchers a secure way to access administrative data. CPL aims to build trusting partnerships with government entities and enable them to make empirically supported

---

[460] *CMS Virtual Research Data Center (VRDC)*, RSCH. DATA ASSIST. CTR., https://resdac.org/cms-virtual-research-data-center-vrdc (last visited Feb. 28th, 2022).
[461] *Request for Information (RFI) Seeking Stakeholder Input on the Need for an NIH Administrative Data Enclave*, NAT'L INST. OF HEALTH (Mar. 1, 2019), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-085.html.
[462] *See FASEB Response to NIH Request for Information (RFI): Seeking Stakeholder Input on the Need for an NIH Administrative Data Enclave*, FED'N OF AM. SOC'YS FOR EXPERIMENTAL BIOLOGY (2019), https://www.faseb.org/Portals/2/PDFs/opa/2019/FASEB_Response_Data_Enclave_RFI_NOT-OD-19-085.pdf; AM. SOC'Y OF BIOCHEMISTRY & MOLECULAR BIOLOGY (May 30, 2019), https://www.asbmb.org/getmedia/e3401ed5-3210-4ed2-a82a-7363cb86071d/ASBMB-Response-to-NIH-RFI-NOT-09-19-085.pdf.
[463] *What We Do,* CAL. POL'Y LAB, https://www.capolicylab.org/what-we-do/ (last visited Feb 28, 2022).
[464] *Id.*

policy decisions.

CPL enters data-use agreements with various government entities around California, including, for example, the California Department of Public Health and Los Angeles Homeless Services Authority.[465] These agreements allow CPL to store administrative data in a linkable format, promoting broad longitudinal analyses across various public sector domains.

To help manage the requirements of the various data-use agreements and simplify compliance, CPL applies the strictest requirements for any individual data across all data it stores.[466] Each set of administrative data is thus subject to strict technical restrictions and thorough audits.[467] CPL manages the data in an on-premises data hub at UCLA. This data hub uses "virtual enclaves" modeled after air-gapped clean rooms typically used for sensitive government data.[468] Virtual enclaves are virtual machines that forbid any outbound connections.

CPL creates a new virtual enclave for each research project and only gives specific researchers access to specific datasets for each project.[469] Researchers can only work with the data in the enclave and can only use tools provided in the environment. Data access processes vary, based on the requirements of the government entities, and most of CPL's data-use agreements are purpose limited and thus require approval from the relevant government entity before being used in a project.[470]

Generally, CPL helps researchers understand how to gain access to various types of administrative data. For some datasets, CPL has

---

[465] *CPL Roadmap to Government Administrative Data in California*, CAL. POL'Y LAB, https://www.capolicylab.org/data-resources/california-data-roadmap/ (last visited Feb. 28, 2022).
[466] Interview with Evan White, Executive Director, California Policy Lab (Apr. 29, 2021).
[467] *Id.*
[468] *Id.*; *see, e.g., Policy Evaluation and Research Linkage Initiative (PERLI)*, CAL. POL'Y LAB, https://www.capolicylab.org/data-resources/perli/ (last visited Feb. 28, 2022); *University of California Consumer Credit Panel*, CAL. POL'Y LAB, https://www.capolicylab.org/data-resources/university-of-california-consumer-credit-panel/ (last visited Feb. 28, 2022).
[469] Interview with Evan White, *supra* note 466.
[470] *Id.*

formalized applications on its website.[471] CPL prescreens project proposals and sends promising projects to its government partners for final approval. Researchers then conduct these approved projects on CPL's secure infrastructure. For other datasets without formalized access processes, CPL directs researchers toward individuals within the government entities.[472] CPL can then take over management of approved projects that aim to use data stored on its hub under their standing data-use agreements. Alternatively, the government entities and researchers themselves may craft new data-use agreements for specific projects.

C. Implications for the NRC

1. Dedicated Staff

As discussed above, it will be critical for the NRC to maintain dedicated professional staff who specialize in privacy technologies. First, not all agencies or departments that seek to place data into the NRC will have the expertise to both determine the privacy method that meets data utility expectations and data privacy demands and apply it to the dataset of interest. Specialized NRC staff will be essential to assisting such agencies and departments. Second, even where agencies and departments do possess the requisite expertise, NRC staff will bring a unique perspective from their collaborations across the government. Where a specific department's staff may only foresee risks specific to the dataset, NRC staff will be able to foresee instances where the presence of other data in the NRC may raise other concerns. In fact, by working with Affiliated Government Agencies and agency representatives, the NRC staff can also help these agencies internalize such benefits, such as helping them understand the full range of privacy risks with respect to their data.[473] Such collaborative governance will be necessary to ensure that privacy assessments consider the full implications of access and privacy technologies. Finally, it must not be overlooked that while data management, in general, requires technical expertise, these various privacy-enhancing technologies also require very specific, highly skilled

---

[471] *See, e.g., Life Course Dataset*, CAL. POL'Y LAB, https://www.capolicylab.org/life-course-dataset/ (last visited Feb 28, 2022).
[472] *See CPL Roadmap to Government Administrative Data in California*, *supra* note 465.
[473] We note that it is possible that the organizational form could affect the authority of NRC staff to speak to the legality of data transfers.

expertise. Using synthetic data sets as an example, NRC staff could be asked to build synthetic data on an agency's behalf, or need to validate the work performed at an agency to ensure it is done properly and well. Whatever the task, there are cascading effects downstream through the research ecosystem if not carefully managed and executed.

2. A Focus on Evaluating and Researching Privacy-Enhancing Technologies

It will be necessary to continually evaluate the state of privacy protections on the NRC, either by NRC staff members or by supporting privacy and security researchers at academic institutions. Technical privacy and security research is by nature adversarial: Researchers adopt the posture of adversaries in order to probe the weaknesses of a system/dataset. In the context of the NRC, this will require simulating attacks as researchers try to reidentify individuals within specific NRC datasets. This type of research is necessary to advance the field, and the NRC may be specially positioned to support a research center devoted to researching privacy-enhancing technologies. Doing so would allow the research community to build stronger privacy methods to ensure anonymity, identify flaws, and self-regulate an evolving data ecosystem.

VII. SAFEGUARDS FOR ETHICAL RESEARCH

The pace of advances in AI has sparked ample debate about the principles that should govern its development and implementation. Despite the technology's promise for economic growth and social benefits, AI also poses serious ethical and societal risks. For example, studies have demonstrated AI systems can propagate disinformation,[474] harm labor and employment,[475] demonstrate algorithmic bias along age, gender, race, and disability,[476] and perpetuate systemic inequalities.[477]

---

[474] Christopher Whyte, *Deepfake News: AI-Enabled Disinformation as a Multi-Level Public Policy Challenge*, 5 J. CYBER POL'Y 199 (2020); Don Fallis, *What Is Disinformation?*, 63 LIBR. TRENDS 601 (2015).
[475] MARY L. GRAY & SIDDHARTH SURI, GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS (2019); *Science Must Examine the Future of Work*, NATURE (Oct. 19, 2017), https://www.nature.com/articles/550301b.
[476] David Danks & Alex John London, *Algorithmic Bias in Autonomous Systems*, 26 INT'L JOINT CONF. ON A.I. 4691 (2017); Buolamwini & Gebru, *supra* note 71; Ben Hutchinson et al., *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities*, 125 ACM SIGACCESS ACCESSIBILITY & COMPUTING 1 (2020).
[477] OSCAR H. GANDY JR., THE PANOPTIC SORT: A POLITICAL ECONOMY OF PERSONAL

This section considers how the NRC should ensure its resources are deployed responsibly and ethically. A growing body of research on AI fairness, accountability, and transparency has raised serious and legitimate questions about the values implicated by AI research and its impact on society.[478] The NRC's focus on increasing access to sources of public data and fostering noncommercial AI research is intended to help address these concerns by enabling broader opportunities for academic research. At the same time, broadening access to resources is not enough to assure that academic AI research does not exacerbate existing inequalities or perpetuate systematic biases. In addition, the NRC must also be prepared to handle and act upon complaints of unethical research practices by researchers.

While there is an abundance of proposed ethics frameworks for AI (see Appendix C for those published by federal agencies), there is not a set of accepted principles enshrined into law, like the Common Rule for human subjects research, that clearly establishes the boundaries for ethical research with AI.[479] Lacking such guidance, a core question for the NRC is how to institutionalize the consideration of ethical concerns. This section starts by discussing two potential approaches for research proposals: *ex ante* review at the proposal stage for access to NRC resources (e.g., compute, dataset), and *ex post* review after research has concluded. Separately, we discuss guidance for the NRC on issues related to research practices. One of the virtues of starting with access by Principal Investigator (PI) status (Section 2) is that researchers will (a) often have undergone baseline training by their home institutions in

---

INFORMATION (1993); EUBANKS, *supra* note 72; Rashida Richardson, *Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities*, 36 BERKELEY TECH. L.J. 101 (2021).

[478] For approaches to improve machine learning practices, see Timnit Gebru et al., *Datasheets for Datasets*, ARXIV (Mar. 19, 2020), https://arxiv.org/pdf/1803.09010.pdf; Margaret Mitchell et al., *Model Cards for Model Reporting*, 2019 PROCEEDINGS ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 220 (2019); Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need?*, 2019 CHI CONF. ON HUM. FACTORS IN COMPUTING SYS. 1 (2019); Michael A Madaio et al., *Co-Designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI*, 2020 CHI CONF. ON HUM. FACTORS IN COMPUTING SYS. 318 (2019). The literature on AI's societal impacts and fairness, accountability, and transparency of AI is vast, but see MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN (2019); EUBANKS, *supra* note 72; SOLON BAROCAS, MORITZ HARDT & ARVIND NARAYANAN, FAIRNESS AND MACHINE LEARNING (2019); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016).

[479] 45 C.F.R §§ 46.101-124.

research compliance, privacy, data security, and practices for research using human subjects; and (b) be subject to research standards and peer review (e.g., through IRB review when applicable). These mechanisms are insufficient to cover many AI research projects, such as when human subjects review is deemed inapplicable. Thus, we tailor our recommendations to the institutional design of the NRC.

First, we recommend that the NRC require including an ethics impact statement for PIs requesting access beyond base-level compute, or for research using restricted datasets. This provides a layer of ethical review for the highest resource projects that are already required to undergo a custom application process. Second, for other categories of research (e.g., research conducted under base-level compute access, where no custom review is contemplated), we recommend that the NRC establish a process for handling complaints that may arise out of unethical research practices and outputs. Third, given the limitations of the prior mechanisms, we recommend the exploration of a range of measures to address ethical concerns in AI compute, such as the approach taken by the National Institutes of Health to incentivize the embedding of bioethics in ongoing research.

A. Ethics Review Mechanisms

1. Ex Ante

Ex ante review assesses research yet to be performed.[480] Funding agencies and research councils worldwide rely on ex ante peer reviews to evaluate the intellectual merit and potential societal impact of research proposals, based on set criteria.[481] Institutional Review Boards (IRBs) commonly assess academic research involving human subjects prior to its initiation.[482] However, much AI-related research may not fall under IRB oversight. For instance, such research might not use human subjects. Alternatively, such research might not rely on existing data (not collected by the proposers) about people that is publicly available,[483] used with

---

[480] J. Britt Holbrook & Robert Frodeman, *Peer Review and the Ex Ante Assessment of Societal Impacts*, 20 RES. EVALUATION 239 (2011).
[481] *Id.*
[482] *Institutional Review Boards (IRBs) and Protection of Human Subjects in Clinical Trials*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/institutional-review-boards-irbs-and-protection-human-subjects-clinical-trials (last updated Sept. 11, 2019).
[483] There are crucial questions with regards to consent even with data considered

permission from the party that collected the data, or is anonymized. Potential ethical issues may, therefore, escape IRB review.[484]

Creating an across-the-board ex ante ethics review process, however, would be challenging. First, as we discuss in Section 2, we recommend against case-by-case review for all PI requests for access to NRC compute and data, as such a process would require substantial administrative overhead. At the stage when researchers are simply applying for compute access, the research may be so varied and early stage, that there is not much concrete substance to review. And to the extent that every PI would require project-specific review, such a process would be onerous.

Second, ex ante review is unlikely to grapple with the many ethical implications of design decisions that take place after research commences.[485] Research design can change substantially from initial proposals as projects progress. Ex ante review could identify some concerns, but unlikely all.[486] The nature of machine learning is inherently uncertain and predictions can be challenging to explain as well as highly dependent on the data used to build and train models.[487] Ex ante proposal review alone may not be sufficient to identify biased outcomes and may in fact require extensive documentation and review of the data used in a specific project to assess with any reliability.[488]

Third, there are unique academic speech concerns about government assessment of research. Authorizing the government to

---

"publicly" available. *See generally* Casey Fiesler & Nicholas Proferes, *"Participant" Perceptions of Twitter Research Ethics*, 4 SOCIAL MEDIA + SOCIETY 1 (2018); Sarah Gilbert, Jessica Vitak & Katie Shilton, *Measuring Americans' Comfort with Research Uses of Their Social Media Data*, 7 SOCIAL MEDIA + SOCIETY 1 (2021).

[484] SARA R. JORDAN, DESIGNING AN ARTIFICIAL INTELLIGENCE RESEARCH REVIEW COMMITTEE, FUTURE OF PRIV. F. (2019), https://fpf.org/wp-content/uploads/2019/10/DesigningAIResearchReviewCommittee.pdf.

[485] Agatta Feretti et al., *Ethics Review of Big Data Research: What Should Stay and What Should Be Reformed?*, 22 BMC MEDICAL ETHICS 1, 6 (2021); Kathryn M. Porter et al., *The Emergence of Clinical Research Ethics Consultation: Insights from a National Collaborative*, 2018 AM. J. BIOETHICS 39 (2018).

[486] Feretti et al., *supra* note 485, 1-3.

[487] *See, e.g.,* Mark Diaz et al., *Addressing Age-Related Bias in Sentiment Analysis*, 2018 PROCEEDINGS CHI CONF. ON HUM. FACTORS IN COMPUTING SYS. 1 (2018); Buolamwini & Gebru, *supra* note 71.

[488] *See, e.g.,* Gebru et al., *supra* note 478; Mitchell et al., *supra* note 478; Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 PROCS. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 610 (2021); Christo Wilson et al., *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, 2021 PROCS. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 666 (2021); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017).

conduct an ethics review (separate from IRB review under the Common Rule, which is typically delegated to academic institutions) with vague standards may implicate academic speech concerns, as well as subject proposals to politically-driven evaluations that can shift from administration to administration.

If the NRC were to create a process for ex ante review of research proposals for ethical concerns, such a board would likely need to be composed of scientific and ethics experts, similar to how the NSF conducts its process, though perhaps with the addition of members from civil society organizations that focus on countering AI harms. The NSF convenes groups of experts from academia, industry, private companies, and government agencies as peer reviewers, led by NSF program officers and division directors.[489] However, the scope and range of NRC research proposals are likely to be both broad and highly interdisciplinary in nature, making ethics assessments challenging.

2. Ex Post

Ex post evaluations provide an assessment after research has concluded.[490] In academia, researchers submit research results to journals or conferences for ex post peer review. It is at this pre-publication stage that ethical issues not identified by ex ante processes may be surfaced by reviewers or editors. In the public sector, for example, the Privacy and Civil Liberties Oversight Board (PCLOB) conducts ex post reviews on counterterrorism practices by executive branch departments and agencies to ensure they are consistent with governing laws, regulations, and policies regarding privacy and civil liberties.[491] PCLOB has also recently begun to evaluate the use of new technologies in foreign intelligence collection and analysis[492] and to identify legislative proposals that strengthen its oversight of AI for counterterrorism.[493]

---

[489] *Phase II: Proposal Review and Processing*, NAT'L SCI. FOUND., https://www.nsf.gov/bfa/dias/policy/merit_review/phase2.jsp#select (last visited Feb. 22, 2022).
[490] Harvey A. Averch, *Criteria for Evaluating Research Projects and Portfolios*, *in* EVALUATING R&D IMPACTS: METHODS & PRACTICE 263 (1993).
[491] NAT'L SEC.Y COMM'N ON A.I, FINAL REPORT 14–54 (2021).
[492] *History and Mission*, U.S. PRIV. & C.L. OVERSIGHT BD., https://www.pclob.gov/About/HistoryMission (last visited Feb 28, 2022).
[493] AI in Counterterrorism Oversight Enhancement Act of 2021, H.R. 4469, 117th Cong. (2021).

B. Recommendations

While we do not recommend an across-the-board ex ante review of research proposals, we do recommend that the NRC establish a process to handle complaints about ethical research practices and outputs. On that point, we recommend the NRC collaborate with the Office of Research Integrity (ORI) at the Department of Health and Human Services to model its processes and procedures for managing issues of research misconduct.[494] The ORI has substantial experience overseeing concerns about ethical research practices. Parties could petition the NRC to revoke access when research is shown to manifestly violate general ethical research standards or practices applicable to a researcher's disciplinary domain. We note that the NRC may want to adopt a high standard for such a violation, given the academic speech considerations. For example, federal agencies or external parties that wish to revoke compute or data access from PIs would need to file a written complaint with supporting evidence. Decisions to revoke access should require input from NRC executive leadership and legal counsel.

For PIs requesting access beyond base-level compute or for restricted datasets, we recommend requiring the completion of ethics impact statements to be submitted with research proposals. A recent proposal to address the lack of "widely applied professional ethical and societal review processes" in computing piloted such a requirement in a grant process, requiring a description of the potential social and ethical impacts and mitigation efforts by researchers.[495] We limit this approach to proposals for compute access beyond default allocation or requests for access to restricted datasets, as the administrability concerns are weaker for researchers who are already applying for compute or data access beyond the default levels. For those applications, a review process of a specific proposal will already occur by an external review panel of experts (Section 2), and much like the NSF requires statements of "Broader Impacts,"[496] statements about the ethical considerations of the work could easily be included. It is important to note that ethics impact

---

[494] In instances where a researcher is using data obtained from one of the agencies that falls under ORI's oversight, it may make sense to have ORI adjudicate those cases directly. For more information about the ORI, see *ORI*, OFF. OF RSCH. INTEGRITY, https://ori.hhs.gov/ (last visited Feb. 22, 2022)

[495] Michael S. Bernstein et al., *ESR: Ethics and Society Review of Artificial Intelligence Research*, ARXIV (July 9, 2021), https://arxiv.org/pdf/2106.11521.pdf.

[496] *Broader Impacts*, NAT'L SCI. FOUND., https://www.nsf.gov/od/oia/special/broaderimpacts/ (last visited Apr. 10, 2022).

statements would be only one component of NRC applications and should be weighed in conjunction with other application materials. In addition to requiring researchers to carefully think through and document the potential impacts of their own work, the statements may also serve as useful documentation of potential negative impacts and be of use to NRC staff when determining whether to provide access to specific types of data. Such assessments may also be helpful for journals, conferences, or universities addressing ex post concerns about ethical impacts.

Next, we recommend that the NRC employ a professional staff devoted to ethics oversight, similar to what we propose regarding data privacy in Sections 5 and 6. In addition to staff devoted to handling legal compliance issues, the NRC needs staff with specialized training in AI ethics (as well as expertise in other subdomains) to provide expert internal consulting to NRC applicants, as well as to aid in evaluating ethics impact statements. Similarly, data privacy experts can identify ethical privacy issues specifically related to data, such as whether consent has been properly obtained and documented. To ensure that decisions are based on the merits, the NRC staff overseeing these issues must operate independently of other federal agencies and be insulated from political interference.

We acknowledge that these ethics review mechanisms may not identify all instances where researchers use the NRC in a way to conduct research that raises ethical questions. Few review mechanisms could do so, particularly in light of the considerable ambiguity present in ethics standards (see Appendix C). Nonetheless, these mechanisms can augment key academic checkpoints (IRB review and peer review) in an administrable fashion that does not raise serious concerns about academic speech.

Lastly, we recommend that non-NRC parties explore a range of measures to address ethical concerns in AI compute. These may include an ethics review process or approaches widely deployed in bioethics by the National Institutes of Health, namely to incentivize the embedding of ethicists in research projects.[497] Such embedded ethics approaches may

---

[497] *See, e.g., Notice of Special Interest: Administrative Supplement for Research and Capacity Building Efforts Related to Bioethical Issues*, NAT'L INST. OF HEALTH (Nov. 17, 2020), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-020.html; *Notice of Special Interest: Administrative Supplement for Research on Bioethical Issues*, NAT'L INST. OF HEALTH (Dec. 30, 2019), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-038.html; *see also*

have the particular advantage of identifying and addressing issues as the research proceeds, in contrast to ex ante review, where it may be too early to spot an issue, and ex post review, which may be too late. We expect this to be an active area of inquiry as new approaches are validated. The NRC, potentially in conjunction with the NSF, should consider offering funding for projects that embed ethics domain experts into teams, in order to support this proposal.

VIII. MANAGING CYBERSECURITY RISKS

While the NRC has the potential to level the playing field for AI research, it will also create an alluring target for a vast array of bad actors.

Cybersecurity—the effort to protect systems against incidents that may compromise operations or cause harm to relevant assets and parties—will be a critical focus of the NRC. It will require a cybersecurity framework that manages potential incidents throughout their *lifecycle,* spanning: (1) preparation; (2) detection and analysis; (3) containment, eradication, and recovery; and (4) post-incident activity, which collectively encompasses incident monitoring, detection, recovery, and reporting.[498] Effective cybersecurity practices complement risk assessment based on impact, immediacy, and likelihood, and will help gain the trust of users and thwart subversion and interference from foreign actors or other adversarial parties. Careful administrative design of the NRC with cybersecurity at the forefront will set a high standard as information systems become more central to our national infrastructure.

In this section, we address these cybersecurity concerns. We first provide an overview of common types of vulnerabilities and attacks and assess their relevance to the NRC. Next, we provide an overview of the federal government's regulatory landscape, as it pertains to cybersecurity, with a special focus on the FISMA and FedRAMP frameworks. Finally, we close with a discussion of the security and

---

Courtenay R. Bruce et al., *An Embedded Model for Ethics Consultation: Characteristics, Outcomes, and Challenges*, 5 AJOB EMPIRICAL BIOETHICS 8 (2014); Sharon Begley, *In a Lab Pushing the Boundaries of Biology, an Embedded Ethicist Keeps Scientists in Check*, STAT (Feb. 23, 2017), https://www.statnews.com/2017/02/23/bioethics-harvard-george-church/. Private foundations also promote the use of embedded bioethicists. *See, e.g., Making a Difference Request for Proposals – Fall 2021*, THE GREENWALL FOUND. (2021), https://greenwall.org/making-a-difference-grants/request-for-proposals-MAD-fall-2021.

[498] PAUL CICHONSKI ET AL., COMPUTER SECURITY INCIDENT HANDLING GUIDE (2012).

system design measures best suited to ensure that the integrity of the NRC is not compromised.

A. Motivations for Potential Attacks

Possible attacks against the NRC could take a number of approaches, each of which would entail substantial consequences for the NRC.[499] First, adversaries could launch an attack against the NRC with the intention of disrupting its operations or its ability to aid research. For example, adversaries could attack the NRC's infrastructure directly by disabling or interfering with NRC servers. As a result, researchers would be unable to access NRC servers or effectively utilize them. By launching such attacks, adversaries may throttle the NRC, thereby raising costs for the federal government.[500] Alternatively, adversaries could seek to attack specific research projects on the NRC, thereby slowing the pace of that research or compromising the quality of the research findings. They may also initiate "data-poisoning" attacks on NRC datasets, thereby compromising the quality of research findings.

Second, bad actors could also launch cyber operations against the NRC, intending to steal computational resources. In this case, the purpose would not be to disrupt the NRC but to repurpose computational power toward illicit purposes (e.g., cryptocurrency mining).[501] For instance, individuals could pretend to be researchers, claiming to use cloud credits for legitimate research purposes while actually using them for alternative ends. Individuals could also infiltrate the NRC's network, siphoning off computational resources from other projects and reducing the functionality for legitimate users.

Third, adversaries might pose a threat to the NRC out of a desire to steal or make use of the data and research products housed within the system. The NRC promises to be an attractive target because it will house data from a range of different agencies. If an adversary wanted to steal

---

[499] Putative attacks could include the deployment of ransomware, phishing schemes, gaining root access (the highest level of privilege available which gives users access to all commands and files by default), exposure of secret credentials, data poisoning, data exfiltration, as well as other types of unauthorized network intrusions.

[500] Karen Hao, *AI Consumes a Lot of Energy. Hackers Could Make it Consume More*, MIT TECH. R. (May 6, 2021), https://www.technologyreview.com/2021/05/06/1024654/ai-energy-hack-adversarial-attack/.

[501] Catalin Cimpanu, *Vast Majority of Cyber-Attacks on Cloud Servers Aim to Mine Cryptocurrency*, ZDNET (Sept. 14, 2020), https://www.zdnet.com/article/vast-majority-of-cyber-attacks-on-cloud-servers-aim-to-mine-cryptocurrency/.

equivalent data from the agencies themselves, they would need to break into each agency independently. However, the potential combination of datasets on the NRC, including researcher-owned datasets, may increase the potential gains from accessing this information. Additionally, adversaries may attempt to break into the NRC in order to steal products generated by NRC researchers. This could include trained machine-learning models or specific research findings.

Relatedly, bad actors could determine that executing intrusions into the NRC is an effective way to target Affiliated Government Agencies. Because a participation incentive for agencies is the computing support that the NRC will offer, one of the biggest cyber risks is malicious actors attempting to use the NRC to hack into their systems. For that reason, the cybersecurity risk to the government may be substantial. On the other hand, as we discussed in Section 3, the NRC also presents an opportunity to enhance and harmonize security standards compliance, as agencies move into the cloud.

A range of other motivations may exist. Successful operations against the NRC, as a federal entity, would carry symbolic value and capture attention. Ransomware attacks could result in significant payoffs. The NRC could also be a target for espionage, both on the part of nation-state actors seeking to acquire sensitive datasets (e.g., energy grid infrastructure) and on the part of private sector entities looking to steal intellectual property or to monitor the latest technological advances.

If successful, any attack could undermine the NRC. For example, researchers would be deterred from using the NRC and may invest their efforts into alternate private clouds. This could occur because researchers believe the NRC would be ineffective to use (e.g., on account of frequent server outages), or because they believe their research products would be inadequately protected. Federal agencies and departments could be deterred from entrusting the NRC with sensitive datasets. Federal entities could risk embarrassment and face obstacles in executing their policy objectives if datasets were accidentally leaked. If the NRC is insufficiently secure, such entities may choose to avoid sharing data altogether.

B. FISMA, FedRAMP, and Existing Federal Standards

As a federal entity, the NRC will be subject to federal standards and regulations. In this section, we provide a high-level overview of the two most relevant regulations: the Federal Information Systems

Management Act (FISMA) and the Federal Risk and Authorization Management Program (FedRAMP).[502] FISMA traditionally applies to non-cloud systems that support a single agency, whereas FedRAMP authorization is required for cloud systems.[503] We finish by discussing critiques of these regulations.

1. FISMA

The Federal Information Systems Management Act (FISMA) was first passed in 2002, with the purpose of providing a comprehensive framework for ensuring the effectiveness of security controls for federal information systems.[504] The law was later amended in 2014, and has since been augmented through other individual legislative and executive actions. Our discussion focuses on the collective impact of FISMA compliance regulations.[505]

FISMA applies to all federal agencies, contractors, or other sources that provide information security for information systems that support the operations and assets of the agency.[506] It invests responsibility in several different entities. First, the National Institute of Standards and Technology (NIST) is tasked with developing uniform standards and guidelines for implementing security controls, evaluating the riskiness of different information systems, and other methodologies.[507] Second, the Office of Management and Budget (OMB) is tasked with overseeing agency compliance with FISMA and reporting to Congress on the state of FISMA compliance.[508] Third, the Department of Homeland Security is tasked with administering the implementation of agency information security policies and practices.[509] Finally, federal

---

[502] We note that the NRC will likely need to comply with data specific security regulations as well. For instance, medical data security will need to comply with HIPAA, and financial data will need to comply with The Gramm-Leach-Bliley Act.
[503] Ray Dunham, *FISMA Compliance: Security Standards & Guidelines Overview*, LINFORD & CO. (Nov. 29, 2017), https://linfordco.com/blog/fisma-compliance/.
[504] AMY J. FRONTZ, REVIEW OF THE DEPARTMENT OF HEALTH AND HUMAN SERVICES COMPLIANCE WITH THE FEDERAL INFORMATION SECURITY MODERNIZATION ACT OF 2014 FOR FISCAL YEAR 2020 (2021).
[505] U.S. SENATE COMM. ON HOMELAND SEC. & GOV'T AFFS., FEDERAL CYBERSECURITY: AMERICA'S DATA AT RISK 18 (2019).
[506] *Federal Information Security Modernization Act (FISMA) Background*, NAT'L INST. STANDARDS & TECH., https://csrc.nist.gov/projects/risk-management/fisma-background (last updated Aug. 4, 2021).
[507] Dunham, *supra* note 503.
[508] U.S. SENATE COMM. ON HOMELAND SEC. & GOV'T AFFS., *supra* note 505 at 19.
[509] *Id.* at 18.

agencies are required to develop and implement a risk-based information security program in compliance with NIST standards and OMB policies.[510] Agencies are further required to conduct periodic assessments to ensure continued efficiency and cost-effectiveness.[511]

Several NIST requirements are worth mentioning here. Pursuant to NIST SP 800-18, agencies are required to identify relevant information systems falling under the purview of FISMA. Agencies must also categorize each of these systems into a risk level, following the guidance laid out in FIPS 199 and NIST 800-60.[512] NIST 800-53 outlines both the security controls that agencies should follow and the manner in which agencies should conduct risk assessments.[513] Agencies must further summarize both the security requirements and implemented controls in "security plans," as outlined in NIST 800-18.[514] Finally, organization officials are required to conduct annual security reviews in accordance with NIST 800-37.

2. FedRAMP

In the late 2000s, federal agencies began expressing security concerns as a barrier to cloud computing adoption.[515] In response, Congress passed the 2011 Federal Risk and Authorization Management Program (FedRAMP) to provide a cost-effective, risk-based approach for the adoption and use of cloud services by the federal government.[516] FedRAMP approval is exempted where: (i) the cloud is private to the agency; (ii) the cloud is physically located within a federal facility; and,

---

[510] *Id.* at 19.
[511] *Id.* at 20.
[512] KEVIN STINE, ET AL., GUIDE FOR MAPPING TYPES OF INFORMATION AND INFORMATION SYSTEMS TO SECURITY CATEGORIES (2008). Specifically, FISMA defines compliance in terms of three levels: low impact, moderate impact, and high impact. Low impact indicates that the loss of confidentiality, integrity, or availability of the system will have a limited adverse effect, while high impact indicates that such losses will have severe or catastrophic effects. *See* Sarah Harvey, *3 FISMA Compliance Levels: Low, Moderate, High*, KIRKPATRICKPRICE (Apr. 24, 2020), https://kirkpatrickprice.com/blog/fisma-compliance-levels-low-moderate-high/.
[513] NAT'L INST. STANDARDS & TECH., SECURITY AND PRIVACY CONTROLS FOR INFORMATION SYSTEMS AND ORGANIZATIONS (2020).
[514] MARIANNE SWANSON ET AL., GUIDE FOR DEVELOPING SECURITY PLANS FOR FEDERAL INFORMATION SYSTEMS (2006).
[515] McLaughlin, *supra* note 243.
[516] *Program Basics*, FEDRAMP, https://www.fedramp.gov/program-basics/ (last visited Feb. 22, 2022); *see also* STEVEN VANROEKEL, SECURITY AUTHORIZATION OF INFORMATION SYSTEMS IN CLOUD COMPUTING ENVIRONMENTS (2011).

(iii) the agency is not providing cloud services from the cloud-based information system to any external entities.[517] Like FISMA, FedRAMP security requirements are governed by NIST standards, including NIST SP 800-53, FIPS 199, NIST 800-37, and others.[518] However, unlike FISMA, FedRAMP's two tracks to receiving an authority-to-operate mean that vendors working with multiple agencies do not necessarily need to undergo the full approval process with each agency. This means that cloud services providers and agencies alike are able to save significant time and money.

3. Criticisms of FISMA and FedRAMP

These regulations are not without fault. Most notably, critics point to the fact that despite their existence, cyber intrusions on government infrastructure are common and accelerating.[519] A 2019 report by the U.S. Senate Committee on Homeland Security and Governmental Affairs investigating eight agencies noted that the federal government is failing its legislative mandate from FISMA.[520] The errors identified included a failure to protect personally identifiable information, inadequate IT documentation, poor remediation of bugs, a failure to upgrade legacy systems, and inadequate authority vested in agency chief information officers.[521] Reports by the Government Accountability Office (GAO) have reached similar conclusions.[522] In turn, some have criticized the government's approach to cybersecurity wholesale, arguing it places too much emphasis on merely detecting

---

[517] *FISMA vs. FedRAMP and NIST: Making Sense of Government Compliance Standards*, FORESITE, https://foresite.com/fisma-vs-fedramp-and-nist-making-sense-of-government-compliance-standards/ (last visited Feb. 22, 2022). However, we note that FedRAMP approval is exempted for certain types of cloud models: (i) where the cloud is private to the agency, (ii) where the cloud is physically located within a federal facility, (iii) where the agency is not providing cloud services from the cloud-based information system to any external entities. *See* VANROEKEL, *supra* note 516.

[518] FEDRAMP, FEDRAMP SECURITY ASSESSMENT FRAMEWORK 5 (2017).

[519] Doina Chiacu, *White House Warns Companies to Step Up Cybersecurity: 'We Can't Do it Alone'*, REUTERS (June 3, 2021), https://www.reuters.com/technology/white-house-warns-companies-step-up-cybersecurity-2021-06-03/; *see also Significant Cyber Incidents*, CTR. STRATEGIC & INT'L STUD., https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents (last visited Aug. 19, 2021).

[520] U.S. SENATE COMM. HOMELAND SEC. & GOV'T AFFS., *supra* note 505, at 5.

[521] *Id.* at 6.

[522] FRONTZ, *supra* note 504.

intrusions.[523] They argue for a framework of "zero trust," which assumes that intruders will penetrate a network and instead focus on security controls limiting the ability of those intruders to navigate the network.[524]

FedRAMP faces its own criticisms. A recent study noted that securing authorization can be time-consuming and expensive, taking up to two years and costing millions of dollars in some cases.[525] Even though parts of FedRAMP are designed to be reusable across agencies, agencies often delay the process by imposing separate, additional requirements. A variety of reasons for these deficiencies have been noted, including an understaffed Joint Authorization Board, a lack of trust between agencies with regards to Authorization to Operate (ATOs), and an overly complex authorization process that leads to errors by agencies and Cloud Services Providers.[526] Proposed recommendations to address these deficiencies include increased funding for FedRAMP's Joint Authorization Board, incentives to encourage the reuse of ATOs, and mechanisms to improve the efficiency of the authorization process.[527]

On May 12, 2021, the Biden administration released an Executive Order (EO) on Improving the Nation's Cybersecurity,[528] and OMB published a draft federal strategy for public comment on September 7, 2021.[529] Signed in the aftermath of the breach of the software vendor SolarWinds, and the ransomware attack on Colonial Pipeline, the EO presents several new initiatives. First, it calls on the federal government to embrace "zero-trust architecture" and improve post-attack investigation processes. Second, it seeks to improve collaboration between the public and private sectors by improving disclosure requirements and establishing a private-public Cybersecurity Safety Review Board (modeled after the National Transportation Safety Board). Finally, it seeks a more cohesive government-wide approach to cybersecurity, calling for the creation of a playbook to standardize cyber response across federal agencies, alongside a government-wide detection and response system for attacks.

---

[523] Jonathan Reiber & Matt Glenn, *The U.S. Government Needs to Overhaul Cybersecurity. Here's How.*, LAWFARE (Apr. 9, 2021), https://www.lawfareblog.com/us-government-needs-overhaul-cybersecurity-heres-how.
[524] NAT'L SEC. AGENCY, EMBRACING A ZERO TRUST SECURITY MODEL (2021).
[525] McLaughlin, *supra* note 243.
[526] *Id.*
[527] *Id.*
[528] Exec. Order No. 14,028, 86 Fed. Reg. 26633 (May 17, 2021).
[529] U.S. OFF. OF MGMT. & BUDGET, MOVING THE U.S. GOVERNMENT TOWARDS ZERO TRUST CYBERSECURITY PRINCIPLES (2021).

Though it is too soon to determine whether the EO and the proposed strategy will be effective, it appears to address deficiencies identified in the existing landscape. It seeks to improve documentation and responsiveness to attacks and suggests a shift in cybersecurity thinking. It is unclear, however, whether it will address the underlying procurement issues and lack of interagency trust that critics believe have hampered the effectiveness of FedRAMP. But given the potential for highly sensitive data to be stored on the NRC, embracing a zero-trust architecture at the outset is a crucial consideration for ensuring its integrity.

C. NRC Security Standards and System Design Measures

Here, we present recommendations on cybersecurity policy for the NRC informed by the landscape of the existing federal regulations and unique considerations that a national research cloud will pose.

1. Process for Risk and Security Determinations

Under the current regulatory landscape, agencies are responsible for determining the appropriate risk categorizations and security controls for the datasets located on their servers. However, this raises a potential challenge as agencies begin to share their data with the NRC—making it unclear who will maintain authority for categorizing the risk of these datasets and determining appropriate security controls.

On the one hand, agencies themselves could continue to retain discretion over the security classification and controls for datasets they place into the NRC. In this decentralized approach, much of the security responsibilities assigned by FISMA would remain with the agencies, irrespective of whether the data existed on NRC servers. On the other hand, the NRC could take responsibility for all security decisions. Datasets added to the NRC would then be classified according to the NRC's assessment of risk and protected with controls that the NRC staff deems appropriate. This approach "centralizes" security responsibilities by vesting it with the NRC after the onetime negotiation for each dataset.

Though both approaches have their merits, we recommend the centralized approach for several reasons. First, the centralized approach ensures internal uniformity. The paradox of federal cybersecurity regulation is that although NIST has articulated a set of standards pertaining to risk and controls, agencies interpret these standards differently, leading to discrepancies in implementation and classification

across the federal government. Following each agency's security classifications for data on the NRC would produce unnecessarily complex and incoherent classifications for a single system. This threatens to diminish the usability of the NRC, and the added complexity could arguably weaken security by increasing the likelihood of errors. Permitting the NRC to impose its own classifications allows for uniformity within the NRC and alignment with the access tiers suggested in Section 3 of this Section. This approach may also simplify managing security practices across a potential mix of cloud compute providers.

Second, the NRC represents a valuable opportunity to harmonize federal cybersecurity standards across different agencies. The assessments and implementations adopted by the NRC must generalize to the full diversity of federal datasets. Hence, the NRC's practices can serve as a template for NIST's guidelines, which any agency is free to adopt.

Third, the centralized approach will remove hurdles for data sharing. Security concerns often impede agency data sharing. In a scheme where agencies retain control over all security determinations, agencies could demand security classifications that are excessively high or impractical to implement. The centralized approach would place the burden on agencies to articulate with specificity why the NRC's security policies or classification guidelines are inadequate for a particular dataset.

Finally, researchers should also have a voice in determining the appropriate security controls, since a public resource of this magnitude that cannot attract users is bound to fail. As security controls implicate usability, the NRC should not opt for controls that substantially inhibit or disincentivize researchers from leveraging its resources. The NRC needs to strike the right balance between usability and security.

2. Technical Considerations

The federal government already possesses a range of technical options and countermeasures to cyberattacks. Cybersecurity threats and defenses are, of course, actively evolving, so we discuss these only as a starting point—robust, long-term cybersecurity comes through continued vigilance and prioritization that recognizes the shifting nature of the field.

3. Data Storage

Data storage mechanisms should ensure proper protection from outside access. Encryption can be used to protect sensitive data at rest, to be later unencrypted when needed. Physical isolation through air-gapped environments is another design feature that can remove the possibility of wireless network interfaces from being used to connect the data to malicious outside threats. However, even air gapping is not a foolproof solution. There are ways to "jump" air gaps such as through hiding in USB thumb drives (which is allegedly how the Stuxnet malware famously compromised Iranian nuclear centrifuges).[530] More recent attacks bypass the need for electronic transmission altogether by leveraging other signals that leak data, such as FM frequencies, audio, heat, light, and magnetic fields. These kinds of threats bring home the need for a comprehensive and evolving approach to cybersecurity.

4. Networking Protocols

Data packets sent over networks are transmitted according to a set of internationally standardized internet protocols. Following the Open Systems Interconnection (OSI) model, the conceptual layers involved in computer networking can be categorized into seven dimensions: physical, data link, network, transport, session, presentation, and application layers.[531]

5. Runtime Security

When considering runtime security technologies, three design features that are relevant for the cloud environments are the use of confidential clouds, federated learning, and cryptography-based measures such as homomorphic encryption and secure multiparty

---

[530] *See, e.g.*, David Kushner, *The Real Story of Stuxnet*, IEEE SPECTRUM (Feb. 26, 2013), https://spectrum.ieee.org/the-real-story-of-stuxnet.

[531] HTTP is the protocol at the highest level of abstraction targeting the application layer, and its secure variant HTTPS additionally encrypts the data using an encryption protocol. Without encryption, HTTP is insecure and should not be used. The encryption protocol in original use was SSL but this has since been deprecated in the realm of network security in favor of its newer version, TLS. Both SSL and TLS rely on public key certificates signed by a trusted certificate authority. When these certificates have expired, the websites providing them can no longer necessarily be trusted. Although these measures have their own limitations, not adopting them can only be less secure.

computation. A growing number of vendors offer "confidential cloud" options as an emerging technical solution to fully cyber secure cloud computation that is secure throughout execution.[532] Confidential clouds offer high-security, end-to-end, isolated operation by executing workloads within trusted execution environments. For example, virtualization enables an operating system to run another operating system within it as a virtual environment with additional firewall or other network barriers, effectively simulating another device within the host computer.

6. Distributed Computing and Federated Learning

Another computing paradigm, known as distributed computing or federated learning, considers situations where multiple parties have individual shards of data they are interested in leveraging in aggregate, without sharing outright. Federated learning addresses this situation, for example, demonstrating how users' mobile phones can send information—possibly differentially private—to central servers without exposing the precise details of any one individual's information. A second scenario more relevant to the large-scale decentralized nature of the NRC is distributed computing—in which many institutions collectively share compute, akin in some respects to crowd-sourced computing. These approaches enable multiple parties to leverage existing computational infrastructure, while retaining some guarantees on privacy.

7. Cryptography-Based Measures

Finally, there are two types of cryptography-based measures worth noting.

Cryptography researchers have developed ways of computing mathematical operations over *encrypted* data, known as homomorphic encryption. This impressive feat has valuable implications because it obviates the need for decryption, which can potentially expose the

---

[532] *See, e.g., Azure Confidential Computing*, MICROSOFT, https://azure.microsoft.com/en-ca/solutions/confidential-compute/ (last visited Feb. 22, 2022); Nataraj Nagaratnam, *Confidential Computing*, IBM (Oct. 16, 2020), https://www.ibm.com/cloud/learn/confidential-computing; *Confidential Computing*, GOOGLE CLOUD, https://cloud.google.com/confidential-computing (last visited Feb. 22, 2022).

intermediate values of computation, and grant access to public and secret encryption keys during computation. Initially, only partially homomorphic encryption schemes that supported limited arithmetic operations like addition and multiplication were possible. But fully homomorphic encryption schemes have recently been developed that enable what is known as "arbitrary" computation for promising use cases in predictive medicine and machine learning. That said, standardization is still underway to broader adoption, and homomorphic encryption (by design) is malleable—a property in cryptography that is usually undesirable as it allows attackers to modify encrypted ciphertexts without needing to know their decrypted value. These and other limitations of any technical approach are worth taking into account when considering which technologies to adopt and for what purpose.

Complementing the distributed, decentralized computing model discussed throughout this Section is the subfield known as secure multiparty computation (also known as privacy-preserving computation), which presents methods for multiple parties to jointly compute a function over all their respective inputs, while keeping those inputs private from other parties. These methods have matured in their origins from a theoretical curiosity to techniques with practical application in studies on tax and education records, cryptographic key management for the cloud, and more.[533] This makes secure multiparty computation methods a potential candidate for applications pertaining to secure, distributed computation.

Ultimately, it will be central for the NRC to continuously learn about the most effective security standards (including such other creative strategies as red teaming or bug bounties[534] to identify vulnerabilities) in this rapidly evolving space.

IX. INTELLECTUAL PROPERTY

Who should own the IP rights to outputs developed using NRC resources?[535] When private research is funded, subsidized, or influenced

---

[533] David Archer et al., *From Keys to Databases—Real-World Applications of Secure Multi-Party Computation*, 61 COMPUTER J. 1749 (2018).

[534] Amit Elazari Bar On, *We Need Bug Bounties for Bad Algorithms*, MOTHERBOARD (May 3, 2018), https://www.vice.com/en/article/8xkyj3/we-need-bug-bounties-for-bad-algorithms.

[535] Importantly, this Section discusses the extent to which researchers should be required to share their research outputs, *not* the extent to which researchers should be required to share their private data. The latter was discussed in Section 3.

by the federal government, the laws and rules have evolved, so that both the researcher and the government have certain rights in the intellectual property developed under the research. While IP protection is theoretically designed to incentivize research and innovation, some signs indicate that AI researchers in particular are already amenable to sharing the fruits of their research. Indeed, over two thousand researchers signed a 2018 petition to boycott a new machine intelligence journal started by *Nature*, because it promised to place its articles behind a paywall.[536] The Open Science and Open Research movements have also encouraged AI researchers to make their machine-learning software and algorithms publicly available.[537] Furthermore, as we discuss below, the advancement of techniques, like transfer learning, depends on researchers being able to distribute the fruits of their research freely.

This section surveys the existing IP-sharing agreements between researchers and the government, and explores whether and to what extent the government should retain IP rights over researchers' outputs, as a condition of using the NRC.[538] While the evidence on optimal IP rights varies, we recommend that: (1) academic researchers and universities should retain the same IP rights as the Bayh-Dole Act provides for patents developed under federally funded research; (2) the government should retain its copyrights and data rights under the Uniform Guidance, but contract around those rights where applicable to incentivize NRC usage and AI innovation; and (3) the government should consider conditions for requiring researchers to share their research outputs under an open-access license.

A. Patent Rights

A core question is whether NRC users should retain patent rights in inventions supported by the NRC. The Bayh-Dole Act regulates patent rights for inventions developed under federal funding agreements and its applicability depends on the nature of NRC access; for instance, if cloud credits are apportioned using federal grants, as described in Section 2,

---

[536] Dan Robitzski, *AI Researchers Are Boycotting A New Journal Because It's Not Open Access*, FUTURISM (May 3, 2018), https://futurism.com/artificial-intelligence-journal-boycot-open-access.

[537] *See generally* MIKIO L. BRAUN & CHENG SOON ONG, OPEN SCIENCE IN MACHINE LEARNING (2014).

[538] Since researchers using the NRC are not "contractors" under FAR/DFARS, and since evidence is lacking on the value of Other Transactions to AI researchers, we do not cover FAR/DFARS and Other Transactions in this section.

they may be considered federal funding agreements.[539] In such cases, Bayh-Dole Act permits researchers to hold the title to the patent and to license the patent rights.[540] However, these patent rights come with certain restrictions: For example, the funding agency has a free, nonexclusive license to use the invention "for or on behalf of the United States," and the agency may use "[m]arch-in rights" to grant additional licenses.[541]

The broader policy question about the government's exercise of its patent rights is whether and how patents stimulate innovation in AI. Some commentators have argued that the U.S. suffers from over-patenting in software,[542] and AI is no exception.[543] The total number of AI patent applications received annually by the U.S. Patent and Trademark Office more than doubled from thirty thousand in 2002 to over sixty thousand in 2018,[544] and some argue that this proliferation of broad AI patents, especially those filed by commercial companies, is hindering future innovation.[545] In the Bayh-Dole context, researchers have also found that the benefits of university patenting may justify the costs only where industry licensees need exclusivity to justify undertaking the costs of commercialization, as, for instance, in the

---

[539] Under the Bayh-Dole Act, Aa "federal funding agreement" is defined as "any contract, grant, or cooperative agreement entered into between any Federal agency, other than the Tennessee Valley Authority, and any contractor for the performance of experimental, developmental, or research work funded in whole or in part by the Federal Government." 35 U.S.C. § 201.

[540] 35 U.S.C. § 202.

[541] 35 U.S.C. § 203.

[542] *See, e.g.*, Mark A. Lemley & Julie E. Cohen, *Patent Scope and Innovation in the Software Industry*, 89 CAL. L. REV. 1 (2001); Mark A. Lemley, *Software Patents and the Return of Functional Claiming*, 2013 WIS. L. REV. 905 (2013).

[543] Jeremy Gillula & Daniel Nazer, *Stupid Patent of the Month: Will Patents Slow Artificial Intelligence?*, ELEC. FRONTIER FOUND. (Sept. 29, 2017), https://www.eff.org/deeplinks/2017/09/stupid-patent-month-will-patents-slow-artificial-intelligence.

[544] U.S. PATENT & TRADEMARK OFF., INVENTING AI: TRACING THE DIFFUSION OF ARTIFICIAL INTELLIGENCE WITH PATENTS 2 (2020).

[545] *See, e.g.*, Mike James, *Google Files AI Patents*, I PROGRAMMER (July 8, 2015), https://www.i-programmer.info/news/105-artificial-intelligence/8765-google-files-ai-patents.html. This is especially problematic because companies represent twenty-six out of the top thirty AI patent applicants worldwide, while only four are universities or public research organizations. WORLD INTELL. PROP. ORG., ARTIFICIAL INTELLIGENCE 7 (2019).

pharmaceutical context.[546] For the substantial portion of university patenting, including AI, this rationale may not carry much weight.[547]

Some research shows that patents actually may not actually have any net effect on the amount or quality of AI research conducted in the university context. In an empirical study of faculty at the top computer science and electrical engineering universities in the United States, research has found that the prospect of obtaining patent rights to the fruits of their research does not motivate researchers to conduct more or higher-quality research.[548] Eighty-five percent of professors reported that patent rights were not among the top four factors motivating their research activities, and 57 percent of professors reported that they did not know whether or how their university shares licensing revenue with inventors.[549] The patent scheme adopted by the NRC, therefore, may not have a strong influence on researcher adoption.

That said, as a practical matter, there is a virtue to treating innovations stemming from NRC usage in a fashion that is consistent with Bayh-Dole. Particularly if cloud credits are awarded through the expansion of programs like NSF CloudBank, it would be confusing to have distinct patent rights out of the research and cloud grant. In addition, many university tech transfer offices appear to have strong preferences for patent rights.[550] To the extent that universities view retaining patent rights as a condition for using the NRC, aligning NRC patent rights with Bayh-Dole may be preferred, but the evidence underpinning this recommendation is not strong.

---

[546] Lisa Ouellette & Rebecca Weires, *University Patenting: Is Private Law Serving Public Values?*, 2019 MICH. ST. L. REV. 1329 (2019).
[547] *Id.* at 1331; *see also* Arti Kaur Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 NW. U. L. REV. 77, 136 (1999).
[548] *See* Brian J. Love, *Do University Patents Pay Off? Evidence From a Survey of University Inventors in Computer Science and Electrical Engineering*, 16 YALE J. L. & TECH. 285 (2014).
[549] *See id.* at 286.
[550] *See, e.g.*, *Tech Transfer FAQ*, U. MICH., https://techtransfer.umich.edu/for-inventors/resources/inventor-faq/ (last visited Feb. 19, 2022) ("We carefully review the commercial potential for an invention before investing in the patent process. However, because the need for commencing a patent filing usually precedes finding a licensee, we look for creative and cost-effective ways to seek early protections for as many promising inventions as possible"); *What is Technology Transfer*, PRINCETON U., https://patents.princeton.edu/about-us/what-technology-transfer (last visited Apr. 10, 2022) ("[T]echnologies and everyday products are possible because of technology transfer . . . Because the discoveries emerging from university research tend to be early-stage, high-risk inventions, successful university technology transfer transactions require a patent system that protects such innovations.").

B. Copyright, Data Rights, and the Uniform Guidance

The Uniform Guidance streamlines and consolidates government requirements for receiving and using federal awards to reduce administrative burden.[551] Grants.gov describes it as a "government-wide framework for grants management," a groundwork of rules for federal agencies in administering federal funding.[552] The Uniform Guidance includes provisions on, for instance, cost principles, audit requirements, and requirements for the contents of federal awards.[553]

The Uniform Guidance is applicable to "federal awards,"[554] but IP provisions do not *require* the government to assert its rights over researcher outputs.[555] Whether and how the government allocates its IP rights under the Uniform Guidance is therefore an important question.

This section first covers government copyright and data rights to IP under the Uniform Guidance and discusses how sharing copyright and data rights might impact the AI innovation landscape. We then examine the extent to which the government should retain its rights to research generated using the NRC. While the evidence is mixed, we ultimately recommend that the government retain its copyrights and data rights under the Uniform Guidance, but contract around those rights where applicable, to incentivize NRC usage and further AI innovation.

---

[551] The Uniform Guidance for intellectual property is laid out in 2 C.F.R. § 200.315.
[552] *Uniform Administrative Requirements, Cost Principles, and Audit Requirements for Federal Awards*, GRANTS.GOV, https://www.grants.gov/learn-grants/grant-policies/omb-uniform-guidance-2014.html (last visited Aug. 27, 2021).
[553] *See Key Sections of the Uniform Guidance*, ASS'N INT'L CERTIFIED PROFESSIONAL ACCOUNTANTS, https://www.aicpa.org/interestareas/governmentalauditquality/resources/singleaudit/uniformguidanceforfederalrewards/key-sections-uniform-guidance.html (last visited Mar. 20, 2022).
[554] 2 C.F.R. § 200.315. A "federal award" under the Uniform Guidance includes, among other things, "the federal financial assistance that a recipient receives directly from a Federal awarding agency or indirectly from a pass-through entity;" or "the cost-reimbursement contract under the Federal Acquisition Regulations;" or a "grant agreement, cooperative agreement, [or] other agreement [for federal financial assistance]." 2 C.F.R. § 200.1.
[555] 2 C.F.R. § 200.315(b), (c) (These provisions specify that the government merely "reserves" its "right" to copyright and data rights over research produced under the federal award).

1. Copyright

Under U.S. copyright law, NRC researchers can obtain copyrights over various aspects of their work. For instance, NRC researchers may wish to copyright the software they used to build the model since software is considered a literary work under the Copyright Act.[556] Researchers may even obtain copyrights over various aspects of the model, including the choices of training parameters, model architectures, and training labels, if they can show that those choices required creativity.[557] Many scholars have even opined, without reaching consensus, on whether outputs such as text and art that are artificially generated can be copyrighted.[558]

Under the Uniform Guidance,[559] the recipient of federal funds may copyright any work that was developed or acquired under a federal award. However, even if researchers are permitted to maintain copyrights, the federal awarding agency reserves a "royalty-free, nonexclusive and irrevocable right to reproduce, publish, or otherwise use the work for federal purposes, and to authorize others to do so."[560] Notably, this right is limited to "federal purposes," meaning that third parties who acquire licenses to the researchers' copyrighted works cannot use them for exclusively commercial purposes.[561]

It is unclear to what extent copyrights over NRC outputs should be fully vested in the researcher to stimulate basic AI research. One class of AI research and development output that has received significant academic attention has been whether AI-generated creative works, like music from OpenAI's Jukebox,[562] can or should receive copyright

---

[556] U.S. COPYRIGHT OFFICE, COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES 35 (2021, 3d ed.).
[557] Wil Michiels, *How Do You Protect Your Machine Learning Investment?*, EETIMES (Mar. 31, 2020), https://www.eetimes.com/how-do-you-protect-your-machine-learning-investment-part-ii/.
[558] *See, e.g.*, Tabrez Y. Ebrahim, *Data-Centric Technologies: Patent and Copyright Doctrinal Disruptions*, 43 NOVA L. REV. 287, 304; Daryl Lim, *AI & IP: Innovation & Creativity in an Age of Accelerated Change*, 52 AKRON L. REV. 813, 835 (2018).
[559] 2 C.F.R. § 200.315(b).
[560] *Id.*
[561] For a comprehensive report on how artificial intelligence is used in various government agencies, see DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020).
[562] *Jukebox*, OPENAI (Apr. 30, 2020), https://openai.com/blog/jukebox/.

protection.[563] However, the technology and copyright community has hardly reached a consensus on whether the public interest in AI research requires granting copyright in these scenarios. On one hand, in a survey of AI scientists, tech policy experts, and copyright scholars, roughly 54 percent of respondents agreed that copyright protection is an important incentive for authors to make their work commercially available, and 63 percent agreed that an increase in the number of commercially available AI-produced works would stimulate further AI growth and research.[564] On the other hand, in the same survey, approximately 56 percent of respondents agreed that the U.S. Copyright Office should *deny* copyright protection to creative works produced independently by AI without creative intervention from a human author.[565]

Notwithstanding the prominent debate about copyright over creative works generated by AI models, such works are only a subset of possible copyright protection in the AI context. As discussed above, researchers could theoretically seek additional copyright protection over, among other things, their code, architecture, or model. Here, AI innovation may depend on sharing these copyrightable elements. For instance, transfer learning uses existing ML models and "fine-tunes" those models for a related target task,[566] and various fine-tuning approaches have emerged to perform transfer learning on different classes of tasks.[567]

2. Data Rights

Under the Uniform Guidance, when "data" is "produced" under a federal award, the government reserves the right to: (1) obtain, reproduce, publish or otherwise use such data; and (2) authorize others to receive, reproduce, publish or otherwise use such data.[568]

---

[563] *See, e.g.*, Shlomit Yanisky-Ravid, *Generating Rembrandt: Artificial Intelligence, Copyright, and Accountability in the 3A Era—the Human-Like Authors are Already Here—a New Model*, 27 MICH. ST. L. REV. 659 (2017); Kalin Hristov, *Artificial Intelligence and the Copyright Dilemma*, 57 J. FRANKLIN PIERCE CTR. INTELL. PROP. 431 (2017).
[564] Kalin Hristov, *Artificial Intelligence and the Copyright Survey*, 16 J. SCI. POL'Y & GOVERNANCE 1, 14-15 (2020).
[565] *Id.* at 16.
[566] *See What is Transfer Learning?*, TENSORFLOW (Mar. 31, 2020), https://www.tensorflow.org/js/tutorials/transfer/what_is_transfer_learning.
[567] *See, e.g.*, Yunhui Guo et al., *SpotTune: Transfer Learning Through Adaptive Fine-Tuning*, ARXIV (Nov. 2018), https://arxiv.org/pdf/1811.08737.pdf.
[568] 2 C.F.R. § 200.315(d).

Notably, this does *not* limit the use of such data for federal government purposes. In other words, such data can be promulgated for *any* use. The outstanding question, therefore, is whether this "data," which is not explicitly defined in the Uniform Guidance, covers data generated for AI and machine-learning purposes. Below, we examine two classes of data generated for AI purposes—synthetic data and data labels—and how sharing this data could impact AI innovation.

One class of data generated for AI purposes is synthetic data. Researchers have turned to deep generative models such as Variational Autoencoders[569] and Generative Adversarial Networks[570] to generate synthetic data to train their machine learning models. As noted by the World Intellectual Property Organization, synthetic data is an entirely new class of data that does not fit neatly under existing IP law.[571] While a researcher may seek copyright protection over the subset of synthetic data that is "creative," therefore implicating the copyright provisions of the Uniform Guidance (described above), the broad class of synthetic data, whether "creative" or not, may also implicate the data rights provision. On the one hand, training data is often carefully guarded,[572] so requirements to share synthetic data, which is often used to train AI models, may be a non-starter for NRC users. On the other hand, many scholars have written about the promise of synthetic data to actually *enable* further data sharing by preserving privacy and researchers' trade secrets.[573] In fact, sharing synthetic datasets would spur additional research and innovation in fields such as healthcare, where data sharing has been limited.[574]

Another class of data generated for AI is labeled data, namely data that has been tagged and classified to provide ground truth for

---

[569] *See* Zhiqiang Wan, Yazhou Zhang & Haibo He, *Variational Autoencoder Based Synthetic Data Generation for Imbalanced Learning*, 2017 IEEE SYMP. SERIES ON COMPUTATIONAL INTEL. 1 (2017).
[570] *See* Noseong Park, Mahmoud Mohammadi & Kshitij Gorde, *Data Synthesis Based on Generative Adversarial Networks*, 11 PROC. VLDB ENDOWMENT 1071 (2018).
[571] *See* RON BAKKER, IMPACT OF ARTIFICIAL INTELLIGENCE ON IP POLICY 12 (2020).
[572] *See* MARTA DUQUE LIZARRALDE, A GUIDELINE TO ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND INTELLECTUAL PROPERTY 4-7 (2020).
[573] Steven M. Bellovin et al., *Privacy and Synthetic Datasets*, 22 STAN. TECH. L. REV. 1, 2-3 (2019); *see also* Fida K. Dankar & Mahmoud Ibrahim, *Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation*, 5 APPLIED SCI. 11 (2021); *but see* Theresa Stadler et al., *Synthetic Data – Anonymisation Groundhog Day*, ARXIV (July 8, 2021), https://arxiv.org/pdf/2011.07018.pdf.
[574] *See, e.g.*, Daniel S. Quintana, *A Synthetic Dataset Primer for the Biobehavioural Sciences to Promote Reproducibility and Hypothesis Generation*, 9 ELIFE 1 (2020).

supervised machine learning models.[575] While techniques have been developed to decrease the costs associated with data labeling,[576] it nevertheless remains a resource and time-intensive task. For example, Cognilytics Research reports that 25 percent of the total time spent building machine learning models is devoted to data labeling.[577] Researchers using the NRC may therefore seek to protect their investment in data labeling by opting not to share their labels with others, especially if the underlying data is proprietary.[578] However, recognizing the difficulty of data labeling, some researchers have built online platforms for sharing data labels.[579] In the case of ImageTagger, a data labeling and sharing platform for RoboCup Soccer, the developers wanted to solve the problem that no single team, acting alone, could easily build its own high-quality training sets.[580] Similarly, in the NRC's case, the sharing of labeled government data—where labeling may have been augmented by NRC resources[581]—could act as a rising tide that lifts all boats, improving the quality of not only the government data as a training dataset, but also all subsequent research using that data. Furthermore, sharing data labels could be instrumental in conducting

---

[575] Yuji Roh et al., *A Survey on Data Collection for Machine Learning*, ARXIV (Aug. 12, 2019), https://arxiv.org/pdf/1811.03402.pdf.

[576] *See, e.g.*, Hang Qiu et al., *Minimum Cost Active Labeling*, ARXIV (June 24, 2020), https://arxiv.org/pdf/2006.13999.pdf; Eric Horvitz, *Machine Learning, Reasoning, and Intelligence in Daily Life: Directions and Challenges*, 18 PROC. CONF. ON UNCERTAINTY A.I. 3 (2007).

[577] COGNILYTICS RESEARCH, DATA ENGINEERING, PREPARATION, AND LABELING FOR AI 2019 3 (2019).

[578] *See* Wil Michiels, *How Do You Protect Your Machine Learning Investment?*, EETIMES (Mar. 26, 2020), https://www.eetimes.com/how-do-you-protect-your-machine-learning-investment/. In fact, in the European Union, labeled datasets are awarded with database rights protections. Mauritz Kop, *Machine Learning & EU Data Sharing Practices*, STAN.-VIENNA TRANSATLANTIC TECH. L. F. (Mar. 24, 2020), https://ttlfnews.wordpress.com/2020/03/24/machine-learning-eu-data-sharing-practices/.

[579] *See, e.g.*, Niklas Fiedler et al., *ImageTagger: An Open Source Online Platform for Collaborative Image Labeling*, *in* ROBOCUP 2018: 11374 LECTURE NOTES ON A.I. 162 (Dirk Holz et al. eds., 2019).

[580] *Id.* at 162.

[581] Researchers may, for instance, use NRC data and compute resources to implement active learning strategies, procedures to manually label a subset of available data and infer the remaining labels automatically using a machine learning model. *See, e.g.*, Oscar Reyes et al., *Effective Active Learning Strategy for Multi-Label Learning*, 273 NEUROCOMPUTING 494 (2018). Similarly, researchers may augment existing public sector data with valuable labels.

bias and fairness of NRC research outputs where necessary, as discussed in Section 7.[582]

3. Retaining IP Rights in the Uniform Guidance

As the preceding discussion suggests, sharing AI research output covered by copyrights and data rights could be beneficial to AI innovation. We, therefore, recommend that the NRC at least retain the same rights to copyrights and data rights as under the Uniform Guidance, yielding several additional benefits. First, similar to our recommendation in Section 3 that federal agencies should be allowed to use the NRC's compute resources, retaining the same Uniform Guidance IP allocation scheme could produce welfare benefits by improving government decision-making using AI. For instance, federal agencies can reduce the cost of core governance functions and increase agency efficiency and effectiveness by using data labels shared by NRC researchers or by fine-tuning models generated by NRC researchers. Second, retaining the Uniform Guidance IP allocation scheme would result in more consistency across the federal award landscape. Indeed, as mentioned above in the patent context, it could be confusing to diverge from the Uniform Guidance, especially if the cloud credit grant is apportioned through programs like CloudBank but the research grant is administered as a federal award.

In sum, we recommend that the government at least retain its copyrights and data rights under the Uniform Guidance. However, we also reiterate that the Uniform Guidance serves merely as a helpful framework, not as an immutable rule. Where the Uniform Guidance IP allocation would dissuade researchers from using the NRC or hinder AI innovation in specific scenarios, the government can and should explicitly modify its rights and contract separately with researchers on what rights the government retains, if any.

C. Considerations for Open-Sourcing

Should the government go *beyond* its rights and mandate that researchers share their NRC research outputs with others under an open-

---

[582] *See, e.g.*, Pedro Saleiro et al., *Aequitas: A Bias and Fairness Audit Toolkit*, ARXIV (Apr. 29, 2019), https://arxiv.org/pdf/1811.05577.pdf; Florian Tramèr et al., *FairTest: Discovering Unwarranted Associations in Data-Driven Applications*, ARXIV (Aug. 16, 2019), https://arxiv.org/pdf/1510.02377.pdf.

source license? As an initial matter, we note that agencies can modify the IP allocation schemes under the Uniform Guidance,[583] but not under the Bayh-Dole Act.[584] Some federal agencies supplement and/or replace the IP rights set out in the Uniform Guidance with restrictions that are more specific to the IP being developed for that particular agency or under a specific award.[585] For instance, the Department of Labor requires that intellectual property developed under a federal award must not only comply with the terms specified in the Uniform Guidance, but also be available for open licensing to the public.[586] NSF grantees are also expected to share their data with others.[587] However, the government cannot change the allocation of patent ownership under the Bayh-Dole Act, unless the Act itself is modified or unless the NRC isn't administered as a federal award, rendering the Act inapplicable.

Requiring researchers to open-source their research outputs may be possible, but the considerations around it are complex. On the one hand, an open-source requirement could negatively affect downstream commercialization, given the wide range of potential AI research.[588] While the NRC might protect commercialization to some degree by adopting a restrictive open-source license,[589] the mere divergence from the Uniform Guidance or the Bayh-Dole Act could be confusing for researchers in navigating federal awards and understanding open-source

---

[583] *See* 2 C.F.R. § 200.101(b); 2 C.F.R. §§ 200.315 (a), (c)

[584] *See* 35 U.S.C. §§ 202, 203.

[585] While we do not discuss the idiosyncratic modifications to the Uniform Guidance that vary from agency-to-agency, we encourage the task force to assess these modifications if it decides to implement the NRC through a particular agency. If the NRC is administered through multiple agencies, the complex amalgam of agency-specific IP rules may increase the friction in using the NRC if researchers must context-switch from one set of regulations to the next depending on the funding agency.

[586] 2 C.F.R. § 2900.13. Previously, the Department of Labor explicitly required IP generated under a federal award to be licensed under a Creative Commons Attribution license, but this rule was changed in April 2021 to replace the proprietary term "Creative Commons Attribution license" with the industry-recognized standard "open license." 86 Fed. Reg. 22107 (Apr. 27, 2021).

[587] *Dissemination and Sharing of Research Results - NSF Data Management Plan Requirements*, NAT'L SCI. FOUND., https://www.nsf.gov/bfa/dias/policy/dmp.jsp (last visited Mar. 21, 2022).

[588] *See, e.g.*, Aidan Courtney et al., *Balancing Open Source Stem Cell Science with Commercialization*, NATURE BIOTECHNOLOGY (Feb. 7, 2011), https://www.nature.com/articles/nbt.1773.

[589] *See* Klint Finley, *When Open Source Software Comes with a Few Catches*, WIRED (July 31, 2019), https://www.wired.com/story/when-open-source-software-comes-with-catches/; *Guide to Open Source Licenses*, SYNOPSYS (Oct. 7, 2016), https://www.synopsys.com/blogs/software-security/open-source-licenses/.

licensing interactions across multiple situations.[590] Furthermore, requiring researchers to share research outputs comes with its own host of privacy and cybersecurity issues.[591] If researchers are permitted to use the NRC to conduct classified research,[592] for instance, then keeping research outputs proprietary would serve the national interest.[593] In this case, however, the NRC should consider limiting any open-source requirement to research that has fewer privacy and security implications.

On the other hand, as discussed, sharing research outputs with other NRC researchers could be beneficial, and many scholars argue that AI researchers should open-source their software to stimulate innovation.[594] A requirement to open-source software code, which can be the subject of both copyrights and patent rights,[595] may contravene Bayh-Dole and face challenges from universities that seek to retain their patent rights, but software patent disclosures alone are often limited and over-broad, and fail to enhance social welfare.[596] Requiring fuller disclosure of code generated on the NRC can therefore decrease the risk of over-patenting and increase AI innovation. The growth of the robust open-source and open science movements also suggests that an open-

---

[590] *See* Daniel A. Almeida et. al, *Do Software Developers Understand Open Source Licenses?*, 25 IEEE INT'L CONF. ON PROGRAM COMPREHENSION 1 (2017) (finding that software developers "struggle[] when multiple [open-source] licenses [are] involved" and "lack the knowledge and understanding to tease apart license interactions across multiple situations.").

[591] *See, e.g.,* ALEXANDRA THEBEN ET AL., CHALLENGES AND LIMITS OF AN OPEN SOURCE APPROACH TO ARTIFICIAL INTELLIGENCE 14 (2021); Stadler et al., *supra* note 573; Milad Nasr et al., *Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks Against Centralized and Federated Learning*, ARXIV (June 6, 2020), https://arxiv.org/abs/1812.00910.pdf.

[592] Some universities have decided to eliminate classified research. *See, e.g.*, *At the Hands of Radicals*, STAN. MAG. (Jan. 2009), https://stanfordmag.org/contents/at-the-hands-of-the-radicals.

[593] *See* Donald Kennedy, *Science and Secrecy*, 289 SCI. 724 (2000); Peter J. Westwick, *Secret Science: A Classified Community in the National Laboratories*, 38 MINERVA 363 (2000).

[594] *See* BRAUN & ONG, *supra* note 537; Sören Sonnenburg et al., *The Need for Open Source Software in Machine Learning*, 8 J. MACH. LEARNING RES. 2443 (2007); *see also* Katie Malone & Richard Wolski, *Doing Data Science on the Shoulders of Giants: The Value of Open Source Software for the Data Science Community*, HDSR (May 31, 2020), https://hdsr.mitpress.mit.edu/pub/xsrt4zs2/release/4.

[595] *See* Laura A. Heymann, *Overlapping Intellectual Property Doctrines: Election of Rights Versus Selection of Remedies*, 17 STAN. TECH. L. REV. 239, 240 (2013); *Oracle Am. Inc. v. Google Inc.*, 750 F.3d 1339 (Fed. Cir. 2014) (accepting that software is both patentable and copyrightable).

[596] Robert E. Thomas, *Debugging Software Patents: Increasing Innovation and Reducing Uncertainty in the Judicial Reform of Software Patent Law*, 25 SANTA CLARA COMP. & HIGH TECH. L.J. 191, 222-23 (2008).

sourcing requirement for the NRC would not be a complete barrier to NRC usage.[597]

A strong argument for mandating open-sourcing also comes from the increasing private-sector reliance on trade secrets for IP protection in AI.[598] Some argue that this heightened emphasis on trade secret protection constitutes "artificial stupidity,"[599] as it has stifled innovation in AI by preventing disclosure, providing protection for a potentially unlimited duration, and attaching immediately and broadly to any output with perceivable economic value.[600] The reliance on secrecy, therefore, contravenes many of the principles described above—which argue that sharing code and data is crucial in AI—and results in significant AI industry consolidation and suboptimal levels of AI innovation.[601] This harkens back to the goal of the NRC discussed in Section 1: addressing problems with AI research being concentrated in the hands of a few private-sector players. Because the NRC should explicitly avoid replicating these private-sector challenges, this lends additional support to a recommendation that the NRC should contemplate requiring researchers to share their research outputs.

In sum, while AI raises a host of novel IP issues (e.g., whether AI output is itself eligible for IP protection), we think the government can steer clear of many of these complications by tracking Bayh-Dole and the Uniform Guidance. The government should also consider conditions for

---

[597] *See, e.g.*, Joaquin Vanschorin et al., *OpenML: Networked Science in Machine Learning*, ARXIV (Aug. 1, 2014), https://arxiv.org/pdf/1407.7722.pdf (developing a collaboration platform through which scientists can automatically share, organize and discuss machine learning experiments, data, and algorithms); *see also* Sarah O'Meara, *AI Researchers in China Want to Keep the Global-Sharing Culture Alive*, NATURE (May 29, 2019), https://www.nature.com/articles/d41586-019-01681-x; Shuai Zhao et al., *Packaging and Sharing Machine Learning Models via the Acumos AI Open Platform*, 17 IEEE INT'L CONF. ON MACH. LEARNING & APPLICATIONS 841 (2018).

[598] Jeanne C. Fromer, *Machines as the New Oompa-Loompas: Trade Secrecy, the Cloud, Machine Learning, and Automation*, 94 N.Y.U. L. REV. 706, 712 (2019); JORDAN R. RAFFE ET AL., THE RISING IMPORTANCE OF TRADE SECRET PROTECTION FOR AI-RELATED INTELLECTUAL PROPERTY 1, 5-6 (2020); Jessica M. Meyers, *Artificial Intelligence and Trade Secrets*, AM. BAR ASS'N (Feb. 2019), https://www.americanbar.org/groups/intellectual_property_law/publications/landslide/2018-19/january-february/artificial-intelligence-trade-secrets-webinar/; *AIPLA Comments Regarding "Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation"*, AM. INTELL. PROP. L. ASS'N (Jan. 10, 2020), https://www.uspto.gov/sites/default/files/documents/AIPLA_RFC-84-FR-58141.pdf.

[599] Clark D. Asay, *Artificial Stupidity*, 61 WM. & MARY L. REV. 1187, 1197, 1241-42 (2020).

[600] *See id.*; AM. INTELL. PROP. L. ASS'N, *supra* note 596, at 16.

[601] *See* Asay, *supra* note 597, at 1242.

requiring NRC researchers to disclose or share their research outputs under an open-access license.

CONCLUSION

As we have articulated in this Article, the ambitious call for an NRC has transformative potential for the AI research landscape.

Its biggest promise is to ensure more equitable access to core ingredients for AI research: compute and data. Leveling this playing field could shift the current ecosystem from one that focuses on narrow commercial problems to one that fosters basic, noncommercial AI research to ensure long-term national competitiveness, to solve some of the most pressing problems, and to rigorously interrogate AI models.

As we have spelled out in this Article, the NRC does raise a host of policy, legal, and normative questions. How can such compute resources be provided in a way that is expeditious and user-friendly, but does not preclude the potential cost savings from a publicly owned resource? How can the NRC be designed to adhere to the Privacy Act of 1974, which was animated by concerns about a national system of records that surveils its citizens? How can we ensure that NRC mitigates, rather than heightens, concerns about the unethical use of AI? And how can one prevent the NRC from becoming the biggest target for cyberattacks?

These are tough questions, and we hope to have sketched out our initial attempt at answers above. We are hopeful, if designed well, the NRC could help to realign the AI innovation space from one that is fixated on short-term private profit to one that is infused with long-term public values.

APPENDIX

I. COMPUTING INFRASTRUCTURE COST COMPARISONS

This Appendix provides a sample cost-estimate comparison between a commercial cloud service, AWS, and a dedicated government HPC system, Summit. In sum, our estimations show that AWS P3 instances with comparable hardware to Summit would be 7.5 times as expensive as estimated costs under constant usage, and 2.8 times Summit's estimated costs under fluctuating demand.

Table 3 lists the three infrastructure models used in this comparison. Summit was used as the reference government HPC system because it is one of the DOE's newest systems and has hardware well-suited for AI research.[602] The other infrastructure model used is AWS EC2 P3.[603] Both are commonly used in AI research and general HPC applications. Other commercial cloud platforms, such as GCP or Azure, could also feasibly provide the infrastructure for the NRC. AWS EC2 P3 was used here because AWS has a robust cost calculator that allows for variable workloads.

The number of AWS instances was set such that those models would have the exact same number of GPUs as Summit. GPUs were the fixed variable because GPUs are the most important hardware for AI research applications, specifically deep learning. Both Summit and AWS P3 instances use NVIDIA V100 GPUs.

We conduct our cost comparison for the two infrastructure models over five years, as Summit's initial RFP documents include a five-year maintenance contract. AWS, however, only provides one-year or three-year pricing plans, so we extrapolated the five-year cost based on its three-year plan.

For the cost estimate of Summit, we based our calculation on the budget details in the original Department of Energy (DOE) Request for Proposal (RFP) in January 2014.[604] The RFP includes a $155 million

---

[602] *Department of Energy Awards $425 Million for Next Generation Supercomputing Technologies*, ENERGY.GOV (Nov. 14, 2014), https://www.energy.gov/articles/department-energy-awards-425-million-next-generation-supercomputing-technologies.

[603] *Amazon EC2 P3 Instances*, AMAZON, https://aws.amazon.com/ec2/instance-types/p3/ (last visited Sept. 9, 2021).

[604] *CORAL Request for Proposal B604142*, LAWRENCE LIVERMORE NAT'L LAB'Y (2014), https://web.archive.org/web/20140816181824/https://asc.llnl.gov/CORAL/. We note that we were not able to locate the final award documents, nor is Summit

maximum budget for building Summit, an expected $15 million maximum for the non-recurring engineering cost,[605] and around $15 million for five-year maintenance,[606] plus interest based on the U.S. Treasury securities at five-year constant maturity as specified in the price schedule.[607] Upon calculation, we estimated Summit costs around $192 million in total, which is consistent with public reporting of the cost of Summit.[608]

For the cost estimate of AWS, we used the AWS pricing calculator, choosing U.S. East (N. Virginia) as the data center and publicly available rates under the cheapest possible pricing plan (EC2 Instance Savings Plans). To approximate a negotiated discount, we applied a 10 percent discount based on the negotiated rate of one major university.

Since commercial cloud platform costs scale with how many instances are actually in use, two costs were calculated for each AWS model representing usage extremes: (1) with the infrastructure under constant usage; (2) with the infrastructure under dramatically fluctuating usage each day. For the daily spike traffic calculation, we set the model to run five days a week with 8.4 hours each day at peak performance. The maximum number of instances used is the same as

---

budgeted in sufficient detail to back out cost from the DOE budget statements. Our cost estimates here, however, are comparable to publicly reported estimates for the total cost of the Summit system.

[605] This is based on a $30 million maximum in the DOE Office of Science contract for non-recurring engineering (NRE) costs for the systems at Argonne National Laboratory and Oak Ridge National Laboratory.

[606] This is based on the difference in the RFP terms between the inclusion of maintenance under the Lawrence Livermore National Laboratory system (with a maximum budget of $170 million) and the exclusion of maintenance under the systems for the Oak Ridge National Laboratory and the Argonne National Laboratory (with a maximum budget for the build contract of $155 million). This is likely an upper bound on maintenance, given that the difference reflects the combination of NRE and five-year maintenance.

[607] *See CORAL Price Schedule*, LAWRENCE LIVERMORE NAT'L LAB'Y (2014), https://web.archive.org/web/20140816181824/https://asc.llnl.gov/CORAL/RFP_components/04_CORAL_Price_Schedule_ANL_ORNL_tabs.xlsx. We used 1.62 percent as the interest rate to calculate the cost over sixty months. It is the five-year Treasury constant maturity rate on November 14, 2014, *see Selected Interest Rates (Daily) – H.15*, FED. RES., https://www.federalreserve.gov/releases/H15/default.htm, when DOE announced the award of the HPC system, *see* ENERGY.GOV*, supra* note 600.

[608] For instance, this estimate is in line with the cost of $200 million reported by the *New York Times*. Steve Lohr, *Move Over, China: U.S. is Again Home to World's Speediest Supercomputer*, N.Y. TIMES (June 8, 2018), https://www.nytimes.com/2018/06/08/technology/supercomputer-china-us.html. Some reporting conflates the procurement of multiple systems that occurred contemporaneously.

one would use for constant use while the minimum number is zero. This workload setting is based on the assumption that GPUs used for training AI models sit idle 30 percent of the time.[609] These estimates should provide hard upper and lower bounds on costs for using each instance type.

Figure 1 plots cost on the *y*-axis over a five-year period on the *x*-axis. The turquoise line indicates the cost of a Summit-like system and the purple and blue lines indicate the cost of the same AWS instances under variable and constant usage. Overall, this simple analysis corroborates the analysis conducted by Compute Canada, which found that commercial cloud "ranged from 4x to 10x more than the cost of owning and operating our own clusters."[610] Over five years and under constant usage, AWS P3 instances with comparable hardware to Summit would be 7.5 times as expensive as estimated costs. Under fluctuating demand, AWS P3 instances would cost 2.8 times Summit's estimated costs.

We note that this simple analysis omits many potential factors (see discussion in Section 2), but provides a starting point to understanding the considerable cost implications for the make-or-buy decision.

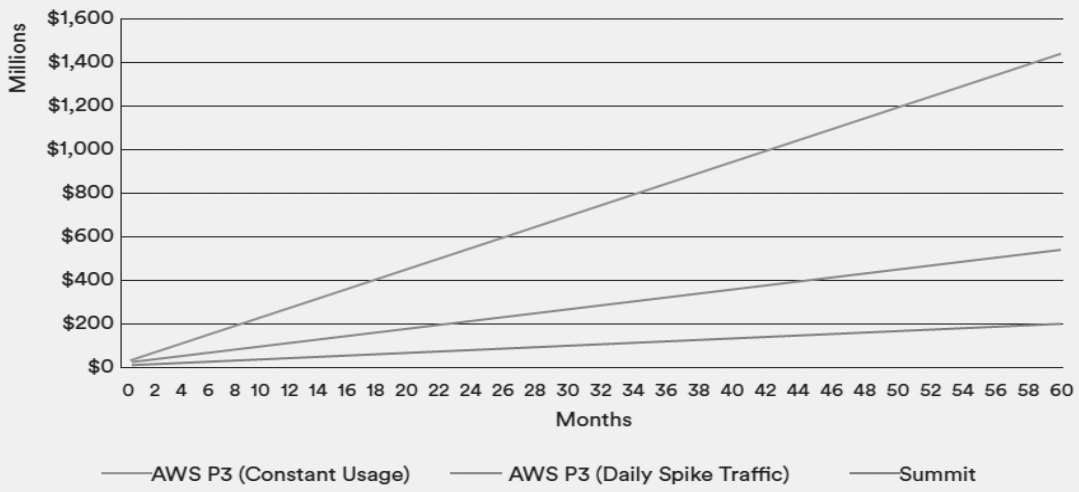
---

[609] Research shows that for training compute-intensive deep learning models, such as ResNet-101, the GPU utilization is around 70 percent. Jingoo Han et al., *A Quantitative Study of Deep Learning Training on Heterogeneous Supercomputers*, 2019 IEEE CONF. ON CLUSTER COMPUTING 1, 5 (2019). However, ResNet-50 has a GPU utilization of approximately 40 percent, *see id.*, and other accounts report that GPUs are utilized only 15-30 percent of the time, *see, e.g.*, Lukas Biewald, *Monitor and Improve GPU Usage for Training Deep Learning Models*, TOWARDS DATA SCI. (Mar. 27, 2019), https://towardsdatascience.com/measuring-actual-gpu-usage-for-deep-learning-training-e2bf3654bcfd; Janet Morss, *Giving Your Data Scientists a Boost with GPUaaS*, CIO (June 2, 2020), https://www.cio.com/article/3561090/giving-your-data-scientists-a-boost-with-gpuaas.html.

[610] COMPUTE CAN., CLOUD COMPUTING FOR RESEARCHERS 1 (2016), https://www.computecanada.ca/wp-content/uploads/2015/02/CloudStrategy2016-2019-forresearchersEXTERNAL-1.pdf.

Table 3: Summit & AWS Comparison			
	GPUs	**RAM**	**Network Bandwidth**
Summit IBM AC922	27,648 (NVIDIA Volta V100)	2.8 PB	200 Gb/s
AWS P3dn.24xlarge (3456 nodes)	27,648 (NVIDIA Volta V100)	2.6 PB	100 /s

FIGURE 1 – ESTIMATED COST OF AWS INSTANCES COMPARED TO SUMMIT OVER 3 YEARS



II. FACILITATING PRIVATE DATASET SHARING

Unique IP challenges arise if researchers are permitted to share their own private datasets with the NRC. Indeed, researchers who "upload" proprietary data may be concerned about how other NRC users

utilize that data.[611] Through interviews conducted for this Article, corporate stakeholders representing the entertainment industry, as well as other creative industries, have further expressed fear that researchers may upload and share data to which they do not hold rights. However, if the NRC does decide to facilitate private data-sharing, it should consider adopting two requirements to address these concerns: (1) the NRC should require all users to affirm they either have the original IP rights to the data or the data is already in the public domain; and (2) the NRC should have a scheme for its users to license their data.

A. NRC users must own IP rights to the data they are uploading

Researchers uploading data need to agree that they own the intellectual property rights to the data prior to upload, or that the data is already in the public domain. This should be the case whether researchers share the data broadly with other researchers or simply use their data for their own private use.

Of course, despite mandating that uploaders guarantee legitimate ownership or public domain status of their uploaded IP, uploaders may nevertheless upload data they don't own the IP rights to. This may happen because computer engineers and researchers are not informed about IP law, anticipate that fair use will excuse their behavior, or simply hope not to get caught.[612] Industry stakeholders were also concerned that AI researchers would pull out "facts" from a copyrighted work (e.g., certain melodies in the chorus of a song) or apply certain algorithms to the work and "wrongly" claim a copyright over the transformed work. Whatever the case may be, this assembly of protected input data represents the "clearest copyright liability in the machine learning process" because assembling protected data violates the right to reproduction, and any preprocessing of the data could violate the right to derivative works.[613]

In interviews, corporate stakeholders expressed a desire to stymie the upload of copyrighted works by having the NRC itself assess whether uploaded data is already protected by copyright. Data can be reviewed

---

[611] Jennifer Shkabatur, *The Global Commons of Data*, 22 STAN. TECH. L.R. 407, 407-09 (2019).
[612] Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 61 (2017).
[613] *Id.*

manually, or by using such automated systems as *Content ID*, which is also used by corporations such as YouTube.[614] The former option would be very labor-intensive,[615] whereas the latter may be prohibitively expensive,[616] so the value of addressing these concerns must be weighed against these burdensome costs.

Finally, it is unclear the extent to which uploading and sharing copyrighted data for machine learning amounts to fair use.[617] The most analogous case is *Author's Guild v. Google Books*.[618] In that case, Google scanned over 20 million books, many of which were copyright-protected, and assembled a corpus of machine-readable texts to power its Google Books service.[619] The Second Circuit held that Google Books' unauthorized reproductions of copyrighted works was transformative fair use, largely because Google Books provided information *about* books through small snippets, without threatening the rights-holders' core protectable expression in the books.[620] While some have opined that the *Author's Guild* holding categorically protects using copyrighted material in datasets for machine learning purposes,[621] many legal scholars are not

---

[614] *See Protecting What We Love About the Internet: Our Efforts to Stop Online Piracy*, GOOGLE PUB. POL'Y BLOG (Nov. 7, 2019), https://www.blog.google/outreach-initiatives/public-policy/protecting-what-we-love-about-internet-our-efforts-stop-online-piracy/.

[615] *See* JENNIFER M. URBAN, JOE KARAGANIS & BRIANNA M. SCHOFIELD, NOTICE & TAKEDOWN IN EVERYDAY PRACTICE 39 (2017) (illustrating the difficulty that online service providers face in manually evaluating a large volume of data for potential infringement; for example, one online service provider explained that "out of fear of failing to remove infringing material, and motivated by the threat of statutory damages, its staff will take "six passes to try to find the [identified content]."); *see also* Letter from Thom Tillis, Marsha Blackburn, Christopher A. Coons, Dianne Feinstein et. al, to Sundar Pichai, Chief Executive Officer, Google Inc. (Sept. 3, 2019), https://www.ipwatchdog.com/wp-content/uploads/2019/09/9.3-Content-ID-Ltr.pdf ("We have heard from copyright holders who have been denied access to Content ID tools, and as a result, are at a significant disadvantage to prevent repeated uploading of content that they have previously identified as infringing. They are left with the choice of spending hours each week seeking out and sending notices about the same copyrighted works, or allowing their intellectual property to be misappropriated.").

[616] *See* GOOGLE, HOW GOOGLE FIGHTS PIRACY 6 (2016). To illustrate the costs of implementing Content ID on a large-scale platform, Google announced in a report in 2016 that YouTube had invested more than $60 million in Content ID.

[617] *See* Sobel, *supra* note 610, at 66-79.

[618] *See Authors Guild v. Google Inc.*, 804 F.3d 202 (2d Cir. 2015).

[619] *Id.*

[620] *Id.* at 216-17.

[621] Matthew Stewart, *The Most Important Court Decision For Data Science and Machine Learning*, TOWARDS DATA SCI. (Oct. 31, 2019), https://towardsdatascience.com/the-most-important-supreme-court-decision-for-data-science-and-machine-learning-44cfc1c1bcaf.

so sure about such a broad holding, especially because fair use is so fact-intensive.[622] Indeed, while Google Books used copyrighted works for a non-expressive purpose, Sobel notes that machine learning models may increasingly be able to glean value from a work's expressive aspects.[623] Therefore, until courts and legislators provide more clarity on the applicability of fair use in the machine learning context, the NRC should still require data uploaders to attest that they own the rights to the data.

B. Users must be able to license their data to other users.

If the NRC enabled private data sharing, users would need to make clear what rights other NRC users have over the uploaders' shared data. The NRC would have two basic options for creating IP licensing schemes: (1) the NRC could permit researchers to use whatever IP license they wish when sharing their private data; or (2) the NRC could mandate a uniform license across the board for all data that is uploaded.

1. Researcher's Choice of License

Allowing researchers to craft their own IP licensing agreements when sharing private data with other researchers would be the most frictionless solution from the perspective of the uploader; it would allow them to share exactly what they want and restrict use to only certain contexts. This choice of license seems to be important to data sharers.[624] Indeed, many data scientists and engineers have written guides advising members of the open-source community on how they should go about choosing specific licenses for their work.[625] GitHub, an open-source code-sharing platform, permits its users to choose from dozens of

---

[622] *See, e.g.*, James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 661 (2016); Sobel, *supra* note 610, at 51-57.
[623] *See* Sobel, *supra* note 610, at 57.
[624] *See* Anna I. Krylov et. al., *What is the Price of Open Source Software?* 6 J. PHYSICAL CHEMISTRY LETTERS 2751, 2753 (2015) (explaining that budding researchers considering commercialization may be particularly concerned about what licenses are available, since a "strictly open-source environment may furthermore disincentivize young researchers to make new code available right away, lest their ability to publish papers be short-circuited by a more senior researcher with an army of postdocs poised to take advantage of any new code.").
[625] *See, e.g., A Data Scientist's Guide to Open-Source Licensing,* TOWARDS DATA SCI. (Nov. 4, 2018), https://towardsdatascience.com/a-data-scientists-guide-to-open-source-licensing-c70d5fe42079; *Choose an Open-Source License*, https://choosealicense.com (last visited Apr. 10, 2022).

licenses,[626] and FigShare, a data-sharing platform for researchers, likewise supports a host of different Creative Commons licenses.[627] Some datasets even have their own custom IP licensing agreements. The Twitter academic dataset, for instance, is licensed according to Twitter's own developer agreement and noncommercial use policies, not to an existing open-source license.[628]

However, there are disadvantages to such flexibility. Just because different licenses might be allowed doesn't mean these licenses will be fully understood by all users. Adopting multiple licenses may result in increased accidental infringement. Indeed, a study conducted by the Institute of Electrical and Electronics Engineers found that "although [software] developers clearly understood cases involving one license, they struggled when multiple licenses were involved,"[629] and in particular, were found to "lack the knowledge and understanding to tease apart license interactions across multiple situations."[630]

In particular, researchers unfamiliar with the allowances provided by different data licenses, in contexts where more than one license is implemented, may lead to certain licenses being violated. For example, when researchers were surveyed regarding their understanding of copyright transfer agreements in the IP commercialization process, they only demonstrated an average 33 percent score on a knowledge-testing survey.[631]

2. Uniform Licensing Agreement

The second option available to the NRC would be to mandate that all private data be licensed under a single uniform license. For the NRC

---

[626] *Licensing a Repository,* GITHUB, https://docs.github.com/en/github/creating-cloning-and-archiving-repositories/licensing-a-repository (last visited Mar. 21, 2022).
[627] *What is the Most Appropriate Licence for My Data?*, FIGSHARE, https://help.figshare.com/article/what-is-the-most-appropriate-licence-for-my-data (last visited Mar. 21, 2022).
[628] *See Developer Agreement*, TWITTER (Mar. 10, 2020), https://developer.twitter.com/en/developer-terms/agreement; *Non-commercial Use of the Twitter API*, TWITTER, https://developer.twitter.com/en/developer-terms/commercial-terms (last visited Mar. 21, 2022).
[629] *See* Daniel A. Almeida et. al, *Do Software Developers Understand Open Source Licenses?*, 25 IEEE INT'L CONF. ON PROGRAM COMPREHENSION 1 (2017).
[630] *Id.* at 9.
[631] Alexandra Kohn & Jessica Lange, *Confused About Copyright? Assessing Researchers' Comprehension of Copyright Transfer Agreements,* 6 J. LIBRARIANSHIP & SCHOLARLY COMMC'N. 1, 9 (2018).

administration itself, this may be the more straightforward option, since users could be notified upon login about the appropriate use of data. The disadvantage of this strategy is that it may deter would-be researchers who would share data under a narrower license.[632] Given the desire to allow researchers to innovate freely, there may be concerns about adopting a restrictive licensing agreement. Nonetheless, several options of licensing agreements would still be available for adoption, and this pathway would require choosing a uniform agreement from these options, with the possibility of allowing an opt-out of this default license.

If the NRC were to implement a uniform license, it could look to the licensing agreements leveraged by institutional research clouds, such as the Harvard Dataverse as an analogy in determining best practices for its own licensing agreements. The model adopted by the Dataverse is a default use of the CC0 Public Domain Dedication "because of its name recognition in the scientific community" and its "use by repositories as well as scientific journals that require the deposit of open data."[633] Like an unrestricted Creative Commons or Open Data license, a public domain license would allow the data it governs to be used in any context, even commercial ones, and would also allow reproduction and creation of derivatives from the data.

Alternatively, the NRC could have a default open license while also permitting researchers to choose from a handful of more restrictive licenses if they wish. For example, the Harvard Dataverse notably allows uploaders to opt out of the CC0 if needed and specify custom terms of use. The Australian Research Data Commons and data-sharing platform FigShare[634] also use a default CC0 license but nevertheless permit researchers to use a conditioned Creative Commons license. These conditioned licenses can, for instance, require attribution to the original owner, prevent exact reproduction, or only allow use for noncommercial contexts. This may also help accommodate researchers who seek to upload datasets incorporating third-party data that holds a more

---

[632] *See* WILL FRASS, JO CROSS & VICTORIA GARDNER, TAYLOR & FRANCIS OPEN ACCESS SURVEY JUNE 2014 15 (2014). Note that lack of IP literacy could act as an additional deterrent to uploaders. The Taylor and Francis Open Access Survey of 2014 found that "63% of respondents indicated a lack of understanding of publisher policy as an important or very important factor in failing to deposit an article in an IR [Institutional Repository]." *Id.*

[633] *Dataverse Community Norms*, HARV. DATAVERSE, https://dataverse.org/best-practices/dataverse-community-norms (last visited Mar. 21, 2022).

[634] *Copyright and License Policy*, FIGSHARE, https://help.figshare.com/article/copyright-and-license-policy (last visited Mar. 21, 2022).

restrictive license, since a "combined dataset will adopt the most restrictive condition(s) of its component parts."[635]

If the NRC goes down this route of giving users the choice of a narrower license, it would also shift some liability to users—or to the NRC itself—by relying on users to abide by the license. Approaches to enforcement would vary, depending on the amount of responsibility in enforcement and, by extension, the liability the NRC seeks to take on. For example, in the Harvard Dataverse, if an uploader decides to opt out of a default open license and pursue their own custom licensing agreement over uploaded data, the Dataverse's General Terms of Use absolve this particular cloud from resource-heavy enforcement responsibilities by stating that it "has no obligation to aid or support either party of the Agreement in the execution or enforcement of the Data Use Agreement's terms."[636]

III. CURRENT STATE OF AI ETHICS FRAMEWORKS

AI ethics frameworks (or principles, guidelines) attempt to address the ethical concerns related to the development, deployment, and use of AI within prospective organizations. We briefly discuss the current landscape of AI ethics frameworks, while noting that this is still an emergent topic without broad consensus.

Between 2015 and 2020, governments, technology companies, international organizations, professional organizations, and researchers around the world have published some 117 documents related to AI ethics.[637] These frameworks aim to tackle the disruptive potential of AI technologies by producing normative principles and "best practice" recommendations.[638] Due to the prominence of essentially contested concepts in AI ethics—i.e., words such as fairness, equity, privacy that have different meanings for different audiences[639]—as well as the lack of binding professional history and accountability mechanisms, those

---

[635] AUSTL. DATA RSCH. COMMONS, RESEARCH DATA RIGHTS MANAGING GUIDE 6 (2019).
[636] *See Harvard Dataverse General Terms of Use*, HARV. DATAVERSE (2021), https://dataverse.org/best-practices/harvard-dataverse-general-terms-use (last visited Mar. 21, 2022).
[637] STAN. U. INST. OF HUM.-CENTERED A.I., ARTIFICIAL INTELLIGENCE INDEX REPORT 2021 125-34 (2021).
[638] Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, 30 MINDS & MACHS. 99 (2020).
[639] Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J.L. & TECH. 117, 182 (2021).

frameworks are often high level and self-regulatory, posing little threat to potential breaches to ethical conduct.[640]

A. Federal Frameworks

In the United States, there is no central guiding framework on the responsible development and application of AI across the federal government. Some government agencies have adopted or are in the process of adopting their own AI framework, while others have not published such guidelines. The following are published federal AI ethical frameworks as of August 2021:

- After 15 months of deliberation with leading AI experts, the Department of Defense (DOD) adopted a series of ethical principles for the use of AI in February 2020 that align with the existing DOD mission and stakeholders.[641]

- The General Services Administration (GSA), tasked by the Office of Management and Budget (OMB) in the Federal Data Strategy 2020 Action Plan, developed a Data Ethics Framework in February 2020 to help federal personnel make ethical decisions as they acquire, manage, and use data.[642]

- The Government Accountability Office (GAO) developed an AI accountability framework in June 2020 for federal agencies and other entities involved in the design, development, deployment, and continuous monitoring of AI systems to help ensure accountability and responsible use of AI.[643]

- The Office of the Director of National Intelligence (ODNI)

---

[640] Brent Mittlestadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 NATURE MACH. INTEL. 501 (2019).
[641] *DOD Adopts Ethical Principles for Artificial Intelligence*, U.S. DEP'T DEFENSE (Feb. 24, 2020), https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.
[642] PRESIDENT'S MGMT. AGENDA, FEDERAL DATA STRATEGY: DATA ETHICS FRAMEWORK (2020).
[643] *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, U.S. GOV'T ACCOUNTABILITY OFF. (June 30, 2021), https://www.gao.gov/products/gao-21-519sp.

> released the *Principles of AI Ethics for the Intelligence Community* in July 2020 to guide the intelligence community's (IC) ethical development and use of AI to solve intelligence problems.[644]
>
> • The National Security Commission on Artificial Intelligence (NSCAI) published a set of best practices in July 2020 (later revised and integrated into the Commission's 2021 Final Report) for agencies critical to national security to implement as a paradigm for the responsible development and fielding of AI systems.[645]

While these frameworks can help guide the NRC's approach to ethics, we refrain from recommending a specific framework for several reasons. First, despite growing calls for applied ethics in the AI community, developing an AI ethics framework is still an emerging area. The lack of a unified government standard poses challenges to the establishment of the NRC's ethics review process.

Second, there are, in fact, significant differences among ethics frameworks published by various federal agencies. For example, NSCAI laid out differences between its recommended practices and those by DOD and IC.[646] Moreover, among the five frameworks above, the GSA Framework focused only on the ethical conduct of federal employees when dealing with data while others focused on the ethical development and application of AI systems specifically.

Third, the ethics framework for adopting AI technology may be different from a framework for assessing research. Most federal agencies develop frameworks to guide the use of AI-driven solutions for agency-specific tasks. For example, DOD's ethical principles only apply to

---

[644] *Principles of Artificial Intelligence Ethics for the Intelligence Community*, OFF. DIR. NAT'L INTELL., https://www.odni.gov/index.php/features/2763-principles-of-artificial-intelligence-ethics-for-the-intelligence-community (last visited Mar. 21, 2022).

[645] *Key Considerations for Responsible Development and Fielding of Artificial Intelligence*, NAT'L SEC. COMM'N A.I. (2021), https://www.nscai.gov/key-considerations/.

[646] *Recommended Practices*, NAT'L SEC. COMM'N A.I., https://www.nscai.gov/wp-content/uploads/2021/01/Key-Considerations-Supporting-Visuals.pdf (last visited Mar. 21, 2022).

defense-specific combat or noncombat AI systems.[647] In the absence of a central federal guideline, the NRC should not adopt a framework by a particular agency because these frameworks are not necessarily designed for the wide range of research contemplated for the NRC. The work on frameworks may nonetheless provide a useful starting point for NRC's ethics process.

IV. STAFFING AND EXPERTISE

As noted throughout this Article, the success of the NRC will depend on human resources—both within the NRC as well as across government—to resolve the many challenges the NRC promises to tackle. While we refrain from providing an organizational chart, we list the dimensions where staffing and expertise will be critical to the success of the NRC. This list is not meant to be exhaustive, but to highlight the vital importance of human resources.

Human Resource Areas:

- Computing
  - System administrators
  - Data center engineers
  - Research software engineers
  - Research application developers
- Data
  - Data officers
  - Agency liaisons
  - Data architects
  - Data scientists
- Grant administrators
- Contracting officers
- Support and training staff
- Privacy staff (technical and legal)
- Ethics staff
- Cybersecurity staff

---

[647] DEFENSE INNOVATION BD., AI PRINCIPLES: RECOMMENDATIONS ON THE ETHICAL USE OF ARTIFICIAL INTELLIGENCE BY THE DEPARTMENT OF DEFENSE (2019).

VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

ARTICLES

IN DEFENSE OF (VIRTUOUS) AUTONOMOUS WEAPONS

Don Howard

IN DEFENSE OF (VIRTUOUS) AUTONOMOUS WEAPONS

*Don Howard**

INTRODUCTION

In 2012, Human Rights Watch (HRW) issued a call for a global ban on autonomous weapons.[1] A new NGO, the Campaign to Stop Killer Robots (CSKR) was formed in October 2012 to promote such a ban. In 2015, the Future of Life Institute (FLI) issued a new call for a ban, though now restricted to offensive autonomous weapons.[2] The FLI proposal garnered the support of tens of thousands of signatories, including such prominent figures as Elon Musk and Stephen Hawking, and generated considerable attention in the international press and on social media. Meanwhile, the CSKR helped to organize "informal meetings of experts" starting in 2014 in Geneva under the auspices of the UN's Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (CCW) for the purpose of exploring the possibility of adding an autonomous weapons ban to existing bans on land mines and blinding lasers, among other banned or restricted weapons.[3] In 2017 these sessions were elevated to the level of annual and still ongoing meetings of a formally constituted Group of Governmental Experts (GGE).[4] Against the background of these developments on the international legal front, an extensive literature on the ethics and policy of autonomous weapons has emerged and media attention to the debate has intensified. At least in the public arena, momentum seems to be building for some kind of ban.

Is a ban the right way to go? I think not. There are obvious questions of law, policy, and ethics that must be weighed regarding autonomous weapons. But, in my opinion, imposing a total ban, even if

*Professor, Department of Philosophy, University of Notre Dame.

[1] *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2019), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

[2] *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, FUTURE OF LIFE INST. (July 28, 2015), http://futureoflife.org/open-letter-autonomous-weapons/.

[3] Campaign (2015). "Step up the CCW Mandate." Campaign to Stop Killer Robots. http://www.stopkillerrobots.org/2015/06/mandateccw/.

[4] *Convention on Certain Conventional Weapons – Group of Governmental Experts on Lethal Autonomous Weapons*, UNITED NATIONS, https://meetings.unoda.org/meeting/ccw-gge-2017.

only a ban on offensive autonomous weapons, risks our depriving ourselves of tools that can continue the progress already made with the advent of "smart" weapons in reducing the suffering that will always be part of war, especially by way of still further reductions in harm to non-combatants. Moreover, as I will argue, we can construct effective means for norming the use of autonomous weapons short of a total ban by building upon the foundation of existing requirements stipulated in Article 36 of Protocol I to the Geneva Conventions that all new weapons technologies be reviewed for compliance with the International Law of Armed Conflict (ILOAC) and International Humanitarian Law (IHL).

I begin with a critical review of several of the most commonly encountered arguments in favor of a ban. That is followed by a discussion of the moral opportunities afforded by enhanced autonomy. I conclude with a concrete policy proposal based upon the principle of Article 36 review.

I. ARGUMENTS FOR A BAN ON AUTONOMOUS WEAPONS

Many arguments have been adduced for some kind of ban on autonomous weapons. They are too numerous and diverse all to be reviewed here. I have, therefore, chosen to focus on six of the most compelling arguments, as judged by their prominence in the literature and their seeming effectiveness in moving public opinion.

A. Morality, Emotions, and Robots

The original HRW call for an autonomous weapons ban placed surprisingly heavy emphasis on an argument that invites skepticism if not outright scorn. The argument is this: Morality requires an emotional capacity. Robots cannot feel emotions. Therefore, robot weapons are inherently immoral.[5]

One understands the idea behind this argument. In many situations, the ability to feel emotions makes possible an empathic relation to those affected by our actions, which includes an appreciation of their needs and fears. One feels oneself into the place of the other. And my Roomba cannot do that. Moreover, it is an empirical fact of

[5] *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2019), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots. In fairness, this is my distillation of an extended argument that includes acknowledgment of, if not an adequate response to, some of the critical points that I make. But it is an accurate representation of the main thrust of the report's argument.

considerable importance that emotional responses can be powerful enablers of moral action and powerful brakes on immoral action. One might well argue that merely knowing the good does not suffice for doing the good, that knowledge is ineffective without the will to act. The HRW report touches upon all of these points. But there are still at least three serious problems with this argument.

First, there is a long tradition in moral philosophy, from Plato to Kant and beyond, that holds that emotion is an impediment, not an aid to morality, because emotion clouds reason. That that can be so is obvious from long experience. Sometimes emotions of misplaced sympathy lead one to act more kindly toward some than reason would dictate, as when one male faculty member declines to report a case of possible sexual abuse or gender discrimination by another male colleague out of sympathy for that friend, whose career might suffer. Emotions do not always connect us in proper measure to everyone whose interests are involved. Second, and far more importantly, not all emotions move us to sympathy or kindness. Some move us to do truly horrible things, as when fear motivates racist violence. To this point I will return a bit later.

The third problem with this argument is that there can be no first principles proof for the claim that robots cannot sense or express emotions, unless one simply defines emotions as something distinctly human. But that is an evasion, not an argument. No, this is an empirical question, the answer to which depends on progress in research and development. In the ten years since the original HRW call, some developers have claimed considerable progress in designing robots that are said to be able to read human emotions and respond in emotionally appropriate ways. The most widely publicized early example was the robot, Pepper, that was announced in 2014 and brought to market in 2015 by Aldebaran.[6] And while she prefers the language of "sociability" to that of "emotion," the development of such robots has long been the focus of Cynthia Breazeal's highly innovative Personal Robotics group in MIT's Media Lab.[7] It goes without saying that none of these robots yet evince anything like a full, human-like, emotional capacity. But a lot of progress has been made, and that is just the point. Only time will tell to what extent robot emotions will be realized.

---

[6] *SoftBank Mobile and Aldebaran Unveil "Pepper" – the World's First Personal Robot that Reads Emotions*, SOFTBANK (June 5, 2014), https://www.softbank.jp/en/corp/group/sbm/news/press/2014/20140605_01/.

[7] *Cynthia Breazeal*, MIT MEDIA LAB PEOPLE, https://www.media.mit.edu/people/cynthiab/overview/ (last visited . . .).

Serious conceptual confusions also plague discussions of the potential emotional capacities of robots. That current robotic technology cannot produce in robots the kind of emotional capacity that we recognize in ourselves is, as noted, not worth disputing, if only because human emotion requires the biology of an endocrine system. But is that the kind of competence needed in autonomous weapons? Do we really need weapons that cry? No. If an emotional capacity is needed, it might only be the ability to read human emotion and to respond in emotionally appropriate ways. One can well imagine that a sentry-bot might do its job more reliably were it able to sense fear, nervousness, or anger, even if it does not, itself, experience such. It is important to keep the difference in mind, because designing robots that read emotion and respond in emotionally appropriate ways is, from an engineering point of view, a much more tractable problem than designing robots that genuinely feel sadness or remorse. So the argument about emotional capacity proves little or nothing about the wisdom of developing and fielding autonomous weapons. That might be why one hears it less frequently today.

B. Discrimination and Proportionality

The other major argument in the 2012 HRW call for an autonomous weapons ban was that robots are inherently incapable of respecting the International Law of Armed Conflict and International Humanitarian Law because they lack the ability to distinguish combatants from non-combatants and the ability to make judgments about proportionality.[8] There is no disputing the fact that no current weapons system has the ability to make all of the subtle distinctions that human combatants must and often do make between, say, a nervous suicide bomber walking up to a checkpoint in Tikrit and a pregnant woman on her way home from shopping made anxious by all of the foreign force on display everyday in what was once her happy home town. But that obvious fact does not settle the question.

First, as with the question of robots and emotion, what capabilities we might engineer into weapons systems in the future is an empirical question, not one of principle. Will a robot ever be able to make the distinction just discussed between the suicide bomber and the pregnant shopper? Only time will tell, but it has to be noted that one of the areas

[8] *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2019), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

of most rapid progress in artificial intelligence is pattern recognition. If a deep learning system can teach itself the difference between cats and dogs, why is it not conceivable that such a system can also learn to distinguish a gathering of Taliban leaders from a wedding party?

Second, the question is not whether perfection is possible in a robot's making such discriminations. The question is whether an autonomous weapons system can reach a reasonable threshold of success. After all, however high our expectations, we have to acknowledge that humans make far too many mistakes, sometimes costing the lives of the innocent, sometimes costing the lives of our own personnel. If an autonomous weapons system can consistently outperform human soldiers in distinguishing combatants from non-combatants, then there would be a moral gain. I would set the threshold higher still. But wherever that threshold lies, whether it can be met is an empirical question to be answered only by further research and development.

Third, the argument as stated seems to assume that discrimination is a context-independent competence. But this is not true. In fact, the kind of discrimination that is needed is highly context dependent. Consider, for example, the British Brimstone air-launched ground-attack missile system.[9] First deployed in 2005, it was originally designed as a fire-and-forget missile for use mainly against tanks and other mobile, armored vehicles. The original design assumed operation within a highly-circumscribed fire zone, one in which there was a reasonably low probability of encountering non-combatants. On-board sensor systems and programming, including active radar homing, handled target identification, acquisition, tracking, and firing, all based upon a set of situation specific targeting data uploaded before launch by a weapons system officer (WSO). Most importantly, the Brimstone system was designed to be capable of distinguishing between, say, a tank and a passenger vehicle, with the decision to fire based entirely on that distinction. If a suitable target was not found, the missile would self-destruct. The crucial fact is that, in this original configuration, Brimstone is an autonomous offensive weapons system capable of making context-specific discriminations between permissible and impermissible targets.

But the rules of engagement in Afghanistan required a person-in-the-loop, precluding the use of Brimstone in its original form. This led to

---

[9] *Brimstone*, MISSILE THREAT (July 30, 2021), https://missilethreat.csis.org/missile/brimstone/; *Brimstone Advanced Anti-Armour Missile*, ARMY TECH. (July 16, 2021), https://www.army-technology.com/projects/brimstone/.

the development in 2008 of a new, dual-mode model, with an added laser-targeting system that could be used by the pilot of the launch aircraft (so far only British Tornado and Typhoon aircraft) to guide the munitions to the target, the choice of mode being in the hands of the pilot.

Brimstone has now been modified for use also as a ground-based, antitank weapon, with the capacity of being mounted on unmanned ground vehicles. A variant model, Sea Spear, has been developed for use against swarms of small boats, in either a ship-launched or helicopter-launched version. Dual-mode Brimstone systems have been sold to Saudi Arabia, and there has been discussion of supplying the Sea Spear system to both Estonia and Ukraine.

There are many questions that one might ask about Brimstone. Should mode selection be in the hands of the pilot of the launch aircraft? What should be the constraints on the targeting data uploaded by the WSO? In what kinds of conflict arenas is such a system appropriate? But the main point, again, is that Brimstone is an example of an autonomous offensive weapons system about which it is claimed that, within an appropriately circumscribed context, it is capable of making the kind of discrimination required by ILOAC and IHL.

Whether the claimed discrimination capability is as robust as has been asserted and whether still more stringent constraints are appropriate are, of course, relevant questions. But I want to defer those questions to when I take up the proposal of an Article 36 based certification system for autonomous weapons. For now, let us just use the Brimstone example to illustrate the point that discrimination is a context-dependent issue and that, in some contexts of deployment, we might already have hardware and software capable of making the necessary discrimination.

C. Human Dignity

Of all of the arguments against autonomous weapons that are known to me, the most moving is perhaps that which asserts that the decision to kill must be left to a human being because, only thus, do we respect the essential human dignity of the human target and of all of those humans otherwise implicated in the use of violence in war. The idea is that a combatant makes him- or herself less than human by delegating a kill decision to an artificial system that cannot understand the victim's suffering and that one also, thereby, denies the human dignity of the victim. This argument takes center stage in the 2018 HRW report

updating the call for a ban on autonomous weapons and it has been widely discussed in the literature.[10]

That the argument from human dignity has the power to persuade is obvious. But is it a cogent argument? One curious feature of many invocations of the argument from dignity is the frequency with which its proponents openly acknowledge the difficulty of clearly articulating the core concept of human dignity. For example, Amanda Sharkey, one of the leaders of CSKR, devotes four dense pages of her 2019 paper, "Autonomous Weapons Systems, Killer Robots, and Human Dignity," to a surprisingly detailed cataloguing of the contradictions, ambiguities, and other muddles to be found in the literature, concluding that "it should be apparent that not only have some specific questions been raised about the impact of AWS on human dignity, but also that there is a lack of a clear consensus about what dignity is."[11] Equally noteworthy, however, is the fact that, having acknowledged the inherent lack of clarity of the concept of human dignity, the proponents of the dignity argument still commend its usefulness from a rhetorical point of view. Sharkey is straightforward about this. Having asked whether the dignity argument would help the campaign against killer robots, she responds:

> "There could be some campaigning advantages. Saying that something is against human dignity evokes a strong visceral response. Even though dignity is difficult to define clearly, people have an intuitive understanding of its meaning, and of the importance of maintaining and preserving it. Reference to human dignity can highlight a repugnance to the idea of machines having the power of life or death decisions over humans."[12]

Elvira Rosert and Frank Sauer make a similar rhetorical point in their 2018 paper, "Prohibiting Autonomous Weapons: Put Human

---

[10] *Heed the Call: A Moral and Legal Imperative to Ban Killer Robots*, HUMAN RIGHTS WATCH (Aug. 21, 2018), https://www.hrw.org/report/2018/08/21/heed-call/moral-and-legal-imperative-ban-killer-robots. *See e.g.*, Michael Horowitz, *The Ethics and Morality of Robotic Warfare: Assessing the Debate Over Autonomous Weapons*, 145 J. OF THE AM. ACAD. OF ARTS & SCI., no. 4, 2016, at 25–36 (2016); Amanda Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, 21 ETHICS & INFO. TECH. 75, 75–87 (2018); Elvira Rosert & Frank Sauer, *Prohibiting Autonomous Weapons: Put Human Dignity First*, 10 GLOBAL POLICY, no. 3, 2019, at 370–75.

[11] Amanda Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, 21 ETHICS & INFO. TECH. 75, 82 (2018).

[12] Amanda Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, 21 ETHICS & INFO. TECH. 75, 83 (2018).

Dignity First," writing: "From a strategic communication point of view, adjusting the message toward the infringement on human dignity would have the general benefit of dampening the overall level of contention."[13]

Those dedicated to a cause are not to be faulted for thinking carefully about the rhetorical impact of their arguments. But we must remember that the ultimate aim of the campaign for a ban on autonomous weapons is the crafting of new international law or other ways of norming the use of such weapons, and premises that work by eliciting a visceral response might not serve well as a basis for that latter enterprise, one in which clarity is most definitely a virtue. Some champions of the dignity argument, such as the authors of the 2018 Human Rights Watch call for a ban,[14] will respond by claiming that the appeal to human dignity already serves well as a basis for International Humanitarian Law (IHL) and the International Law of Armed Conflict (ILOAC) in the form of the Martens Clause, which was incorporated in the 1899 Hague Convention and added to the Geneva Conventions in Additional Protocol 1 of 1977:

> "In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience."[15]

Appeals to the Martens clause have played an important role in the arguments leading to the adoption of several additions to ILOAC, such as the ban on blinding lasers. But it is well to remember what led to the adoption of the Martens clause in the first place. It was added to the Hague Convention precisely to paper over issues about which the delegates could not reach consensus by reasoning from other, clear, legal principles, and there has since been a long history of debate and disagreement over how to interpret the clause, precisely because of the mentioned unclarity in such notions as essential human dignity.[16]

---

[13] Elvira Rosert & Frank Sauer, *Prohibiting Autonomous Weapons: Put Human Dignity First*, 10 GLOBAL POLICY, no. 3, 2019, at 370–75.
[14] *Heed the Call, supra* note 10, 8-43.
[15] Geneva Conventions in Additional Protocol 1 of 1977
[16] Rupert Ticehurst, *The Martens Clause and the Laws of Armed Conflict* 317 INT'L REV. OF THE RED CROSS, April 1997, at 125–34.

Still, such arguments work well by way of stirring the emotions. When I think about the argument from human dignity, my mind goes immediately to a remarkable moment in the climactic battle sequence of the movie, "Saving Private Ryan," where, at the end of a harrowing, hand-to-hand struggle, a tough German soldier rolls atop the exhausted Private Stanley Mellish, taking a bayonet from Mellish's hand. Mellish begs the German, "Listen to me. Listen to me. Stop. Stop." But the German slowly pushes the bayonet into Mellish's heart, holding him almost tenderly and gently whispering to him, "Shhh. Shhh," like a father to a frightened son, until Mellish breathes his last. The German understood Mellish's suffering and fear, and one wants to think that Mellish might have taken comfort at the end from the warmth of the German's embrace. I think that Steven Spielberg was trying to make a complicated point about morality in war with that scene. We are supposed to despise the German soldier, but, ironically, his act of killing becomes an act of love. There can be no more essentially human moment in war than such an intimate, face-to-face act of violence.

I am so moved by such a scene that even just describing it leaves me emotionally and psychologically drained. I have to take a deep breath. I have to recenter and relax. Only then can I stop and think clearly.

What do I think? When emotion subsides and my head clears, I am horrified by the suggestion that, because Mellish was killed by a human who sought to comfort him in his dying moment, there was, therefore, in that act, respect for human dignity of a kind that would be missing were Mellish killed by a robot. On the contrary, one can argue that killing in any form, even in war or self-defense, entails the denial of human dignity, if there is such. But the problem is that killing in war and killing in self-defense are sometimes necessary, however fundamentally inhumane that killing might be. Kill we must, but let's not make killing out to be anything other than what it really is, namely, a horrible, if unavoidable, denial of both our own and the victim's humanity. This is why even people fighting on the "good" side in a perfectly just, defensive war experience killing in war as morally corrosive. I think that any attempt to make it appear that humans killing humans in war is more humane than robots killing humans in war is to lose sight of our humanity in a most profound way.

What, then, of the argument against autonomous weapons from the premise of essential human dignity? I think that, killing in war being the denial of human dignity, the morally responsible thing to do is to minimize it, to do no more killing, to inflict no more harm than is absolutely necessary for the achievement of proper ends. That principle

has long been fundamental in International Humanitarian Law and the International Law of Armed Conflict going all the way back to the 1868 St. Petersburg Declaration, which banned weapons and practices that cause unnecessary suffering, and it is now codified as Rule 70 of Customary IHL.[17] I think that I respect the dignity of my enemy and of all of those who suffer in war by doing everything that I can to minimize violence and the harm that I do to others. If autonomous weapons further that end, then so be it. Were I in Mellish's situation, knowing that I am going to die, what I would want most would be for it to be a quick and painless death. Soothing words from my killer would only add to the insult.

D. Increasing the Temptation to Engage in Conflict

If autonomous weapons promise both to minimize a nation's own casualties and to minimize harm to non-combatants, will there not be an added incentive to initiate conflict, say by intervening in conflict situations where, previously, the threat to one's own troops or worries about collateral casualties would have made the risk not worth the gain or the intervention politically unacceptable? Could we imagine that a high-minded effort to minimize death and suffering might, in this way, ironically, increase death and suffering by increasing the number of fights in which we engage?

The worry is not new to autonomous weapons. The same concern has often been expressed about the "smart" weapons that featured so prominently already in the First Gulf War. Thus, more than one critic of US military policy has argued that we would not have intervened in the Libyan conflict had it required troops on the ground and that Obama judged it politically feasible to intervene because "smart" munitions gave us an ability to assist the anti-Gaddafi forces without seriously risking the lives of US troops.[18] While that intervention toppled the Gaddafi regime, the long-term consequences, including bloody civil war and Libya's becoming a terrorist haven, proved to be catastrophic.

That the availability of autonomous weapons might increase the temptation to engage in conflict cannot be denied. But, as with so many of the other arguments against autonomous weapons, the first response

---

[17] *Rule 70. Weapons of a Nature to Cause Superfluous Injury or Unnecessary Suffering*, IHL DATABASE, https://ihl-databases.icrc.org/customary-ihl/eng/docindex/v1_rul_rule70 (last visited . ..).

[18] *See, e.g.* Lawrence Kaplan, *More Questions than Answers: Obama, Libya, and the Dubious Ethics of Modern Air Wars*, THE NEW REPUBLIC (Mar. 22, 2011), https://newrepublic.com/article/85555/obama-libya-air-war-qaddafi-ethics.

is that, whether in fact such an effect occurs is an empirical question, as is the question of the magnitude of the effect. However, in this case, it is also a political and a moral question. It is not just whether such actions do occur, but whether they should. There are other examples of military intervention made politically and militarily easier by technology that have a very different moral valence than the Libyan conflict. The NATO intervention in Kosovo in 1998 is one such.[19] Opinion differs strongly about the net benefit of NATO intervention, but I am on the side of the argument that sees NATO's role in Kosovo as an exemplary model for the future. Mistakes were made and innocent civilians suffered. But an ethnic war of possibly catastrophic proportions was prevented. European and US public opinion would not have tolerated a massive NATO ground involvement in Kosovo. The good that was achieved was made possible by our ability to apply force with minimal risk to our own personnel and to non-combatants. Did we kill civilians who otherwise would not have died? We did. But how many Kosovar and Serbian lives did we save in the process? That is the proper question. And my reading of the evidence suggests that we probably saved many tens of thousands of lives.[20]

So the question is not whether the even greater reduction in suffering promised by autonomous weapons would lead to more military interventions. The question is, rather, what kinds and numbers of interventions would such a capability facilitate. If such a capability could have made it politically and militarily feasible to stop the slaughters in Rwanda, Cambodia, and Biafra - to name only the most horrific wars of the last several decades - then that would have been a moral gain.

E. An Autonomous Weapons Arms Race

The 2015 Future of Life Institute call for an offensive autonomous weapons ban foregrounded an argument mentioned but not as much emphasized in the 2012 HRW call for a total ban. This is the argument that, absent a ban, we will see a global autonomous weapons arms race that will make the nuclear weapons arms race pale by comparison.[21]

That there would be an autonomous weapons arms race is likely. After all, it is declared US policy to seek and maintain technological

---

[19] BENJAMIN LAMBETH, NATO'S AIR WAR FOR KOSOVO: A STRATEGIC AND OPERATIONAL ASSESSMENT (2001).

[20] Agon Maliqi, *Remembering the U.S. Intervention That Worked*, WASH. POST. (June 8, 2019), https://www.washingtonpost.com/opinions/2019/06/08/remembering-us-intervention-that-worked/.

[21] *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, FUTURE OF LIFE INST (July 28, 2015), http://futureoflife.org/open-letter-autonomous-weapons/.

dominance over all of our potential adversaries, and adversaries such as China and Russia have made clear their determination to narrow the gap if not to surpass US capabilities in at least some modes of conflict, space-based weapons being an especially noteworthy example.[22] We know that Russia has been developing various types of autonomous weapons. Other actors are also getting into the game. For example, in October 2015 reports on a recent military exercise, Iran announced that it was testing what it called "kamikaze robots," whatever that means.[23] In June 2021, it was reported that the Libyan government used Turkish-made, autonomous, weaponized drones in an attack on rebels.[24] Moreover, history has shown that adversaries capable of competing with innovative US military technologies have done so. Soviet era competition with the US in ballistic missile and space technology is probably the most famous example, because, the US did not always lead in that competition, certainly not in its earliest years, with Sputnik having been the first earth satellite and Yuri Gagarin the first human in space (see Wolfe 2013), and some argue that the US is now trailing behind Russia and China in the development of hypersonic weapons.[25] But the history of competition in weapons technology goes back far beyond the Cold War to the earliest days of technologized warfare. One thinks of competition in submarine and tank technology in World War II, or the tragic competition in poison gas weapons in World War I.

Competition in weapons development has, thus, been the norm for a long time. Why, then, would one think that there would be something importantly different about an autonomous weapons arms race? Cost might be one factor, some robotic systems being cheap by comparison with both conventional arms and human combatants. So there might be more players in a robot weapons arms race. But the cheap

---

[22] *See* GIAN GENTILE ET AL., A HISTORY OF THE THIRD OFFSET, 2014–2018 (2021); Abraham Mahshie, *Russia and China Could Team Up to Challenge US Space Superiority, Experts Say*, AIR FORCE MAG. (June 29, 2021), https://www.airforcemag.com/russia-china-team-up-challenge-us-space-superiority/.

[23] *Straight Truth, 'Kamikaze' robots debut in Iran Army Drill*, TEHRAN TIMES (Oct. 21, 2015), https://www.tehrantimes.com/news/250250/Kamikaze-robots-debut-in-Iran-Army-drill).

[24] Joe Hernandez, *A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says*, NPR (June 1, 2021), https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d.

[25] McLeary, Paul and Alexander Ward (2021). "U.S. 'Not as Advanced' as China and Russia on Hypersonic Tech, Space Force General Warns." *Politico*. November 20, 2021. https://www.politico.com/news/2021/11/20/hypersonic-technology-us-behind-china-russia-523130.

weapons are likely to be the less worrisome ones, there being a rough correlation between cost and destructive potential. But the price of entry for large-scale autonomy – with, say, integrated, autonomous command and control combined with autonomous weapons platforms for multiple levels and modes of combat across a large field of combat – will keep out all but a few actors, the US, Russia, and China being the main candidates.

Competition at that level could be worrisome. And one would expect to see greater levels of autonomy and increased system integration. But then the question is whether such development is likely to take those actors to a level where serious fears about a loss of human control is conceivable. Here, again, history is helpful. For, during the Cold War, both the US and the Soviet Union took steps in the direction of automating their nuclear attack response capabilities, the idea being that, if human operators do not survive, then the computers can launch the retaliatory strikes. Hollywood had fun with this theme, in movies like "War Games." But the attendant risks of such automated response capabilities were well understood, which is why we never went too far down that road and why we engineered multiple layers of checks and controls. We learned important lessons about the vulnerabilities of engineered systems to unanticipated failure modes. To be sure, we came close to nuclear Armageddon on too many occasions, but those were mostly human failures, and we came close to serious nuclear accidents on many more, and from those near-misses we learned still more about how to engineer against failure (see Schlosser 2013).

One final feature of the analogy between the nuclear arms race and an autonomous weapons arms race puzzles me greatly. The destructive capability of nuclear weapons is such that even a medium-scale, regional nuclear exchange could have globally catastrophic consequences. But autonomous weapons are, from one point of view, the next phase in a history of steadily dialing back destructive power thanks to our technology's making possible the ever-more-accurate delivery of force on a target. This trend line is no accident. It is deliberate policy at least in the US military. If competition in autonomous weapons development were to accelerate this trend, would that not be a moral gain rather than a loss (U.S. Mission Geneva 2019)?

F. Autonomous Weapons and an Artificial Intelligence Apocalypse

I can construct only one scenario through which an autonomous weapons arms race could leave us in a worse place than the nuclear

weapons arms race did, and that scenario is the one envisioned in the newest, and, I think, most curious, argument for an autonomous weapons ban. This argument asks us to imagine a time when, the singularity having arrived, the artificial intelligence is smarter than us and decides to use enhanced autonomous weapons capabilities either to eliminate humankind altogether or wreak comparable mayhem in service of a goal that we mere humans cannot comprehend. This is the Skynet apocalypse, famous from the "Terminator" movie series.

When first I saw this argument, I could not believe that serious people would promote it, because I tell all my students and all of my audiences never to look to Hollywood science fiction for guidance, for the obvious reason that Hollywood purveys, well, fiction, not fact, and fiction that preys on our deepest irrational fears, not reasonable extrapolations from current technology. Imagine my even greater surprise, therefore, when, in 2015, AI specialist, Toby Walsh, the main engine behind the new Future of Life Institute call for an autonomous weapons ban wrote, in an op-ed at CNN: "Once this genie is out of the bottle, there will be an arms race to improve on the initially rather crude robots. And the end point of such an arms race is precisely the sort of terrifying technology you see in 'Terminator. Hollywood got that part right."[26] Seriously? Are we really debating such an important issue on the basis of Hollywood nightmare films?

But let us be serious about the question and ask whether the "Terminator" apocalypse is a realistic scenario of such a kind that it should guide our thinking about weapons development policy. Is a "Terminator" apocalypse possible? Of course it is, from a purely logical point of view. There is nothing inherently contradictory in the concept of such a future. But if it is possible, and if it would mean the end of all human life, then must we not do everything possible to prevent it, starting with an immediate ban on all autonomous weapons development? However unlikely the possibility, the consequences would be so dire that all other possible futures are irrelevant. That seems like a reasonable argument. No?

The very reasonableness of the argument, or its seeming reasonableness, is the problem. If, in any policy debate, one assigns an infinite negative utility to a given possible outcome, such as the death of all humankind, then, no matter how tiny the probability, the product of negative infinity times that tiny probability totally overwhelms every

---

[26] Toby Walsh, *The Rise of the Killer Robots - And Why We Need to Stop Them*, CNN (October 26, 2015), http://www.cnn.com/2015/10/26/opinions/killer-robots-walsh/index.html.

other term in the expected utility calculation, rendering such a calculation useless for policy purposes. In other words, the invocation of an apocalypse means – and this is sometimes the goal – the end of rational deliberation.[27]

Another way to think about this is to realize that there are many conceivable apocalypse scenarios. Global climate change might render the planet uninhabitable for all higher life forms within a few hundred years if we pass a climate tipping point in the very near future. That is another, possible future. Must we, therefore, immediately subordinate all other human purposes to effecting not just an immediate end to CO2 equivalent emissions but also the active removal of CO2 equivalents from the atmosphere?

Of course it is also possible that a new "terminator" pathogen might evolve tomorrow, one vastly more virulent and lethal than the Spanish flu of 1918 or Ebola or COVID-19, one that could eliminate all human life. Therefore, instead of redirecting all of our resources to combating climate change, we should stop all travel, all meetings of two or more strangers, all animal farming, all raising of pets, all activities that might facilitate viral transmission among individuals and species. And we should redirect all of our research efforts to studying viral evolution and to the development of new vaccines and disease treatments. But wait a minute. We cannot do such research, because the research, itself, might accidently create such a "terminator" pathogen that might be accidentally released into the wild. It could happen.

Perhaps our demise might be caused not by our actions but by our inaction. It is possible that a heretofore undiscovered space rock of a size capable of causing an extinction-level event might be found next year to be hurtling toward a collision with Earth that could cause a catastrophe on the scale of that which produced the cretaceous extinction. Again, Hollywood loves this scenario. But it is possible, as witness the sudden appearance in 2015 of a previously unknown asteroid, 2015 TB145, large enough, at 400m, to cause continent-scale devastation, that passed nearer to Earth on Halloween than any other object of that size since

---

[27] Don Howard, *On the Moral and Intellectual Bankruptcy of Risk Analysis: Garbage In, Garbage Out,* SCIENCE MATTERS BLOG. (Sept. 26, 2014), http://donhoward-blog.nd.edu/ 2014/09/26/on-the-moral-and-intellectual-bankruptcy-of-risk-analysis-garbage-in-garbage-out/#.VjeO-H6rT4Y; Casadevall, Arturo, Michael Imperiale, Don Howard, *The Apocalypse as a Rhetorical Device in the Influenza Virus Gain-of-Function Debate*, MBIO: AN OPEN ACCCESS JOURNAL PUBLISHED BY THE AMERICAN SOCIETY FOR MICROBIOLOGY 5 (5) e01875-14 (Oct. 14, 2014), http://mbio.asm.org/content/5/5/e02062-14.full.

1999.[28] If our doing nothing would seal the fate of humankind, should we not redirect all of our resources to the most rapid possible development of a technology for deflecting such space objects from a collision course with Earth?

I trust that the point is clear. Invocations of apocalypse make rational policy decisions impossible. It is well to be mindful of all such possible catastrophes. But balanced good judgment would place more emphasis on the extremely low probabilities than on the infinite, negative utilities of such events. Prudence might well dictate our taking steps to minimize those probabilities still further or to mitigate harm in the case of events beyond our control. But neither prudence nor reason should lead us to act merely on the basis of possibility.

What, then, should we say about Terminator-AI apocalypse scenarios? What we should say is that, contrary to Toby Walsh's confident assertion that such an apocalypse is the "end point" of an autonomous weapons arms race, such an extrapolation from current technology is not supported by any evidence or compelling argumentation.

History has shown that forecasting technology development is a nearly impossible task. This point is forcefully made in a 1983 paper by Charles Townes, co-inventor of the transistor, in which he reflected on our poor record of technology forecasting in the twentieth century. He points to a 1937 report of a committee of experts assembled at the behest of President Roosevelt to assess technology trends as they might affect national policy and planning. Among the revolutionary technologies of the near future totally missed by the committee were: nuclear energy, radar, antibiotics, jet aircraft, rocketry, space exploration, computers, microelectronics, and genetic engineering. And Townes notes that the scientific and technical bases for nearly all of these developments were already in place in 1937.[29]

But what about the development of AI in particular? In fact, opinion is strongly divided over the pace and nature of advances in AI. There have been notable achievements in recent years, thanks, especially, to machine learning algorithms and neural nets. Some predict that the AI singularity – the point at which AI is supposed to surpasses human

---

[28] Todd Leopold, John Newsome, Jareen Imam, *Halloween Asteroid Resembling Skull Narrowly Misses Earth*, CNN (Oct. 21, 2015), https://www.cnn.com/2015/10/21/us/asteroid-earth-nasa-halloween-feat/index.html.

[29]Charles H. Townes, *Science, Technology, and Invention: Their Progress and Interactions*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES (1983) at 80: 7679–7683.

intelligence – might arrive within a couple of decades. But others push the date back to the end of the century.[30] Still others - and count me in this group - point out that the very question is probably not well posed, since human intelligence is not some one thing that will be achieved in artificial systems at a magical moment when the light of consciousness suddenly turns on in Watson. My expectation is that more and more human abilities will be better approximated, some even eclipsed, in the coming years, but that this will happen in a context-dependent and task specific way, not in the form of general AI. We should closely monitor all of these developments for their potential impact on human well being, and we need to be able to respond nimbly and quickly should serious threats emerge. But no one can pretend now to know that a Terminator-AI apocalypse is inevitable or even likely.

G. Differences Between Offensive and Defensive Weapons Systems.

An interesting twist in the 2015 Future of Life Institute call for an autonomous weapons ban is that it proposes to ban only offensive autonomous weapons. One might guess that one reason for this modification is that defensive autonomous weapons have been proving their effectiveness and basic safety for a number of years. The US Navy's fully autonomous, Phalanx, ship-borne, anti-missile defense system, which fires 20 mm projectiles at a rate of between 3,000 and 4,500 rounds per minute from six, revolving barrels, was first developed in late 1970s.[31] Israel first used its autonomous, Iron Dome anti-missile defense system in 2011.[32] And South Korea introduced the Samsung SGR-A1 border patrol robot, which has both a fully autonomous and a person-in-the-loop mode, in September of 2014.[33]

---

[30] In November 2015, Microsoft's head of research, Eric Horvitz, opened MIT's annual EmTech conference by noting that "the mastery of AI has been much harder than expected." (http://www.techrepublic.com/article/mastery-of-ai-has-been-harder-than-expected-and-future-is-uncertain-says-microsofts-ai-chief/)

[31] John Pike, MK 15 Phalanx Close-In Weapons System (CIWS), FAS Miltary Analysis Network (January 9, 2003), https://man.fas.org/dod-101/sys/ship/weaps/mk-15.htm.

[32] Missile Defense Project, "Iron Dome (Israel)," *Missile Threat*, Center for Strategic and International Studies (April 14, 2016), https://missilethreat.csis.org/defsys/iron-dome/.

[33] David Crane, *Samsung SGR-A1 Armed/Weaponized Robot Sentry (or 'Sentry Robot') Remote Weapons Station (RWS). Finally Ready for Prime Time?,* DEFENSE REVIEW (September 17, 2014), https://defensereview.com/samsung-sgr-a1-armedweaponized-robot-sentry-or-sentry-robot-remote-weapons-station-rws-finally-ready-for-prime-time/.

Many people seem to share the intuition that the rules of defensive warfare might differ from those for offensive action, I suppose on the grounds that killing in self-defense differs from initiating killing in morally relevant ways. That may be so in cases of individual self-defense, as when I am allowed to use deadly force to protect myself and others from an imminent threat of death, when the pre-emptive taking of life would not be permissible. But it is not at all obvious that there is a morally-relevant difference when it comes to the employment of autonomous weapons in war. Start with the fact that the tradition of Just War Theory and the body of international law founded upon it assumes that going to war is morally justified only to right a wrong, meaning that, in a sense, the only permissible war is one of defense against an aggressor. Of course, bad actors initiate conflict for bad reasons all the time (however much they might convince themselves that they are righting wrongs), and ILOAC and IHL seek to norm all such conflict. Consider next the fact that, in war of any kind, there is no perfect or even very clean distinction between offensive and defensive action. If someone shoots at me and I shoot back, that is clearly a defensive act, no? But what if I provoked the first shot by some tactic like reconnaissance in force, aimed at eliciting enemy fire? On the other hand, my initiating combat to secure an objective seems the epitome of an offensive action. But what if the ultimate goal were to secure, say, a high point for better defense against possible future assaults? Examples such as these are the daily bread of courses on ILOAC for young ROTC cadets and students of military law. They all go to prove the point that what makes an act offensive or defensive is highly context dependent and depends also on the larger aims and intentions of the actors.

But what about the weapons themselves? Surely there is no imaginable offensive use for Iron Dome or Phalanx. They were designed as defensive systems and have only been deployed for purposes of defense. Or have they? Iron Dome is an especially interesting example. It has been used so far mainly only in defense against Hamas missile attacks originating from within Gaza. While its effectiveness has been disputed, it has made many impressive kills and has surely prevented damage if not also saved lives. What could be a more morally just use of high technology? In fact, Iron Dome is only the first layer of Israel's evolving, multi-layer, anti-missile, defense system, that also includes the Arrow 2, Arrow 3, Arrow 4, and David's Sling systems that are designed to defend against not only Hamas's crude, short-range missiles but also

against tactical and intermediate-range ballistic missiles of the kind that Iran has developed.[34] Some observers think that the real goal of the overall program is to provide a comprehensive defensive shield against Iranian ballistic missiles so as to insulate Israel against retaliation if, for example, Israel chose to launch a pre-emptive strike against Iranian nuclear weapons facilities.[35] If so, then what appears a defensive capability becomes an offensive one by making possible offensive actions that would otherwise lead to unacceptable risk to one's own nation. The logic here is much like that in the debate about Star Wars in the 1980s. Who could not welcome a perfect defense against nuclear armed ICBMs? The Soviets, for one. They regarded Star Wars as a highly destabilizing technology because they feared that it would embolden the US to launch a pre-emptive strike secure in the faith that a Soviet retaliatory strike would fail.

The Phalanx system challenges the offensive-defensive distinction in the same way as Iron Dome, for a defense against anti-ship missiles facilitates offensive action in, say, the Straits of Hormuz, that otherwise might be too risky. But Phalanx also challenges the distinction in a more straightforward way. During the Iraq War it was already adapted for use by ground forces in such settings as perimeter defense against mortars and other small, fast munitions.[36] Mount it on a mobile platform, alter a few lines of code, and it would become a fearsome offensive weapon, obliterating bodies, buildings, and even heavy armor that might be in its path.

So there is no clear-cut distinction between offensive and defensive autonomous weapons sufficient to support the restriction of the Future of Life Institute's proposed ban to offensive weapons alone. The offensive-defensive distinction is functional and contextual, not structural, a matter not so much of technology as of human intention. The contextual nature of the offensive-defensive distinction reminds us of the point made earlier about the contextual nature of discrimination, and both points will be relevant when, shortly, we turn to the question of an Article 36-based alternative to a wholesale ban.

---

[34] Gili Cohen, *Why Does Israel Need Three Different Missile Defense Systems*?, HAARETZ (April 2, 2015), https://www.haaretz.com/.premium-why-does-israel-need-3-anti-missile-systems-1.5346632.
[35] John Hannah, *Israel Needs Weapons to Stop Iran's Bomb*, FOREIGN POLICY (October 15, 2021), https://foreignpolicy.com/2021/10/15/israel-idf-iran-nuclear-arms-weapons/.
[36] *20 mm Phalanx Close-in Weapon System (CIWS)*, NAVWEAPS (last updated, Jan. 6, 2022), http://www.navweaps.com/Weapons/WNUS_Phalanx.php.

II. MORAL ADVANTAGES OF AUTONOMY

All of the main arguments in favor of an autonomous weapons ban have been found wanting. Let us turn to the other side of the argument and remind ourselves about the claimed moral gains from the introduction of autonomous weapons. By far the most compelling case of this kind is that made by Ronald Arkin in his 2009 book, *Governing Lethal Behavior in Autonomous Robots*.[37] Do not dawdle over the particular architecture that Arkin suggests in that book, some of which is already dated, though his idea of the "ethical governor" is still worthy of attention.[38] Appreciate, instead, his main point, which is that humans are notoriously unreliable systems, that human combatants commit war crimes with frightening frequency, and that what we must ask of autonomous weapons systems is not moral perfection, but simply performance above the level of the average human soldier.

There is not space here to review in detail the study by the United States Army Medical Command's Office of the Surgeon General from the Iraq War upon which Arkin mainly bases his assessment of human combatant performance.[39] Suffice it to say that the numbers of admitted war crimes by US troops, the numbers of unreported but observed war crimes, and the self-reported ignorance about what even constitutes a war crime are staggering. With such empirical evidence as background, Arkin's claim to be able to build a "more moral" robot combatant seems far more plausible than one might initially have thought. Why?

Start with the obvious reasons. Autonomous weapons systems suffer from none of the human failings that so often produce immoral behavior in war. They feel no fear, hunger, fatigue, or anger over the death of a friend. Move on to the slightly less obvious reasons. Thus, a robot, not fearing for its own well-being, can easily err on the side of caution, choosing not to fire in moments of doubt (think of the suicide bomber/pregnant shopper scenario above), where a human might rightly have to err on the side of self-defense. Then consider still more important design constraints, such as those embodied in Arkin's "Ethical Adaptor," into which are programmed all relevant parts of ILOAC, IHL, and the

---

[37] Ronald Arkin, GOVERNING LETHAL BEHAVIOR IN AUTONOMOUS ROBOTS, Boca Raton, FL: Chapman Hall/CRC (2009).
[38] *Id.* at 127-133.
[39] Office of the Surgeon General, *Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05-07. Final Report,* DEPARTMENT OF COMMERCE, NATIONAL TECHNICAL REPORTS LIBRARY (November 7, 2006), https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB2010103335.xhtml#.

rules of engagement specific to given conflict arena or a specific action.[40] The Ethical Adapter blocks the "fire" option unless all of those prescriptions are satisfied. Arkin's robots could not fire (absent an override from a human operator) at all, unless the most stringent requirements are met. In the face of uncertainty about target identification, discrimination, applicability of rules of engagement, and so forth, the robot combatant defaults to the "no fire" option. Of course, other militaries could design the robots differently, say, by making "fire," rather than "no fire," the default. But hold that thought until, again, we turn to the discussion of an Article 36 regulatory regime.

Arkin illustrates the functioning of the Ethical Adaptor with several scenarios, one of which – a Taliban gathering in a cemetery for a funeral[41] – bears an eerie similarity to the horrific US attack on a Doctors without Borders (Médecins Sans Frontières - MSF) hospital in Kunduz, Afghanistan in October of 2015.[42] The rules of engagement as uploaded to the Ethical Adaptor would typically include specific coordinates for areas within which no fire would be permitted, including hospitals, schools, important cultural monuments, and other protected spaces. Likewise, no fire could be directed at any structure, vehicle, or individual displaying the red cross or the red crescent. This assumes, of course, sensor and AI capabilities adequate for spotting and correctly identifying such insignia, but, especially with structures and vehicles, where the symbol is commonly painted in large, high-contrast format on the roof, that is not a difficult problem. A fully autonomous drone designed as per Arkin's model that was tasked with the same action that led to the bombing of the MSF hospital in Kunduz simply would not have fired at the hospital. A human might have overridden that decision, but the robot would not have fired on its own. Moreover, the kind of robot weapon that Arkin has designed would even remind the human operator that a war crime might be committed if the action proceeds.

Another kind of moral gain from autonomous weapons was once pointed out to me by an undergraduate student – an engineering major – in my "Robot Ethics" class. He recalled the oft-expressed worry about the dehumanization of combat with standoff weapons, such as remotely

---

[40] Arkins, *supra* note 36 at 138-143.
[41] *Id*. at 157-161.
[42] Alissa J. Ruben, *Airstrike Hits Doctors Without Borders Hospital in Afghanistan*, NEW YORK TIMES (October 3, 2015), https://www.nytimes.com/2015/10/04/world/asia/afghanistan-bombing-hospital-doctors-without-borders-kunduz.html.

piloted drones. The concern is that the computer-game-like character of operator interfaces and controls, and the insulation of the operator from the direct risk of combat, might dull the moral sensitivity of the operator. But my student argued with deliberate and insightful irony, that the solution to the problem of dehumanization might be to take the human out of the loop, because it is the human operator who is, thus, dehumanized. For the record, I would dispute the dehumanization argument in the first place, because the typical drone operator often watches the target for many minutes, if not hours, and gets to know the humans on the receiving end of the munitions – including the wives, husbands, and children – far better than does, say, an artillery officer, a bombardier in a high-altitude bomber, or even the infantryman who gets, at best, a fleeting and indistinct glimpse of an enemy combatant across a wide, hazy, busy field of combat. That drone operators get to know their targets so well is part of the explanation for the extremely high reported rates of PTSD and other forms of combat stress among them.[43] Still, my student's point was a good one. If dehumanization is the problem, then take the dehumanized human operator out of the loop. This is really just a special case of Arkin's point about how stress and other contextual circumstances increase the likelihood of mistakes or deliberate bad acts by humans in combat and that, since robots are unaffected by such factors, they will not make those mistakes.

One of the most common criticisms of Arkin's model is the same voiced in the original HRW call for a ban, namely, that sensor systems and AI are not capable of distinguishing combatants from non-combatants, so that, even if the principle of discrimination is programmed into a robot weapon, it still cannot satisfy the requirements of international law. But we dealt with that point above, the two main responses having been: (1) what is or is not technically feasible is an empirical question to be decided by further research, not on a priori grounds, and (2) discrimination is usually a highly context-dependent challenge, and in some contexts, such as finding and identifying a Red Cross or Red Crescent symbol, the problem is easily solved.

The other major criticism of Arkin's model is that, since it assumes a conventional, structured, top-down, decision tree approach to programming ethics and law into autonomous weapons, it cannot deal

---

[43] Chappelle, Goodman, Reardon, Thompson, *An analysis of post-traumatic stress symptoms in United States Air Force drone operators*. J ANXIETY DISORD. 2014 Jun;28(5):480-7. doi: 10.1016/j.janxdis.2014.05.003. Epub 2014 May 17. PMID: 24907535.

with the often bewildering complexity of real battlefield situations. The basis of the objection is a simple and old worry about any rule-based or algorithmic approach to ethical decision making, such as deontology or consequentialism. It is that one cannot write a rule or build a decision tree to cover every contingency and that the consequentialist's calculation of benefit and risk is often impossible to carry out when not all consequences can be foreseen. The objection is a good one, at least by way of pointing out the limited range of applicability of Arkin-type autonomous weapons systems.

But Arkin's model for ethical autonomous weapons design is only a beginning. This last objection – that one cannot write a rule to cover every contingency – is the main reason why some of us are hard at work on developing a very different approach to ethics programming for artificial systems, one inspired by the virtue ethics tradition and implemented via neural nets and machine learning algorithms. The idea – already explored in concept by Wendell Wallach and Colin Allen in their 2010 book, *Moral Machines* (Wallach and Allen 2010) – is to supplement Arkin's top-down approach, involving rules and perhaps a consequentialist algorithm, with a bottom-up approach in which we design autonomous systems as moral learners, growing in them a nuanced and plastic moral capacity in the form of habits of moral response, in much the same way that we mature our children as moral agents.[44] There is considerable debate about this approach via moral learning. Arkin, himself, objects that neural nets and learning algorithms "black box" the developed competence in such a way as to make impossible both the robot's reconstructing for us either a decision tree or a moral justification of its choices, which he regards as a minimum necessary condition on moral machines, and the operator's reliably predicting the robot's behavior.[45] We respond that human moral agents are also somewhat unpredictable and that what they produce, when pressed for a justification of their actions, are after-the-fact rationalizations of moral choices. Why should we demand more of moral robots? How to produce after-the-fact rationalizations is an interesting technical question, one currently being vigorously and successfully investigated under such headings as "rule extraction," "interpretable AI," and "explainable AI."[46]

---

[44] Ioan Mutean & Don Howard, *Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency,* PHILOSOPHY AND COMPUTING (Thomas Powers, ed. Cham, Switzerland: Springer, 2017) at 121-159.
[45] Arkins, *supra* note 36 at 67, 108.
[46] Wojciech Samek, et al., eds. (2019). *Explainable AI: Interpreting, Explaining, and Visualizing Deep Learning. Cham, Switzerland: Springer.*

Others object that there is no consensus on what morality to program into our robots, whether through learning or rule sets. We respond that moral diversity among robots should be prized in the same way that we prize human moral diversity. We learn from one another because of our moral differences. But, at the same time, in the constrained space of autonomous weapons, there is consensus in the form of the international support for extant international law and the just war moral theory upon which it is based. Saudi Arabian health care robots might rightly evince different habits with respect to touching and viewing unveiled bodies from those evinced by North American or European health care robots. But Saudi Arabia has ratified the main principles of the Geneva Conventions just as has the United States.

Earlier, we touched directly or indirectly upon other potential moral gains from autonomous weapons, such as facilitating military intervention to prevent genocide or other human rights abuses, minimizing risk of death or injury to our own troops, and sparing drone operators and other personnel both psychological damage and moral corrosion from direct participation in combat. One can imagine still more, such as employing weaponized autonomous escort vehicles to protect aid convoys in conflict zones. The conclusion is that there are, in fact, noteworthy potential moral gains from the development and deployment of both offensive and defensive autonomous weapons. Of course this must be done in such a way as to insure compliance with all existing international law and in a manner that minimizes the likelihood of the technology's being put to the wrong uses by bad actors. Short of a ban on autonomous weapons, how do we do that?

III. AN ARTICLE 36 REGULATORY REGIME

The goal is regulating the development and deployment of autonomous weapons in a way that ensures compliance with international law and minimizes the chance of misuse. Moreover, we need to do this in a politically feasible way, using regulatory structures that will be accepted by the international community. This last point is important, because one common criticism of the proposed ban on autonomous weapons is, precisely, that it stands little chance of ever being incorporated in international law.

Even in the talks under the aegis of the UN's Convention on Certain Conventional Weapons (CCW) that have being going on since 2014 in Geneva, it is mainly only nations with little or no prospect of becoming significant participants in the development and use of autonomous weapons that have shown support for moving forward with consideration of a ban. The major players, including the United States, have repeatedly indicated that they will not support a ban. In December of 2021, the United States representative in Geneva, Josh Dorosin, said it again, while adding that a non-binding, international code of conduct might be appropriate.[47] That sufficiently strong support for a ban was unlikely ever to emerge from the Geneva talks was already clearly sensed six years ago by the most energetic proponents of the ban. Thus, in a 2016 press release, the Stop Killer Robots campaign subtly shifted the discourse, hinting at a tactical retreat, by urging a focus on "meaningful human control" (whatever that might mean), though talk of a ban still dominates the headlines.[48] If the goal is regulating the development and use of autonomous weapons in a politically feasible way, then seven years of talks have been wasted by the continued insistence on a ban.

What could the international community having been discussing instead? The discussion should have focused on what might be done within the compass of extant international law. There is already in place since 1977 Article 36 of Protocol I to the Geneva Conventions, which stipulates:

> "In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."[49]

---

[47] John Bowden, *Biden Administration Won't Back Ban on 'Killer Robots' Used in War*, THE INDEPENDENT. (December 8, 2021), https://www.independent.co.uk/news/world/americas/us-politics/biden-killer-war-robots-ban-b1972343.html.

[48] Clare Conboy, *Focus on Meaningful Human Control of Weapons Systems – Third United Nations Meeting on Killer Robots Opens in Geneva*, STOP KILLER ROBOTS (April 11, 2016), https://www.stopkillerrobots.org/news/press-release-focus-on-meaningful-human-control-of-weapons-systems-third-united-nations-meeting-on-killer-robots-opens-in-geneva/.

[49] *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I),* INTERNATIONAL COMMITTEE OF THE RED CROSS (Jun. 8, 1977), https://ihl-

174 states have ratified Protocol I, including Article 36, and three states, Pakistan, Iran, and the United States, are signatories but have not formally ratified the Protocol.[50] But the United States has promised to abide by nearly all provisions, including Article 36, and has established rules and procedures in all three branches of the military for insuring legal review of new weapons systems.[51] The countries having ratified Protocol I include every other major nation, among them China, the Russian Federation, and all NATO member states. I would argue that, since Article 36 is already a widely accepted part of international law, it is the best foundation upon which to construct a regulatory regime for autonomous weapons.

Concerns have been expressed about the effectiveness of Article 36 in general, chief among them being that the prescribed legal reviews are sometimes perfunctory and that it is too easy to evade an Article 36 review by declaring that a weapon is not new but just a minor modification of an existing and already authorized weapon. Those are serious worries, as evidenced by the recent controversy over whether the US's redesign of the B61 nuclear warhead with a tail assembly that makes possible limited, real-time steering of the warhead, the configuration designated now as B61-12, constituted a new weapon, as critics allege, or merely a modification, as the US asserts.[52] Another worry is that only a small number of states have certified that they are regularly carrying out Article 36 reviews. Equally serious are concerns that have been expressed about the effectiveness of Article 36 specifically with respect to autonomous weapons, as in a briefing report for delegates to the 2016 meeting of experts, which argued that what is at issue with autonomous weapons is not so much the conformity of individual weapons systems with international law, but the wholesale transformation of the nature of warfare wrought by the "unprecedented shift in human control over the

---

databases.icrc.org/applic/ihl/ihl.nsf/Treaty.xsp?action=openDocument&documentId=D9E6B6264D7723C3C12563CD002D6CE4/.

[50] *Id.*

[51] ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, INTERNATIONAL REVIEW OF THE RED CROSS (2006) at 88, 931-956; *see also* U.S. Army, *Legal Review of Weapons and Weapon Systems." Army Regulation 27–53*, DEPARTMENT OF THE ARMY (Washington, DC: Headquarters), (Sept. 23, 2019), https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/ARN8435_AR27-53_Final_Web.pdf.

[52] Adam Mount, *The Case against New Nuclear Weapons*, CENTER FOR AMERICAN PROGRESS (May 4, 2017), https://www.americanprogress.org/article/case-new-nuclear-weapons/.

use of force" that autonomous weapons represent. The magnitude of that change was said to require not individual state review but the engagement of the entire international community.[53] All such concerns would have to be addressed explicitly in the construction of an autonomous weapons regulatory regime based on Article 36.

How would a new Article 36 regulatory regime be constructed? Most important would be the development of a set of clear specifications of what would constitute compliance with relevant international law. This could be the charge to a Group of Governmental Experts under the auspices of the CCW.

First in importance among such guidelines would be a detailed articulation of what capabilities an autonomous weapon must possess for handling the problem of discrimination, bearing in mind the point made repeatedly above that this is not an all-or-nothing capability, but, rather, one specific to the functions and potential uses of an individual weapons system. Thus, as discussed above, for use within its intended missions, the Brimstone missile need only the capability to distinguish different categories of vehicles within its designated field of fire. An autonomous check-point sentry, by contrast, would have to be capable of much more sophisticated discriminations. Similarly detailed specifications would have to be developed for determinations of proportionality, recognition of a human combatant's having been rendered hors de combat, recognition of a target's displaying insignia, such as the Red Cross or Red Crescent, that identify a structure, vehicle, or individual as protected medical personnel, and so forth.

Just as important as developing the specifications would be the development of protocols for testing to insure compliance. Optimal, but politically unachievable, for obvious reasons, would be the open sharing of all relevant design specifications. It is highly unlikely that states and manufacturers are going to let the world community look under the hood at such things as new sensor technologies and accompanying software. The alternative is demonstrations of performance capability in realistic testing scenarios. We already have considerable relevant experience and expertise in safety and effectiveness testing for a wide range of engineered systems, especially pertinent being the testing protocols for

---

[53] CCW, *Article 36 Reviews and Addressing Lethal Autonomous Weapons Systems*, Briefing Paper for Delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, at 11-15 (April 2016), http://www.article36.org/wp-content/uploads/2016/04/LAWS-and-A36.pdf.

certifying control systems in commercial aircraft and industrial systems. One might think that weapons developers would be just as shy about showing off the weapon at work in realistic scenarios, lest adversaries and competitors infer confidential capabilities and technologies. But, in fact, most weapons developers are proud to show off videos of their new systems' doing impressive things and to display and demonstrate their products at international weapons expositions. What would be required would not be the sharing of secrets but simply demonstrations of reliability in complying with the detailed guidelines just discussed.

As with the existing Article 36 requirements, certification of compliance will surely have to be left to individual states. But it is not unreasonable to begin an international conversation about a more public system for declaring that the required certifications have been carried out, even if that consists in little more than asking signatories and states parties to file such certifications with the UN, ICRC, or another designated international entity.

The good news is that, within just the last few years, serious discussion of precisely such concrete elaborations of Article 36 protocols for autonomous weapons has begun to appear in the scholarly, policy, and legal literatures.[54] Equally encouraging is the willingness of some governments to underwrite such work. Thus, the German Auswärtiges Amt (Foreign Office) subsidized a 2015 expert seminar under the auspices of the Stockholm International Peace Research Institute (SIPRI) that had representation from France, Germany, Sweden, Switzerland, the United Kingdom and the United States (Boulanin 2015).[55]

What have been the fruits of such work? Many good ideas have emerged. Especially thoughtful are the main recommendations contained in a 2017 report that was also sponsored by SIPRI covering Article 36 elaborations for cyber weapons, autonomous weapons, and soldier enhancement. Their approach was to focus on advice to reviewing

---

[54] Ryan Poitras, *Article 36 Weapons Review & Autonomous Weapons Systems: Supporting an International Review Standard*, AMERICAN UNIVERSITY INTERNATIONAL LAW REVIEW 34, at 465-495; *see* Cochrane, Jared M. (2020). "Conducting Article 36 Legal Reviews for Lethal Autonomous Weapons." *Journal of Science Policy & Governance* 16;1 (April 2020). https://www.sciencepolicyjournal.org/uploads/5/4/3/4/5434385/cochrane_jspg_v16.pdf.

[55] Vincent Boulanin, *Implementing Article 36 Weapon Reviews in the Light of Increasing Autonomy in Weapon Systems*, STOCKHOLM INT'L PEACE RESEARCH INST. (Nov. 2015), https://www.sipri.org/sites/default/files/files/insight/SIPRIInsight1501.pdf.

authorities in individual member states, and they emphasize two broad categories of advice: (1) Building on best practices already being employed by states that have well-developed review procedures. (2) Strengthening transparency and cooperation among states. Under the first heading, they advise, for example:

1. Start the review process as early as possible and incorporate it into the procurement process at key decision points.
2. Provide military lawyers involved in the review process with additional technical training. Engineers and systems developers should also be informed about the requirements of international law so that they can factor these into the design of the weapons and means of warfare.[56]

About increased transparency and cooperation they say that it would become a "virtuous circle," and they observe that:

1. It would allow states that conduct reviews to publicly demonstrate their commitment to legal compliance.
2. It would be of assistance to states that are seeking to set up and improve their weapon review mechanisms and thereby create the conditions for more widespread and robust compliance.
3. It could facilitate the identification of elements of best practice and interpretative points of guidance for the implementation of legal reviews, which would strengthen international confidence in such mechanisms.

They add:

> Cooperation is also an effective way to address some of the outstanding conceptual and technical issues raised by emerging technologies. Dialogues, expert meetings and conferences can allow generic issues to be debated and addressed in a manner that does not threaten the national security of any state.[57]

---

[56] Vincent Boulanin & Maaike Verbruggen, *Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies*, STOCKHOLM INT'L PEACE RESEARCH INST., viii (Stockholm, Sweden) (2017).
[57] *Id.*

When it comes specifically to Article 36 reviews involving autonomous weapons, they identify as the foremost challenge verifying "the predictability of autonomous weapon systems' compliance with international law".[58]

I am not at all naive about how strict compliance with Article 36 requirements would be. But existing Article 36 requirements have already created a culture of expectations about compliance and a space within which states can and have been challenged, sometimes successfully, to offer proof of compliance, as with the widely expressed concerns about truly indiscriminate weapons, such as land mines and cluster munitions. We begin to norm such a space simply by putting the relevant norms in front of the world community and initiating a public conversation about compliance. This is what we should be talking about in Geneva if we are serious about building some measure of international control over autonomous weapons.

CONCLUSION

War is hell. It will always be an inherently immoral form of human activity. The goal of international law is to minimize the otherwise inevitable death and suffering that war entails. Advances in technology can contribute toward that goal by making weapons more accurate, less lethal, and more selective. The advent of autonomous weapons promises still further moral gains by removing the single most common cause of war crimes, the too often morally incapacitated human combatant. We cannot let unrealistic fears about a Terminator-AI apocalypse prevent our taking advantage of the opportunities for moral progress that properly designed and deployed autonomous weapons afford. We must, of course, ensure that such systems are being used for good, rather than malign purposes, as we must with any technology, and especially technologies of war. Indeed, with autonomous weapons we need to be more vigilant, still. But minimizing death and suffering in war is the ultimate goal. If autonomous weapons can contribute to progress toward that goal, then we must find a way to license their use in full compliance with what law and morality demand.

---

[58] *Id.* at xi.

ACKNOWLEDGMENTS

This paper was much improved thanks to helpful feedback from discussants at conferences and colloquia at the University of Zurich in November of 2015, Purdue University in December of 2015, and the University of Minnesota in April of 2016 where an early version was presented. Special thanks also go to my University of Notre Dame Ph.D. student, Kevin Schieman, for very constructive, critical feedback.

VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

ARTICLES

ETHICAL AI IN AMERICAN POLICING

Elizabeth E. Joh

ETHICAL AI IN AMERICAN POLICING

*Elizabeth E. Joh**

INTRODUCTION

We know there are problems in the use of artificial intelligence in policing, but we don't quite know what to do about them.[1] Artificial intelligence (AI) systems[2] are becoming conventional and widespread in routine policing. License plate reader systems routinely scan thousands of plates per minute.[3] At least 117 million Americans are included in databases where facial recognition searched are conducted.[4] Predictive algorithms try to forecast future places or persons warranting law enforcement attention.[5] Autonomous drones can follow a suspect or record activity with the push of a button.[6] Increasingly the issue is not whether, but under what circumstances, these tools will be used.

---

*Professor of Law, U.C. Davis School of Law. Thanks to the editorial staff of the Notre Dame Journal on Emerging Technologies for their editorial work, and to the inter-journal collaboration at Notre Dame Law School for organizing the Race & the Law: Interdisciplinary Perspectives symposium.

[1] This isn't just an American problem. The director of the UK Police Foundation stated in January 2022 that "national guidance on ethical considerations [for emerging technologies] would be especially welcome." *See* GLORIA GONZÁLEZ FUSTER, EUROPEAN PARLIAMENT POL'Y DEP'T FOR CITIZEN'S RTS. & CONST. AFFS., ARTIFICIAL INTELLIGENCE AND LAW ENFORCEMENT: IMPACT ON FUNDAMENTAL RIGHTS (2020) [hereinafter IMPACT ON FUNDAMENTAL RIGHTS] ("The magnitude and seriousness of challenges triggered by AI in the field of law enforcement and criminal justice . . . do not appear to be conveniently addressed by ongoing reflections."); Claudia Glover, *Policing Minister Rejects Need for Ethical Guidance on Emerging Tech*, TECH MONITOR (Jan. 13, 2022), https://techmonitor.ai/policy/regulating-use-of-technology-in-uk-police.

[2] By using the terns "AI applications" or "AI systems," I refer to the application of algorithms and substantial amounts of computing power to enormous amounts of digitized data.

[3] *See, e.g.*, Ángel Díaz & Rachel Levinson-Waldman, *Automatic License Plate Readers: Legal Status and Policy Recommendations for Law Enforcement Use*, BRENNAN CTR. (Sept. 10, 2020), https://www.brennancenter.org/our-work/research-reports/automatic-license-plate-readers-legal-status-and-policy-recommendations (noting "93 percent of police departments in cities with populations of 1 million or more use their own ALPR systems, some of which can scan nearly 2,000 license plates per minute").

[4] Clare Garvey et al., *The Perpetual Line-Up: Unregulated Police Face Recognition in America*, GEORGETOWN L. CTR. ON PRIV. AND TECH. (Oct. 18, 2016) [hereinafter *Perpetual Line-Up*], https://www.perpetuallineup.org/ (noting that at least 26 states allow the police to run face recognition searches against driver's license and ID photos).

[5] *See infra* Part I.

[6] *See infra* Part I.

With artificial intelligence, the police can perform their traditional functions not just on a faster and larger scale, but in novel ways that have prompted strong criticism. Some of these issues are familiar to a legal audience. If the police can track everywhere you've been in public, what does that mean for the usual lack of constitutional protections in public spaces? If the police can easily identify every face in a public protest, how does that dampen free speech rights? Other voices in this backlash have arisen out of what has been called the algorithmic accountability movement: scholars and activists who have focused on the harms posed by the particulars of the technologies themselves.[7] For instance, the now quite well-documented issue of racial and gender bias in many facial recognition technology programs means that the costs of mistaken matches are borne disproportionately by people of color and women.[8] At the same time, law enforcement officials have embraced these technologies as promising innovations. Automation both in and around policing is growing, with few signs of slowing down.

One can also find many reports and white papers today offering principles for the responsible use of AI systems by governments, civil society organizations, and the private sector. Increasingly common too are calls for the fair use of artificial intelligence across fields like housing, employment, consumer credit, and criminal justice. This comes at a time when automated decision-making might determine whether you'll be hired,[9] whether you'll be fired,[10] whether you'll receive one medical

---

[7] We can also include here the development of the field of Fairness, Accountability, and Transparency in Machine Learning. See, e.g., *Fairness, Accountability, and Transparency in Machine Learning*, FATML, http://fatml.org (last visited Feb. 26, 2022).

[8] Researcher Joy Buolamwini was among the first to identify the issue of bias. Steve Lohr, *Facial Recognition is Accurate, if You're a White Guy*, N.Y. TIMES (Feb. 9, 2018), https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html (citing Buolamwini's work finding up to 35% error rate for darker skinned women compared to 1 percent error rate for white men). The National Institute of Standards and Technology similarly found in 2019 that the facial recognition programs it studied mistakenly identified people of color far more often than white people. *See NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*, NIST (Dec. 19, 2019), https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software (evaluating 189 software algorithms and finding that "for one-to-one matching, the team saw higher rates of false positives for Asian and African-American faces relative to images of Caucasians.").

[9] Rebecca Heilweil, *Artificial Intelligence Will Help Determine if You Get Your Next Job*, VOX (Dec. 12, 2019), https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen ("recruiters are increasingly using AI to make the first round of cuts and to determine whether a job posting is even advertised to you.").

[10] Spencer Soper, *Fired by Bot at Amazon: 'It's You Against the Machine,'* BLOOMBERG (June 28, 2021), https://www.bloomberg.com/news/features/2021-06-28/fired-by-

treatment over another, or whether you'll be granted bail. In 2021, Congress established a National AI Advisory Committee, tasked with providing recommendations about the use of AI and its impact on society.[11] The White House Office of Science and Technology Policy plans to publish an Algorithmic "Bill of Rights."[12] The European Union is preparing to adopt a comprehensive regulatory framework for the use of AI in 2022.[13]

Yet, largely missing from the current debate in the United States is a shared framework for thinking about the ethical and responsible use of AI that is specific to policing.[14] Leading an average-sized law enforcement agency in the United States in the 2020s means responding to very different pressures: to reduce crime, to address bias and

---

bot-amazon-turns-to-machine-managers-and-workers-are-losing-out ("Increasingly, the company is ceding its human-resources operation to machines as well, using software not only to manage workers in its warehouses but to oversee contract drivers, independent delivery companies and even the performance of its office workers.").

[11] The Committee is one of several governance bodies created by the National Artificial Intelligence Initiative Act of 2020. *See National Artificial Intelligence Advisory Committee (NAIAC)*, NIST (Oct. 27, 2021), https://www.nist.gov/artificial-intelligence/national-artificial-intelligence-advisory-committee-naiac.

[12] Eric Lander & Alondra Nelson, *Americans Need a Bill of Rights for an AI-Powered World*, WIRED (Oct. 8, 2021), https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/ ("In the coming months, the White House Office of Science and Technology Policy (which we lead) will be developing such a bill of rights, working with partners and experts across the federal government, in academia, civil society, the private sector, and communities all over the country.").

[13] *2021 Artificial Intelligence and Automated Systems Annual Legal Review*, GIBSON DUNN (Jan. 20, 2022), https://www.gibsondunn.com/wp-content/uploads/2022/01/2021-artificial-intelligence-and-automated-systems-annual-legal-review.pdf ("With the new Artificial Intelligence Act, which is expected to be finalized in 2022, it is likely that high-risk AI systems will be explicitly and comprehensively
regulated in the EU."). The proposed EU regulations focus on "harmonised rules for the development, placement on the market and use of AI system in the Union following a proportionate risk-based approach." *See Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, at 3, COM (2021) 206 final (Apr. 21, 2021).

[14] The same is true elsewhere, as observed by the U.K. government's Centre for Data Ethics & Innovation. *See* CTR. FOR DATA ETHICS & INNOVATION, REVIEW INTO BIAS IN ALGORITHMIC DECISION-MAKING 7 (2020) [hereinafter CDEI] ("Though there is strong momentum in data ethics in policing at a national level, the picture is fragmented with multiple governance and regulatory actors, and no single body fully empowered or resourced to take ownership."). The CDEI is a "government expert body enabling the trustworthy use of data and AI." *See About Us*, CDEI, https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation/about#:~:text=Overview-,The%20CDEI%20is%20a%20government%20expert%20body%20enabling,use%20of%20data%20and%20AI.&text=The%20CDEI%20is%20committed%20to,core%20component%20of%20its%20work (last visited Feb. 27, 2022).

discrimination, to cut costs, and to innovate. In this context, AI systems offer tools that promise faster and more efficient methods of investigation and police administration. But their adoption into police decision-making and tactics also introduces complications. Any police department interested in guidelines for ethical use of AI systems would "find a field with few existing examples and no established guidelines or best practices."[15]

Commitments to ethical and responsible principles in the police use of AI have a role here. They aren't substitutes for regulation or judicial decision-making. However, legislators and judges have been slow. The United States lacks a national, comprehensive approach to the regulation of AI systems.[16] Instead, state and local governments have been left to decide whether and how to regulate AI systems either based on a particular industry or on specific use cases. Similarly, there have been a small number of cases challenging the use of AI systems in the courts, but not enough to conclude that a body of rules have been developed.[17] This means that policing in particular is guided by an uncertain set of rules and legal decisions for the adoption and use of AI-based systems. And while ethical and legal principles share common concerns, ethical principles broaden the set of possible questions police departments should consider.[18]

Many AI policy guidance documents exist now, but their value to the police is limited. Simply repeating broad principles about the responsible use of AI systems are less helpful than ones that 1) take into account the specific context of policing, and 2) consider the American experience of policing in particular. There is an emerging consensus

---

[15] *See Use of New Artificial Intelligence Technologies Policy – Public Consultation*, TORONTO POLICE SERV. BD. (2022) [hereinafter Toronto Police Services Board], https://tpsb.ca/ai.

[16] *See* Heather Sussman et al., *U.S. Artificial Intelligence Regulation Takes Shape*, ORRICK (Nov. 18, 2021), https://www.orrick.com/en/Insights/2021/11/US-Artificial-Intelligence-Regulation-Takes-Shape (contrasting developments in EU while noting "there is currently no federal regulation of AI in the U.S.").

[17] Jessica Field et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, (2020) [hereinafter Berkman Klein Report], https://cyber.harvard.edu/publication/2020/principled-ai ("Litigation over the harmful consequences of AI technology is still nascent, with just a handful of cases having been brought. Similarly, only a few jurisdictions have adopted regulations concerning AI . . .").

[18] *Cf.* Lexo Zardiashvili et al., *AI Ethics for Law Enforcement: A Study into Requirements for Responsible Use of AI at the Dutch Police*, 2 DELPHI 1, 2 (2019) (arguing that "for such spaces left open by the law, the police can, and we advise that they should incorporate 'ethics' through practical measures to ensure responsible use of AI and contribute toward enhancing (rather than limiting) legitimacy of and trust in the police.").

about what ethical and responsible values should be part of AI systems. This essay considers what kind of ethical considerations can guide the use of AI systems by American police.

I. AI SYSTEMS AND THEIR USE IN POLICING

Anyone taking a first look at the use of AI systems in policing would be justifiably confused. New tools are alternatively described as data-driven, based on artificial intelligence, powered by algorithms, or new surveillance technologies. Are these terms meaningfully different? We can begin by looking at what we mean by an AI system, and how police are using these tools.

First, there is no single widely accepted definition of artificial intelligence.[19] But many policy documents from around the world define AI in terms of software that can achieve a complex goal by acting upon collected information and then processing or interpreting that data. Sometimes an AI system will adapt its behavior by analyzing the environment changed by its previous actions.[20] This use of algorithms, combined with cheap and powerful computer processing, and massive amounts of data has also sometimes been referred to as the use of "big data."[21]

To add to these ambiguities, some discussions of AI systems in policing might also use the term "data-driven" policing: a term that captures both AI systems today and earlier efforts dating back to the 1990s that simply emphasize the increasing reliance of police decision-making on statistics.[22] Finally, discussions of AI systems in policing like

---

[19] The term "artificial intelligence" was first coined by John McCarthy in 1955, who defined it as "the science and engineering of making intelligent machines," but that definition is just one among many today. *See* PETER STONE ET AL., STAN. UNIV., ARTIFICIAL INTELLIGENCE AND LIFE IN 2030: REPORT OF THE 2015 STUDY PANEL 50 (2016) ("McCarthy is credited with the first use of the term "artificial intelligence" in the proposal he co-authored for the workshop with Marvin Minsky, Nathaniel Rochester, and Claude Shannon."); *see also* Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 404 (2017) ("There is no straightforward, consensus definition of artificial intelligence.").

[20] This particular definition is derived from the European Commission's High-Level Expert Group on Artificial Intelligence, but many similar ones exist. *See, e.g.*, Berkman Klein Report, *supra* note 17, at 4-5.

[21] *See, e.g.*, IMPACT ON FUNDAMENTAL RIGHTS, *supra* note 1, at 21 (defining AI as including "data, algorithms, and computer power" and acknowledging overlap with "big data").

[22] *See, e.g.*, Annie Gilbertson, *Data-Informed Predictive Policing was Heralded as Less Biased. Is It?*, THE MARKUP (Aug. 20, 2020), https://themarkup.org/ask-the-markup/2020/08/20/does-predictive-police-technology-contribute-to-bias ("Early

predictive policing or facial recognition software sometimes focus on their increased surveillance capacity and are described as new surveillance technologies.[23] All of these terms are sometimes used interchangeably. For simplicity's sake, we can use the term "AI systems" here. All of these technologies introduce some degree of automated decision-making into what had traditionally been the entirely human process of police work.

In theory, AI systems can introduce efficiency and innovation to a field that is as much about the management of risk and the processing of information as it is about stops and arrests. Identifying patterns in large sets of data can help the police prioritize where their officers and dollars go.[24]

In policing, the AI systems that have received the most attention are probably facial recognition software and predictive policing. Predictive policing software can take a variety of forms, but at their most basic they rely on past information to make forecasts about the future: whether crimes are likely to occur in particular places, or whether people are likely to engage in some kinds of crimes or become victims of crime.[25] In 2011, the police department in Santa Cruz, California became one of the first in the United States to pilot a predictive policing program, one developed by the private company PredPol (now Geolitica).[26] That program assessed historical crime data and directed its client, the Santa Cruz police, to those five hundred square foot areas where crime was

---

versions of data-driven policing were used in the 1990s, but it has grown more popular and the technology more sophisticated over the last decade.").

[23] *See, e.g.*, Andrew G. Ferguson, *Surveillance and the Tyrant Test*, 110 GEO. L. J. 205, 210 (2021) (characterizing tools like facial recognition and license plate readers as "new surveillance technologies").

[24] *See, e.g.*, CDEI, *supra* note 14, at 64 ("In theory, tools which help spot patterns of activity and potential crime, should lead to more effective prioritization and allocation of scarce police resources.").

[25] *See, e.g.*, IMPACT ON FUNDAMENTAL RIGHTS, *supra* note 1, at 22 (defining predictive policing as "the algorithmic processing of data sets . . . to reveal patterns of probable future offending and victimization, which can thus be interdicted before they happen"). Examples of predictive software about persons include Chicago's "Strategic Subjects List," which identified persons at high risk of being involved in future gun violence as perpetrators or victims. *See, e.g.*, Mick Dumke & Frank Main, *A Look Inside the Watch List Chicago Police Fought to Keep Secret*, CHI. SUN TIMES (May 18, 2017), https://chicago.suntimes.com/2017/5/18/18386116/a-look-inside-the-watch-list-chicago-police-fought-to-keep-secret (describing risk assessment that scored individuals and listed 398,000 entries in 2017). Another example is the UK Metropolitan Police's use of the Gangs Violence Matrix, a tool to identify those at risk of gang violence as perpetrators or victims. *See* IMPACT ON FUNDAMENTAL RIGHTS, *supra* note 1, at 24.

[26] *See* Erica Goode, *Sending the Police Before There's a Crime*, N.Y. TIMES (Aug. 15, 2011), https://www.nytimes.com/2011/08/16/us/16police.html (describing Santa Cruz's "unusual experiment" to test a prediction method for property crimes).

likely to occur.[27] Dozens of police departments piloted and adopted similar programs in the following years.[28]

Like predictive policing, facial recognition technology is a broad term. The technology uses an algorithm to see if one image can be matched against another in an existing database of images. To deliver results, a facial recognition program must collect images, classify them, train that data, and test these training sets.[29] These comparisons can be used in many ways. For instance, face verification confirms your identity against a stored image.[30] Face identification involves matching a suspect's face to a database of existing images, like a driver's license records.[31] Or, the technology might be used for generalized surveillance, to identify many people in places like airports or public streets.[32]

Predictive policing and facial recognition have received the most public attention in policing, and for good reason. Predictive policing threatens to replace the seemingly unique skill of human police expertise. The assessments of suspicious persons and places by police officers poses its own problems, of course, but turning over some of this decision-making to machines preys on people's suspicions about how trustworthy these assessments are.[33] And the potential of facial recognition to

---

[27] *See id.*

[28] The Electronic Frontier Foundation's Atlas of Surveillance has identified at least 160 agencies using predictive policing as of January 2022. *See Atlas of Surveillance*, ELEC. FRONTIER FOUND. (Jan. 11, 2022), https://atlasofsurveillance.org/atlas.

[29] *See* Andrew Ferguson, Facial Recognition and the Fourth Amendment, 105 MINN. L. REV. 1105, 1112 (2021). Face recognition algorithms learn to identify important facial features by being trained through the comparison of data. An algorithm might be given pairs of face images of the same person; over time, it recognizes that some features act as reliable identifying signals about the same person. *See* CLARE GARVIE, ALVARO M. BEDOYA & JONATHAN FRANKLE, GEORGETOWN L. CTR. ON PRIV.&TECH.,THE PERPETUAL LINE-UP:UNREGULATED POLICE FACE RECOGNITION IN AMERICA 1 (2016), https://www.perpetuallineup.org/sites/default/files/2016-12/The%20 Perpetual%20Line-Up%20-%20Center%20on%20Privacy%20and%20Technology% 20at%20Georgetown%20Law%20-%20121616.pdf.

[30] *See id.* at 109.

[31] *See id.* at 108 (discussing this as "face identification").

[32] *See id.* (discussing this as "face surveillance").

[33] For example, research from DeepMind and the U.K.'s RSA found that sixty percent of survey respondents opposed or strongly opposed the use of automated decision-making in the criminal justice system and the workplace. *See*, BRHMIE BALARAM ET AL., ROYAL SOCIETY FOR THE ENCOURAGEMENT OF ARTS, MANUFACTURERS, AND COMMERCE, ARTIFICIAL INTELLIGENCE: REAL PUBLIC ENGAGEMENT 4 (2018). Similarly, a 2018 Pew Research report found that majorities of Americans surveyed found it "unacceptable" for algorithms to make decisions with "real-world consequences for humans," including criminal risk assessments for people considered for parole. *See Public Attitudes Toward Computer Algorithms*, PEW RSCH. CTR. (Nov. 16, 2018), https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/.

identify hundreds, even thousands of people in moments sparks concerns about unchecked surveillance power.[34]

But AI systems have other roles in policing as well. While many police departments had been using remote-controlled drones, newer versions are more similar to autonomous cars than radio-controlled toy cars.[35] An autonomous police drone can respond quickly to a 911 call and provide the police with details as they assemble their human response.[36] A police drone can also fly into enclosed spaces for surveillance where the police are concerned about unknown threats.[37] Similarly, the inevitable introduction of autonomous cars will mean not just autonomous police cars, but also the possibility of remote stops of cars by the police.[38]

Other AI systems can address police issues that are important but don't generate the same public concern. Most of us don't focus on the administrative parts of policing, but police officers devote enormous amounts of time to pushing paper and filling out forms.[39] The paperwork associated with arrests, for instance, takes up so much time that it can provide a perverse incentive for some officers to use arrests as an excuse for overtime pay.[40] AI systems can make these processes less cumbersome by automating form-filling and aggregating information. Companies like Axon Enterprise and Mark43 offer cloud-based records based management (RMS) systems that automate some of the report-

---

[34] *See, e.g.*, Laura K. Donohue, *Technological Leap, Statutory Gap, and Constitutional Abyss: Remote Biometric Identification Comes of Age*, 97 MINN. L. REV. 407, 415 (2012) (arguing that remotely used biometric technologies like face recognition are "significantly different from that which the government has held at any point in U.S. history").
[35] Cade Metz, *Police Drones Are Starting to Think for Themselves*, N.Y. TIMES (Dec. 5, 2020), https://www.nytimes.com/2020/12/05/technology/police-drones.html.
[36] *See id.*
[37] *See id.*
[38] *See, e.g.*, Elizabeth E. Joh, *Automated Seizures: Police Stops of Self-Driving Cars*, 94 N.Y.U. L. REV. (ONLINE ISSUE) 113 (2019).
[39] *See, e.g.*, Brad W. Smith et al., *Community Policing and the Work Routines of Street-Level Officers*, 26 CRIM. JUST. REV. 17, 31 (2001) (reporting research that "administrative activities consumed a significant portion [of an officer's daily shift].").
[40] *See, e.g.*, EDITH LINN, ARREST DECISIONS: WHAT WORKS FOR THE OFFICER? 1 (2009)(finding that the overtime pay associated with arrest procedures influences police officer behavior).

taking process.[41] Axon's CEO even envisions an entirely automated information flow one day from body camera video to police report.[42]

In sum, AI systems are already a part of ordinary police work. Public attention tends to focus on a few applications that are controversial because they raise the specter of vastly increased police power with new risks and few checks. But AI systems also assume other tasks in policing, including through some seemingly mundane tasks that are nevertheless central to what police do: processing information to investigate crime.

II. THE SECOND WAVE OF AI SYSTEMS IN POLICING

Today AI systems in policing find a very different audience from the one that endorsed predictive policing as one of the fifty "best inventions of the year" in 2011.[43] If the 2010s can be characterized as an enthusiastic embrace of novel police technologies, the 2020s could be deemed a second wave of AI-based systems in policing.[44] It is a second wave not only because there is much more use of AI everywhere, but also because the social and political context has changed as well. Civil rights organizations, policymakers, and scholars have pointed out the shortcomings of those AI systems already in place. And the harms of AI systems in policing are no longer theoretical. People have been mistakenly stopped and arrested because of mistaken AI

---

[41] *See, e.g.*, Thad Rueter, *Mark43 Raises $101M to Expand Police Tech Products*, GOVTECH BIZ (July 12, 2021), https://www.govtech.com/biz/mark43-raises-101m-to-expand-police-tech-products (citing evidence for "increased spending for [cloud based records management] even amid pandemic spending cuts and the broad 'defund the police' movement in the U.S. that calls for government to shift some of law enforcement's responsibilities to other agencies"); Peter Hall, *New Record Keeping Software Will Make It Easier for Lehigh County Police Departments to Share Information*, MORNING CALL (Sept. 13, 2021), https://www.mcall.com/news/local/mc-nws-lehigh-county-police-record-software-upgrade-20210913-ziw73kxxorfsxokseg3w5bjbsy-story.html (describing new $3.6 million dollar three year contract with Mark43 which will provide cloud based report writing software including predictive language use).

[42] *See also* Dana Goodyear, *Can the Manufacturer of Tasers Provide the Answer to Police Abuse?*, THE NEW YORKER (Aug. 20, 2018), https://www.newyorker.com/magazine/2018/08/27/can-the-manufacturer-of-tasers-provide-the-answer-to-police-abuse.

[43] Lev Grossman et al., *The 50 Best Inventions*, TIME MAG. (Nov. 28, 2011), http://content.time.com/time/subscriber/article/0,33009,2099708,00.html.

[44] These terms are loosely based upon Frank Pasquale's description: "While the first wave of algorithmic accountability focuses on improving existing systems, a second wave of research has asked whether they should be used at all—and, if so, who gets to govern them." Frank Pasquale, *The Second Wave of Algorithmic Accountability*, LPE PROJECT (Nov. 25, 2019), https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/.

determinations.[45] And there are likely many, many people who received greater police scrutiny short of a physical encounter because an AI system flagged them for attention.

A. Critiques of AI Systems in Policing

There are now some well-established criticisms of the use of AI systems in policing. We can divide them broadly into three categories: bias, privacy, and secrecy. The data used in these systems may be biased.[46] The design of the systems may reflect the biases of the engineers who created them. These biases can in turn amplify biases against marginalized groups, or even create new forms of bias.[47]

As costs for data collection, storage, and analysis become ever cheaper, the police gain the ability to conduct indiscriminate mass surveillance. These capabilities can chill speech, the ability to freely associate with others, and to remain anonymous.[48] Each of these data points, whether collected directly by the police or by third parties like cellphone apps, may seem unworthy of privacy protection. But in the aggregate, they form the ability to create a time machine into our past movements, and sometimes our real-time movements as well.

Discovering how American law enforcement agencies use AI-based systems has been challenging because of their secrecy and opacity. One type of secrecy happens when some AI systems can make determinations about data in ways that even developers cannot completely explain.[49] This black box problem may have few consequences in some applications, like chatbots for recreation. But there are—and increasingly will be—many situations where people feel

---

[45] As of January 2022, there are at least three known cases where facial recognition technology provided a mistaken match. *See* Kashmir Hill, *Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match*, N.Y. TIMES (Dec. 29, 2021), https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

[46] CDEI *supra* note 14, at 68 ("Police data can be biased due to it either being unrepresentative of how crime is distributed or in more serious cases reflecting unlawful policing practices.").

[47] *Cf.* CDEI, *supra* note 14, at 21 ("There is clear evidence that algorithmic bias can occur, whether through entrenching previous human biases or introducing new ones.").

[48] *See Perpetual Line-Up*, *supra* note 4, at 41-44 ("Despite the fact that leading law enforcement agencies . . . have explicitly recognized the potential chilling effect of face recognition on free speech, we found that almost none of the agencies using face recognition have adopted express prohibitions against using the technology to track political or other First Amendment activity.").

[49] *See, e.g.*, Calo, *supra* note 19, at 414 (observing that deep learning AI systems "can say what will happen but not why").

the real impacts of such inscrutable decisions. They will be turned down for a loan, or stopped by the police. That is why "explainability" is a widely shared principle from AI-guidance proposals from around the world.[50]

Another type of secrecy in many AI systems, particularly in the United States, stems from companies making claims that disclosure will harm their intellectual property rights.[51] This means that trying to find out about the AI-based system—even one that directly impacted your life in some way—may be nearly impossible to find out. The company that developed it may claim that providing important information might divulge a trade secret.[52] A public agency that uses the AI system might also claim that it is bound by a non-disclosure agreement entered into with that same company.[53]

The response to these issues has been uneven. There is widespread agreement that the increasing use of AI systems needs guidance.[54] A survey of more than thirty documents stating AI principles from around the world identified several shared themes.[55] These included values important for policing: privacy,[56] accountability,[57]

---

[50] It's also true that the field of "explainable AI" (XAI) has not achieved consensus on how exactly this value can be implemented in practice. *See, e.g.*, Jessica Newman, *Explainability Won't Save AI*, BROOKINGS (May 19, 2021), https://www.brookings.edu/techstream/explainability-wont-save-ai/ (noting that "the XAI field has generally struggled to realize the goals of understandable, trustworthy, and controllable AI in practice.").

[51] *See generally* Elizabeth E. Joh, *The Undue Influence of Surveillance Technology Companies on Policing*, 92 N.Y.U. L. Rev. (ONLINE ISSUE) 101 (2017) [hereinafter *Undue Influence*]; Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018).

[52] *Undue Influence*, *supra* note 51, at 125-126 (discussing TrueAllelle's citing of trade secrets for non-disclosure).

[53] *Id.* at 104-08 (discussing use of non-disclosure agreements to shield details of cell site simulator technology).

[54] *See, e.g.*, Calo, *supra* note 19, at 411 ("Perhaps the most visible and developed area of AI policy to date
involves the capacity of algorithms or trained systems to reflect human values such as fairness, accountability, and transparency ("FAT")).

[55] Berkman Klein Report, *supra* note 17, at 4-5.

[56] "Privacy" is defined as referring to the idea that "AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it." *Id.* at 4.

[57] "Accountability" is defined as including mechanisms to ensure that those impacted by AI systems have appropriate remedies and that AI's effects are appropriately distributed. *Id.*

transparency and explainability,[58] fairness and non-discrimination,[59] human control of technology,[60] and the promotion of human values.[61] Most of us would agree that these are worthy goals.

B. Attempts to Regulate Police AI Systems

How these values have translated into practice is another matter. The United States has no national legislation on the use of AI-based systems, in any field. What has occurred in this absence is a patchwork of solutions. This section discusses some of the most prominent efforts to regulate AI in policing and their shortcomings.

First, there have been attempts to regulate the police use of surveillance technologies, of which AI-based systems are a part, by enacting local ordinances at the city or county level. In 2016, the ACLU launched an initiative to help local communities pass laws requiring oversight and transparency about the police use of new technologies.[62] Its Community Control Over Police Surveillance (CCOPs) campaign, supported by many civil rights groups,[63] published a model ordinance to serve as a template for local governments to follow.[64] Key features of the model act include the requirement of explicit approval for the purchase or use of new surveillance technologies,[65] the requirement of surveillance

---

[58] "Transparency" and "explainability" include the translation of "operations into intelligible outputs and the provision of information about where, when, and how they are being used." *Id.*

[59] "Fairness" and "non-discrimination" are defined as designing AI "to maximize fairness and [to] promote inclusivity." *Id.*

[60] "Human control of technology" refers to a requirement that "important decisions remain subject to human review." *Id.*

[61] "Promotion of human values" refers to the idea that "the ends to which AI is devote . . . should correspond with our core values and generally promote humanity's well-being." *Id.*

[62] *Community Control over Police Surveillance (CCOPS)*, ACLU, https://www.aclu.org/issues/privacy-technology/surveillance-technologies/community-control-over-police-surveillance?redirect=feature/community-control-over-police-surveillance (last visited Feb. 27, 2022).

[63] *See* Dave Maass, *Join the Movement for Community Control Over Police Surveillance*, ELEC. FRONTIER FOUND. (Sept. 21, 2016), https://www.eff.org/deeplinks/2016/09/join-movement-community-control-over-police-surveillance.

[64] *Community Control Over Police Surveillance (CCOPS) Model Bill*, ACLU (Apr. 2021) [hereinafter CCOPS Model Bill], https://www.aclu.org/legal-document/community-control-over-police-surveillance-ccops-model-bill.

[65] *Id.* at Section 1.

impact reports and other surveillance data,[66] and the creation of community advisory committees.[67]

In practice, this model has not found wide adoption. A 2020 study found that only fourteen local governments had passed local ordinances regulating police use of new surveillance technologies.[68] In other cities, proposals have been defeated or stalled. The reasons are varied, but this kind of intensive local oversight of police can be a difficult political project.[69] Their slow place and infrequent adoption thus far means that local administrative regulations are unlikely to provide significant constraints or guidance soon.

A second important development can also be found in cities around the country. In 2019, the Board of Supervisors voted to ban the use of facial recognition by its police and other public agencies.[70] In 2020, the city of Santa Cruz, California became the first American city to ban the use of predictive policing software.[71] A few dozen other local governments have followed their lead in considering bans or moratoria on the use of specific technologies, particularly facial recognition software.[72]

While civil liberties organizations have lauded these measures as successes, they have limits. On the one hand, bans are blunt tools with an intuitive appeal. They impose easy-to-understand total embargoes. But these bans are problematic. Technology-specific bans can simultaneously be both blunt but too narrow. They address only one specific system, such as facial recognition technology in body cameras, without addressing other AI-based systems that might pose similar

---

[66] *Id.* at Sections 6-7.
[67] *Id.* at Section 8.
[68] Mailyn Fidler, *Local Police Surveillance and the Administrative Fourth Amendment*, 36 SANTA CLARA HIGH TECH. L. J. 481, 545 (2020).
[69] Mailyn Fidler & Lily Liu, *Four Obstacles to Local Surveillance Ordnances*, LAWFARE (Sept. 4, 2020), https://www.lawfareblog.com/four-obstacles-local-surveillance-ordinances (identifying objections from politically strong mayors, police lobbying, an overemphasis on surveillance cameras, and concerns about public safety and overregulation as obstacles that stalled attempts at local oversight of police technologies).
[70] Kate Conger et al., *San Francisco Bans Facial Recognition Technology*, N.Y. TIMES (May 14, 2019), https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html.
[71] Kristi Sturgill, *Santa Cruz Becomes the First U.S. City to Ban Predictive Policing*, L.A. TIMES (June 26, 2020), https://www.latimes.com/california/story/2020-06-26/santa-cruz-becomes-first-u-s-city-to-ban-predictive-policing.
[72] Kashmir Hill, *How One State Managed to Actually Write Rules on Facial Recognition*, N.Y. TIMES (Feb. 27, 2021), https://www.nytimes.com/2021/02/27/technology/Massachusetts-facial-recognition-rules.html?searchResultPosition=3 (noting that Oakland, Portland, San Francisco, and Minneapolis have banned use of facial recognition technology).

harms.[73] There is a larger problem with advocating bans, too. Why should the police be barred from the potential benefits of the digital world, as every other sector of society moves in this direction?[74] Whatever the potential risks that arise from police use of AI systems, it would be strange to conclude that the solution would be a total prohibition on their use in law enforcement.

Third, some courts are beginning to consider the harms of AI-based systems with seriousness. These issues have been considered as traditional criminal procedure claims, such as the Wisconsin Supreme Court's 2016 decision that a proprietary risk assessment tool used to sentence the defendant did not violate his due process rights.[75]

But there are limits to this approach as well. Let's look at the framework of constitutional criminal procedure. By raising claims about the police tactics used in their own cases, defendants help define all of our rights. But defendants are inadequate proxies in the case of AI systems. In order to raise a criminal procedure claim, a defendant has to identify the evidence that came about as a result. But the police might rely on an AI system for the early stages of an investigation without collecting evidence. Or, the police might use an AI system for indiscriminate surveillance that only sometimes leads to the prosecution of individuals. At the same time, most of us would probably agree that the police should not use AI systems without any rules at all.

To be sure, the pursuit of local surveillance oversight mechanisms, the passage of bans for demonstrably flawed AI systems, and increasing judicial awareness of their pitfalls have made progress. Such measures have made the procurement of these tools and their costs more transparent, and thus more amenable to oversight. But ethical guidelines can address a broader set of issues in policing, including those situations where there may be not be harms in a traditional legal sense.

---

[73] *Cf.* Bruce Schneier, *We're Banning Facial Recognition. We're Missing the Point.*, N.Y. TIMES (Jan. 20, 2020), https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html?smid=url-share ("A ban on facial recognition won't make any difference if, in response, surveillance systems switch to identifying people by smartphone MAC addresses. The problem is that we are being identified without our knowledge or consent, and society needs rules about when that is permissible.").

[74] Andrew Ferguson makes a similar observation about what he characterizes as the "trap lens" with regard to new surveillance technologies. Ferguson, *supra* note 23, at 241 (noting that police abolitionists and advocates of bans "need to make an argument about why policing does not deserve to evolve in a digital world" when "every other professional enterprise has benefited from technological innovation").

[75] State v. Loomis, 881 N.W.2d 749, 770 (Wis. 2016) (holding that "if used properly with an awareness of the limitations and cautions, a circuits court's consideration of a COMPAS risk assessment at sentencing does not violate a defendant's right to due process").

III. THE CONTEXT OF AMERICAN POLICING

AI systems are already being used in American policing. Yet few police departments face any significant oversight or regulation. This is where a discussion of ethical and responsible policing might provide further guidance. Guidelines for responsible use of AI systems in policing are already a topic of public debate elsewhere.[76] Any conversation about such guidelines, however, should consider the specific context of American policing. In particular, we should highlight 1) the highly decentralized nature of American policing and 2) the longstanding racial tensions that are part of American police history.

A. Decentralization of Policing

One of the most distinctive aspects of American policing is its extreme decentralization.[77] To speak of "the police" in the United States is really to refer to the more than 18,000 individual law enforcement agencies, most of which are organized at the city and county levels.[78] There are more than 12,000 local police departments alone.[79] The most common type of agency is a small one, with ten or fewer offices, significantly smaller than the 40,000 officers in the New York Police Department.[80] And because most of these agencies are organized at the city or county level, they are controlled at the local level. States can and do impose rules on what police departments do within their borders, but not on every subject, and little has been done to control the police use of AI systems. Although the federal government can regulate, for instance, the private companies that design, sell, and use AI systems, it cannot regulate directly how states control their police agencies.[81] While the

---

[76] The Toronto Police Department, for instance, is currently developing an ethics policy for its use of AI systems. *See* Toronto Police Services Board, *supra* note 15.
[77] In fact, policing is so decentralized we have hard time counting how many agencies even exist. DUREN BANKS ET AL., U.S. DEP'T OF JUSTICE, NAT'L SOURCES OF LAW ENFORCEMENT EMPLOYMENT DATA 1 (2016) ("The decentralized, fragmented, and local nature of law enforcement in the United States makes it challenging to accurately count the number of agencies and officers.").
[78] *See id.*
[79] SHELLEY S. HYLAND & ELIZABETH DAVIS, U.S. DEP'T OF JUSTICE, LOCAL POLICE DEPARTMENTS, 2016: PERSONNEL 1 (2019).
[80] *See* BANKS ET AL., *supra* note 77. *See also* HYLAND & DAVIS, supra note 79, at 2 (observing that 48% of all local police departments employed less than 10 full time officers).
[81] *See, e.g.*, Printz v. United States, 521 U.S. 898, 936 (1997) ("The Federal Government may neither issue directives requiring the States to address particular

federal government can condition federal grants on changes in police conduct, there are few signs that funds used for technology purchases have any real constraints.[82]

When it comes to the tools police use, local officials like mayors and city councils are often the ones with the ability to impose conditions and requirements. Here, local communities can provide input, but as we saw earlier, local control of AI systems in policing has enjoyed limited success. Of the thousands of local governments, fewer than twenty have imposed any sort of regulations or requirements over how police can acquire or use these technologies.[83] While the pandemic has shown that communities can be engaged in and vocal about issues of local government, AI systems generate far less local engagement. This may be for a variety of reasons. People may readily accept police justifications that these systems are necessary innovations for criminal investigations. And many of these AI systems, including any potential for the harms or risks they pose, may be hard to explain and understand.

B. Racial Bias and Inequality

Concerns about bias are, of course, present in policing systems around the world.[84] However, the use of AI systems in American policing should be sensitive to our own particular context, history, and experiences. To raise the concern that AI systems used by the police might harbor bias or exhibit discriminatory behavior is to miss the point. Even as the murder of George Floyd while in police custody provoked

---

problems, nor command the States' officers, or those of their political subdivisions, to administer or enforce a federal regulatory program."). Accordingly, a proposed Algorithmic Accountability Act would direct Federal Trade Commission to require *companies* to reduce bias and improve privacy protections in the algorithms they produce. *See* Press Release, Office of Sen. Ron Wyden, Wyden, Booker, Clarke Introduce Bill Requiring Companies to Target Bias in Corporate Algorithms (Apr. 10, 2019), https://www.wyden.senate.gov/news/press-releases/wyden-booker-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms-.

[82] The millions distributed by the federal government for body cameras is a good example. Regulation over body camera use has been left up to states, cities, and individual departments. *See, e.g.,* Urban Institute, Police Body-Worn Camera Legislation Tracker (2018), at https://apps.urban.org/features/body-camera-update/ (noting that laws "governing how and when police body-worn cameras can be used and whether the footage is releasted vary considerably across the country").

[83] *See* Fidler, *supra* note 68.

[84] For instance, an important 2017 review of deaths in police custody commissioned by the UK Home Secretary stated that "Deaths of people from BAME communities, in particular young Black men, resonate with the Black community's experience of systemic racism." *See* ELISH ANGIOLINI, REPORT OF THE INDEPENDENT REVIEW OF DEATHS AND SERIOUS INCIDENTS IN POLICE CUSTODY 84 (2017).

national, and even global calls for greater accountability in policing, unequal and discriminatory policing is part of the history of American policing.[85] Before George Floyd, there were the deaths of Freddie Gray and Michael Brown. And before that was the death of Amadou Diallo and the abuse of Abner Louima. And before that, the beating of Rodney King. We could add to these individual cases the systematic reporting of racially biased policing against Black and Hispanic drivers,[86] Black and Hispanic pedestrians,[87] and even Black and Hispanic bicyclists.[88] The impacts of inequitable policing, then, are by definition unevenly experienced. Such experiences have left those most vulnerable to over-policing and discriminatory practices "legally estranged" from their own

---

[85] Former Minneapolis police officer Derek Chauvin was convicted of second-degree unintentional murder, third-degree murder and second-degree manslaughter after a widely circulated video captured him pressing his knee against George Floyd's neck on May 25, 2020. Police had responded to a call that Floyd had used a counterfeit twenty-dollar bill to buy cigarettes. *See* John Eligon et al., *Derek Chauvin Verdict Brings a Rare Rebuke of Police Misconduct*, N.Y. TIMES (Apr. 20, 2021), https://www.nytimes.com/2021/04/20/us/george-floyd-chauvin-verdict.html; Amy Forliti, *Explainer: What Next After Chauvin's Conviction on 3 Counts*?, ASSOCIATED PRESS, (Apr. 20, 2021), https://apnews.com/article/derek-chauvin-trial-charges-716fa235ecf6212f0ee4993110d959df.

[86] There are numerous studies here, dating back to the 1990s. A pioneering observational study by John Lamberth found that African Americans made up 13.5% of the population on the New Jersey turnpike and 15% of speeders but represented 35% of those pulled over by the police. In other words, African Americans were 4.85 times as likely to be stopped as others. *See* John Lamberth, *Driving While Black*, WASH. POST (Aug. 16, 1998), https://www.washingtonpost.com/archive/opinions/1998/08/16/driving-while-black/23ecdf90-7317-44b5-ac43-4c9d7b874e3d/ (summarizing his study's methodology and findings); *see also* David A. Harris, *The Stories, the Statistics, and the Law: Why "Driving While Black" Matters*, 84 MINN. L. REV. 265 (1999). The nonpartisan Public Policy Institute of California provides a recent example of similar findings. *See* Magnus Lofstrom et al., *African Americans Are Notably Overrepresented in Police Stops*, PPIC (Aug. 13, 2020), https://www.ppic.org/blog/african-americans-are-notably-overrepresented-in-police-stops/ (finding in review of 1.8 million police stops, "the data clearly shows that African-Americans make up a much larger share of interactions with law enforcement relative to their populations [sic] share than any other racial/ethnic group in California").

[87] *See, e.g.*, Lyndsay Winkley & Teri Figueroa, *Another Report Finds Deep Racial Disparities in Sheriff's Departments Stop Data*, SAN DIEGO UNION-TRIB. (Dec. 9, 2021), https://www.sandiegouniontribune.com/news/public-safety/story/2021-12-09/another-report-finds-deep-racial-disparities-in-sheriffs-department-data (citing Center for Policing Equity study finding "Black pedestrians were stopped by sheriff's deputies 3.5 times as often" compared to Whites).

[88] *See, e.g.*, Alene Tchekmedyian et al., *L.A. Sheriff's Deputies Use Minor Stops to Search Bicyclist, With Latinos Hit Hardest*, L.A. TIMES (Nov. 4, 2021), https://www.latimes.com/projects/la-county-sheriff-bike-stops-analysis/ (documenting more than 44,000 bike stops logged by the Sheriff's Department and finding 7 of 10 stops involved Latino cyclists).

police departments.[89] These and countless other incidents in American policing have generated countless commission reports, lawsuits, and calls for reform for nearly a century.[90]

Thus the risks of AI in policing arise in the context of an institution that has a long history of meting out justice unequally and in a discriminatory way. What follows? First, bias in AI systems can perpetuate existing biases or introduce new ones, but it does so in the context of a social institution with a long history of discrimination, especially against African-Americans. We should not be surprised, then, if the use of an AI system in a community in longstanding tension with its local police department meets skepticism, resistance, or calls for prohibition.

Second, crafting AI ethics for policing requires speaking to two different audiences. Each is important but distinct. One audience is engaged primarily in "tech policy": the drafting and decision-making of rules and policies that engage in the use of technologies across industries and institutions. Advocacy organizations and policymakers engaged in AI policy often address the use of AI in matters that can include online speech, advertising, healthcare, lending, and employment. Policing is only one subject, and subsumed under criminal justice policy, at that. And even when policing is a concern, this tech policy lens tends towards a focus on individual privacy and the harms of mass surveillance.

On the other hand, the Black Lives Matter movement and related campaigns have focused on police violence and addressing longstanding structural problems in the relationship between the police and marginalized communities. Young African-American men make up an overwhelming number of those killed by police, year after year.[91] Many

---

[89] Monica Bell's theory of legal estrangement describes this problem well: one that captures "both legal cynicism-the subjective "cultural orientation" among groups 'in which the law and the agents of its enforcement, such as the police and courts, are viewed as illegitimate, unresponsive, and ill equipped to ensure public safety' and the objective structural conditions (including officer behaviors and the substantive criminal law) that give birth to this subjective orientation." Monica C. Bell, *Police Reform and the Dismantling of Legal Estrangement*, 126 YALE L. J. 2054, 2066-67 (2017).

[90] President Hoover's commission of the Report of the Enforcement of the Prohibition Laws, better known as the Wickersham Report, was among the first national reports focusing on problems in policing. *See* NATIONAL COMMISSION ON LAW OBSERVANCE AND ENFORCEMENT, REPORT NO. 2, REPORT ON THE ENFORCEMENT OF THE PROHIBITION LAWS OF THE UNITED STATES (1931).

[91] Starting in 2015, the Washington Post has tracked every fatal shooting by a police officer in the United States. Among its findings is the observation that African Americans are killed by the police at more than twice the rate of Whites. *See Fatal Force: 1022 People Have Been Shot and Killed by Police in the Past Year*, WASH. POST

Black and Hispanic communities are simultaneously over-policed and under-policed.[92] We know from federal investigations that some municipal budgets literally depend on fines and fees, almost always imposed on the poor, and always meted out by local police.[93] These problems have been rightly identified as reasons for desperately needed police reforms.

To be sure, there are groups and voices that have brought these two concerns together. Some civil rights groups have made explicit the disproportionately borne harms of unregulated AI systems on marginalized communities.[94] This has led, for instance, to a coalition of civil rights groups to publish "civil rights principles for the era of big data."[95]

IV. ETHICAL COMMITMENTS IN AI-SYSTEMS IN POLICING

What then, do we mean by the ethical use of AI in American policing? Police departments should make prior public commitments to the values they adopt as they rely on AI systems of all types. Ethical commitments can serve as meaningful guides, even if they lack penalties or enforcement consequences.[96] These commitments should embody

---

(Feb. 2, 2022), https://www.washingtonpost.com/graphics/investigations/police-shootings-database/.

[92] Alexandra Natapoff, *Underenforcement*, 75 FORDHAM L. REV. 1715, 1775 (2006) ("Our criminal system is rife with inegalitarian enforcement failures—pervasive, yet little-noticed way that the state predictably abandons its constituents by failing to enforce the rules.").

[93] *See e.g.*, U. S. DEP'T OF JUSTICE, C.R. DIV., INVESTIGATION OF THE FERGUSON POLICE DEPARTMENT 2 (2015) ("The City budgets for sizeable increases in municipal fines and fees each year, exhorts police and court staff to deliver those revenue increases, and closely monitors whether those increase are achieved.").

[94] *See, e.g.*, Letter from Am. Civ. Liberties Union et al., to Dr. Eric S. Lander, Dir., White House Office of Sci. & Tech. Pol'y, Exec. Off. of the President, et al., Centering Civil Rights in AI Policy (July 13, 2021) (available at https://www.upturn.org/static/files/2021-07-13%20Coalition%20Letter%20to%20OSTP%20on%20Centering%20Civil%20Rights%20in%20AI%20Policy.pdf) (urging White House Office of Science & Technology Policy to "bring civil rights and racial justice to the forefront of AI policy across the board in areas beyond national security—in housing, in employment, in criminal legal issues, and more.").

[95] *See Civil Rights Principles for the Era of Big Data*, LEADERSHIP CONF. ON CIV. & HUM. RTS. (Feb. 27, 2014), https://civilrights.org/2014/02/27/civil-rights-principles-era-big-data/ (urging that "it is vitally important that these technologies be designed and used in ways that respect the values of equal opportunity and equal justice").

[96] This is the principle underlying soft law: "instruments or arrangements that create substantive expectations that are not directly enforceable, unlike 'hard law' requirements such as treaties and statutes." *See* Gary E. Marchant & Brad Allenby, *Soft Law: New Tools for Governing Emerging Technologies*, 73 BULL. ATOMIC SCIENTISTS 108, 112 (2017) (arguing that one "soft-law category of potential relevance

social values, not just legal or technocratic concerns. This section identifies four ethical commitments we can embrace in policing. These propositions are not meant to be exclusive, but rather a starting point for further development.

A. Transparency and Oversight Mean Little Without Broad Explainability

Think of this principle of the "why" of AI. We can begin with the narrower definition of explainability in AI policy discussions. Explainability refers to the idea that a person subjected to a decision or outcome informed by an AI system should be able to understand how the system works, and why a particular decision was reached in their case.[97] This specific sense of explainability matters because AI systems can be both difficult to explain and understand, and yet also have direct impacts on people's lives.[98]

We can find this call for explainability in AI policy discussions across many fields. That is because explainability can serve multiple goals, including giving users confidence in AI systems, reducing bias, meeting regulatory standards, and helping to improve the AI system itself.[99] But these differing goals mean that the requirement of explainability means different things to different audiences. For developers, explainability might include actions like publishing the algorithm or creating systems that are inherently interpretable rather than creating models that are difficult to understand.[100] For individuals facing an adverse decision made by an AI system, that might mean having the decision-making process made understandable to a layperson.[101]

For the police, explainability matters in several senses. First, there is the individual affected by an adverse decision. In other fields, that might mean the person turned down for a loan or a person who is skipped over for a job interview because of an automated decision. In

---

to many emerging technologies includes various types of private standards, guidelines, codes of conduct, and principles").

[97] *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

[98] *See, e.g.*, THE ROYAL SOCIETY, EXPLAINABLE AI: THE BASICS 5 (2019) ("There has, for some time, been growing discussion in research and policy communities about the extent to which individual developing AI, or subject to an AI-enabled decision, are able to understand how AI works, and why a particular decision was reached.").

[99] *See id.* at 9-10 (discussing justifications for explainability requirement).

[100] *See id.* at 12-13 (explaining how different explainability needs require different actions).

[101] An example of this would be an explanation of why an applicant was turned for a loan through an automated process. *See id.* at 14.

policing, that adverse decision might include decisions like a purported facial recognition match, an assessment of risk during a traffic stop, or a prediction of violent behavior leading to a further investigation. Individuals routinely contest even traditional policing actions. Explainability helps people understand how an automated process came to a particular decision, whether it might contain errors, and thus provide a possible basis for contestation and appeal.

Second, the community being policed is owed a different form of explainability. The responsible use of AI in policing also requires a clear explanation of why any particular AI system is worth adoption. Why should a particular risk assessment tool, for instance, be favored over other approaches to identify persons or places in need of intervention? Why would any AI system be preferred over the existing policing approach? A dominant theory in policing studies focuses on procedural justice: that people view the police as legitimate when they have been treated with fairness and respect.[102] Legitimacy matters in this perspective because it, rather than the risk of punishment, is the basis for why people obey and follow the law. The hasty and secretive introduction of AI systems for policing can only detract from a community's perception of how fairly its police conduct themselves.

Third, there are the police themselves. Artificial intelligence married with robotics may one day lead to nearly total automation in policing. Today, though, police typically *implement* decisions suggested by AI. Whether the police receive forecasts, threat assessments, or image matches, explainability means that officers should understand how these systems work, and their limitations. Without this kind of explainability, police officers face risks. They may blindly follow the assessment of an AI system without taking further steps to verify or confirm.[103] Alternatively, they might balk at a prediction they cannot explain, and follow through with their own intuitive decision.[104]

---

[102] Tom Tyler's scholarship is most closely associated with these insights. *See, e.g.*, Jason Sunshine & Tom R. Tyler, *The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing*, 37 L. & SOC'Y REV. 513 (2003); Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME & JUST. 283 (2003).

[103] CDEI, *supra* note 14, at 68 ("One possibility is that the decisionmaker over-relies on the automated output, without applying their professional judgement to the information.").

[104] *See id.* (noting possibility that a "human decision-maker [may feel] inherently uncomfortable with taking insights from an algorithm to the point where they are nervous to use it at all").

B. Fairness is Not Just the Reduction of Bias in AI Systems Used for Policing

Fairness concerns are common across many AI policy documents, and most discussions consider fairness to mean the impartial and equitable treatment of persons.[105] In its survey of more than thirty source materials from around the world, the Berkman Klein Center found that some principle of "fairness and non-discrimination" was the most highly represented theme.[106] Fairness in AI systems can mean many things, including considerations of how an AI system might visit disproportionate harms, inclusiveness in AI design, and fair representation in data sets used for training models.[107] Though they differ in detail, proposals to reduce bias in AI systems are motivated by the need to establish and increase trust and legitimacy in the public.

But the adoption of AI systems poses a unique challenge for American policing. Although the Obama administration's Twenty-First Policing Report offered hopeful predictions for the future of policing, American policing today finds itself embroiled in crises, along race, class, and political lines.[108] In this context, a narrow concept of fairness is ill-suited to AI systems in policing. Instead, the principle of fairness should consider how the AI system contributes to an improvement in the provision of policing services.

A broader view of fairness includes both attention to specific issues of bias in AI systems, as well as how these systems fit into the broader delivery of fair policing, especially to marginalized communities.[109] We can use facial recognition as an example. Much attention has been given to the high rates of erroneous matches for non-whites. A narrow view of fairness would recognize that this problem stems from the underrepresentation of non-whites in the training data of

---

[105] *See, e.g.*, Berkman Klein Report, *supra* note 17, at 49; CDEI, *supra* note 14, at 3 (noting "urgent need for the world to do better in using algorithms in the right way: to promote fairness, not undermine it").
[106] Berkman Klein Report, *supra* note 17, at 47.
[107] *See id.*
[108] *Cf.* Cynthia Lum & Daniel S. Nagin, *Reinventing American Policing*, 46 CRIME & JUST. 339, 339-340 (2017) (observing that American policing is experiencing a "tumultuous period" and suggesting that new strategies must focus on crime prevention and citizen reaction).
[109] *See* European Commission Community Research and Development Information Service, *Shaping the Ethical Dimensions of Smart Information Systems (SIS) – A European Perspective* (SHERPA) Deliverable No. 1.4, 41 (2019), https://cordis.europa.eu/project/id/786641 ("One of the reasons why the rise of datafication and algorithmic decision-making has an effect on issues of justice is its burden on predominantly poorer members of society").

a facial recognition program and would seek to address it. A broader view of fairness in policing would ask whether certain uses in the community would be unfair, even if the software's identification rates were made fairer. Even a more accurate facial recognition system used indiscriminately during traffic stops would be unlikely to satisfy broad fairness concerns.[110]

C. Privacy and Fairness Represent Different Values

Privacy and fairness are commonly used terms in discussions of ethical AI system use, but they represent different values. We can think of privacy as a form of shielding or controlling individual information from unwanted exposure.[111] It is, at its core, an individual protection. Policing scholars and civil rights advocates have focused on the harms posed by increasingly powerful and ubiquitous surveillance technologies like facial recognition, license plate readers, and a generation before that, closed-caption television cameras. They target these technologies because they collect enormous amounts of data and impact privacy and its associated individual constitutional rights, like free expression and anonymity.

Fairness, however, is different. Fairness can be a value for individuals and communities. And fairness in the use of AI systems can have multiple meanings as well. Fairness might mean that an individual subjected to, say, a facial recognition match is assured that the software has been designed and assessed to minimize bias for race, ethnicity, and gender. But fairness also means where, when, and how that facial recognition technology is used as a policing practice in the community. What is more, because fairness is a principle of police reform outside of AI tech policy, all of these forms of fairness should be compatible with one another.

And we might also imagine instances where privacy and fairness values might exist in conflict. Consider this hypothetical. When autonomous driving technology becomes widespread, should police be

---

[110] *See* Caroline Haskins, *A Popular Workshop for Police Encouraged Cops to use Face Scans to ID People They Pull Over at Traffic Stops*, BUS. INSIDER (Feb. 2, 2022), https://www.businessinsider.com/police-workshop-street-cop-training-podcast-facial-recognition-traffic-stops-2022-2 (describing police instructor advising police "to use facial recognition at traffic stops in order to find out a person's identity and if they have a warrant out for their arrest, even if it's unclear whether that person committed a crime").

[111] There is an enormous literature on privacy and the law and many definitions of privacy. *See generally* DANIEL J. SOLOVE, UNDERSTANDING PRIVACY (2010) (arguing that there is no single workable definition of privacy).

able to conduct remote traffic stops or remote enforcement? Privacy advocates might object that such stops and enforcement actions might collect unnecessarily large amounts of data, might be subject to security breaches, and might intrude upon perfectly lawful behavior. On the other hand, a remote police action vastly reduces the potential for police violence. For some, the reduction in potential violence may outweigh concerns about individual privacy. There may be other reasons communities would object to this increased automation, but this example suggests that privacy and fairness considerations do not always coincide.

D. Responsible AI Use Factors in the Nature and Degree of Private Sector Reliance

Finally, the responsible use of AI systems in policing should consider the risks inherent in privately developed tools. In the U.S., most of the AI systems used in policing are products developed by private companies.[112] Whether a predictive policing tool or a records management system, these tools are marketed to the police who are customers. Police departments may purchase these tools, but increasingly common are subscription-based models in which the public agencies never own either software or hardware.[113] Just like retail customers, police departments may be enticed by the promise of future upgrades, but these newly important relationships may strain a model of responsible policing.

These customer-vendor relationships hold the potential to pose obstacles to responsible policing. Not only is there is an algorithmic "black box" problem that makes it difficult for even developers to explain the AI systems that they have designed, there is the added complication of corporate secrecy. The invocation of trade secrets and non-disclosure agreements, and general claims of proprietary information are common in the commercial world, but unusual in traditional policing. These claims also mean that there is another layer of secrecy around these AI systems.

---

[112] *Cf.* Hannah Bloch-Wehba, *Visible Policing: Technology, Transparency, and Democratic Control,* 109 CAL. L. REV. 917, 919 (2021) (noting new police technologies are "often procured from or otherwise reliant on the private sector").

[113] Axon is increasingly focused on offering a SaaS (Software as a Service) to law enforcement agencies. Brett Schafer, *How the Company Behind TASER Guns is Becoming a SaaS Powerhouse,* MOTLEY FOOL (Mar. 3, 2021), https://www.fool.com/investing/2021/03/03/how-company-behind-taser-becoming-saas-power/.

The more that policing outsources its functions, from the development of suspicion to the most mundane information processing, the more it relies on the judgments of private companies about what responsible AI systems will do and how they will behave. In the United States, Axon Enterprise is a dominant provider of policing platforms. The public may associate Axon with the body cameras it licenses to police departments around the country, but an increasingly larger share of its revenue is invisible to the public. Police department customers pay Axon yearly recurring subscription fees for data storage and software access stored in Axon's cloud servers.[114] As police increasingly must rely on private platforms to collect, store, and analyze the information they process, they become beholden to these companies' decisions.

The need to impose public oversight and enact regulations to curb the influence of these private companies on policing has been recognized by scholars[115] and has been the subject of some local government action.[116] Framing this as an ethical concern, in addition to pushing for traditional regulatory concerns, can help communities in their oversight of their own police departments.

E. AI Systems in Policing Don't Need to End with Policing

The promise of AI systems is that we can sift through the vast amounts of digitized data to identify patterns: patterns of financial irresponsibility, ill health, job unsuitability, and crime. Even if we could successfully address the concerns raised by the current use of AI systems—bias, opacity, and so on—we would still be left with what to do with these insights. In other words, implementation is still a human decision.

Implementation too can be part of an ethical framework for the use of AI systems in policing.[117] If we can forecast crime, is the

---

[114] Dana Goodyear, *Can the Manufacturer of Tasers Provide the Answer to Police Abuse?*, THE NEW YORKER (Aug. 20, 2018), https://www.newyorker.com/magazine/2018/08/27/can-the-manufacturer-of-tasers-provide-the-answer-to-police-abuse (describing Axon as having "an iPod/iTunes opportunity—a chance to pair a hardware business with an endlessly recurring and expanding data-storage subscription plan.").

[115] *See, e.g.*, Catherine Crump, *Surveillance Policy Making by Procurement*, 91 WASH. L. REV. 1591 (2016) ("Surveillance policy making by procurement can short-circuit [the process of local control] when elected officials and the public are left without a meaningful understanding of what technologies their law enforcement agency is acquiring.").

[116] *See supra* part II (discussing local surveillance technology ordinances).

[117] *Cf.* Calo, *supra* note 19, at 412 (noting danger that AI systems can be "selectively applied to . . . marginalized populations").

responsible approach one of increased police presence? If we can identify who might be at high risk for offending or victimization, are police interventions the appropriate consideration?

Such questions speak to a broader audience than those engaged in AI policy. The movement to "abolish the police" is a reaction to distrust and to the call for social solutions beyond traditional law enforcement. Asking mental health specialists to respond to mental health crises is a way of responding to these concerns. So too is asking whether the assessments of AI systems in policing should be met with novel responses rather than traditional police investigations.

CONCLUSION

AI systems are everywhere. Most people are used to them in their daily lives, and they are increasingly important decision-making mechanisms in social services, healthcare, finance, and criminal justice. In this sense, the use of AI systems in policing is part of a larger social transformation.

And just as in many of these fields, the regulation and oversight of AI systems in policing is woefully inadequate. We have no real national standards in the United States. Existing efforts are piecemeal and slow going. One way to address this gap is to introduce ethical principles. Many non-profits and governmental bodies around the world are in the process of drafting ethical guidelines. These guidance documents are not binding or enforceable, but they are far preferable to no standards at all.

The use of AI in policing stands at the intersection of two distinct discussions: the widely acknowledged need for ethical principles in the use of AI systems, and the renewed attention to inequality and bias in American policing. Just as in lending, employment, and healthcare, the use of AI systems in policing needs not just greater regulation, but also a set of principles to guide their use with responsibility. In this way, ethical considerations can contribute to the larger project of police reform and even conversations about envisioning policing entirely differently.

VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

NOTES

STRUCTURED PSYCHOMETRICS IN BIGLAW TALENT ACQUISITION: AI-DRIVEN QUANTITATIVE FIT

Joseph J. Kim

STRUCTURED PSYCHOMETRICS IN BIGLAW TALENT ACQUISITION: AI-DRIVEN QUANTITATIVE FIT

Joseph J. Kim[*]

INTRODUCTION

"The vast majority of hiring practices today are based on 'the way it has always been done' . . . based upon gut feelings, intuition, emotions, subjective beliefs, and common misconceptions about what actually works."[1] This criticism rings true as well for the hiring practices of large law firms in the U.S., which have shown little industry-wide change since the advent of the *Cravath System*, what Cravath, Swaine & Moore LLP calls their "model for developing talent, incentivizing collaboration and client service, and building long-term relationships of trust."[2] The Cravath System is widely emulated by a category of law firms ("Biglaw") that typically are the largest—in both attorney headcount and geographic reach—and compensate competitively amongst each other. The Cravath System seeks to derive partners "from the ranks of [associates]" and to recruit "the most promising students from a diverse array of excellent law schools" while providing "associates with rigorous and expansive training."[3] Such a model for attracting and developing talent has grown to dominate Biglaw and retains an impressive amount of inertia. "Doing something else than the norm requires effort. But it's easy to say that hiring is important. And it's easy to use the same hiring process and screening questions as everyone else."[4] For many decades now, Biglaw has comfortably settled on the Cravath System's hiring philosophy as a sufficient and preferred talent acquisition model.

The Cravath System is not just a talent acquisition model, it is also a talent development model intended to be applied to the same

---

* Candidate for Juris Doctor, Notre Dame Law School, 2023; Bachelor of Science in Human Resource Management, La Sierra University, 2020. I would like to thank Professor Matthew J. Barrett for his guidance and suggestions, Alec Afarian and Malcolm Coffman for their valuable insights, and my family for their encouragement and support. I would also like to thank my colleagues on the Notre Dame Journal on Emerging Technologies for their diligent and thorough editing. All errors are my own.

[1] ATTA TARKI, EVIDENCE-BASED RECRUITING: HOW TO BUILD A COMPANY OF STAR PERFORMERS THROUGH SYSTEMATIC AND REPEATABLE HIRING PRACTICES xiii (2020).

[2] *The Cravath System*, CRAVATH, SWAINE & MOORE LLP, https://www.cravath.com/the-cravath-system/index.html (last visited Jan. 29, 2022).

[3] *Id.*

[4] TARKI, *supra* note 1, at 20.

individuals it attracts. "However, the strategy of developing your own talent requires enormous discipline and bold bets in building the infrastructure needed to succeed in deploying this strategy."[5] The Cravath System has acted as the backbone of Biglaw for many years now and law firms have generally not been perceived to tout incompetent professionals. While the enormous discipline practiced and bold bets currently placed by Biglaw can be argued as more or less effective, improvement is always possible, especially in a business world with ever-evolving goals and competition. Biglaw risks growing complacent, weathering undesirable turnover rates in hopes of producing enough star talent to maintain profit margins and competitive edges. However, "traditional strategies are no longer enough. In today's era, your team's talent and passion should be your competitive advantage."[6] How can Biglaw gain the courage to evolve out of the cautious approach to attain new competitive advantages when the industry as a whole is reluctant to innovate? "The cautious approach is a 'recipe for mediocrity,'"[7] but mediocrity is not what drives the success of law firms. Law firms want rainmakers—profitable partners who have survived unfavorable turnover rates—but little has been done to identify who will or will not become a rainmaker. "If your talent acquisition playbook is the same as most other [firms], you're in trouble. Chances are that another firm is going to run the same plays with more resources and superior talent—and win."[8]

Sadly, this is exactly what has been occurring, except that no firm is truly winning. Biglaw has found itself in a perpetual arms race for talent through compensation. But, even market-leading firms find that "their competitors have followed suit and, in effect, will merely have raised the compensation bar for their industry."[9] The ineffectiveness of salary-raising races can be evidenced by the tech industry's growing

---

[5] *Id.* at 8.
[6] *Id.* at ix.
[7] *Id.* at 8.
[8] *Id.* at 107.
[9] *Id.* at 131; *see Biglaw Salary Scale*, BIGLAW INVESTOR, https://www.biglawinvestor.com/biglaw-salary-scale/ (last visited Jan. 29, 2022) ("The Cravath scale has largely stayed the same across the major law firms because those firms are competing for the best law students from the best law schools. If one firm offers a higher salary, historically the other firms tend to announce salary increases shortly thereafter."); *see also* Dylan Jackson, *The Cost of the Talent War: Bonuses, Raises Drive Up Big Law Compensation Expenses by Double Digits*, THE AM. LAW. (Dec. 10, 2021, 5:00 AM), https://www.law.com/americanlawyer/2021/12/10/the-cost-of-the-talent-war-bonuses-raises-drive-up-big-law-compensation-expenses-by-double-digits/ ("When you look at the expenses of law firms, the No. 1 cost is people.").

capture of graduates from top-ten MBA programs, where "despite lower salaries, tech has been able to extract more talent from these elite programs"[10] like the traditional MBA routes of financial services and consulting firms (two firm types that acquire talent using a similar process as Biglaw). While law firms have not yet faced such threats to as significant of a degree, the future is far from secure. The Big Four accounting firms, despite currently paying less than half of Biglaw's starting salaries, have been perceived for over a decade now as a looming competitor for law school graduates.[11] The accounting firms are primarily prevented from encroaching on Biglaw's business (for now) by the inability to practice law rather than an inability to compensate.[12] Nevertheless, accounting firms have increasingly employed law school graduates in past years.[13] And, in a scenario in which accounting firms begin hiring practicing lawyers, law firms will suddenly have to compete for talent, beyond compensation.[14] Later described in this Note, such a scenario could prove problematic to Biglaw because firms do not screen for associates that openly desire high compensation; they instead interview for the exact opposite—intrinsically motivated employees.[15] Law firms offer competitive compensation, but do not default to selecting compensation-motivated employees.[16] Further, these compensation-motivated employees may drift to firms able to offer a larger variety of

---

[10] TARKI, *supra* note 1, at 133.
[11] Victoria Hudgins, *'Business-Minded' Law School Students Grab Big 4's Hiring Attention*, LAW.COM (Jan. 20, 2021, 11:38 AM), https://www.law.com/legaltechnews/2021/01/20/business-minded-law-school-students-grab-big-4s-hiring-attention/; Aaron Muhly, *Talent Battle: Big Four vs. Big Law*, EVELAW, (June 25, 2019), https://www.evelaw.eu/blog/2019/6/20/talent-battle-big-four-vs-big-law.
[12] *See* Meg McEvoy, *ANALYSIS: The Big 4 Is Knocking – Are State Bars Answering?*, BLOOMBERG L. (Sept. 18, 2019, 5:01 AM), https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-the-big-4-is-knocking-are-state-bars-answering ("U.S. law firms are still somewhat insulated from competition from the Big Four by attorney ethics rules that, in theory, bar them from practicing U.S. law. When and how the large accounting firms will enter the practice of U.S. law is one of the biggest debates in the legal industry today.").
[13] Bruce MacEwen, *Whither the Big 4?*, ADAM SMITH ESQ. (Jan. 23, 2016), https://adamsmithesq.com/2016/01/whither-the-big-4/ (describing how in 2016, "about 5-10% of US law school graduates went to work for an accounting firm").
[14] Bruce MacEwen, *The Associate Comp Wars & Thick/Thin Communities*, ADAM SMITH ESQ. (Mar. 10, 2022), https://adamsmithesq.com/2022/03/the-associate-comp-wars-thick-thin-communities/2/ ("compensation per se ranks as 'one of the three [motivators] at the bottom of the list'" of factors that contribute to job satisfaction).
[15] LAUREN A. RIVERA, PEDIGREE: HOW ELITE STUDENTS GET ELITE JOBS 151 (2015) ("[T]he best paths and values were those presented as having been guided by intrinsic versus extrinsic motivations.").
[16] *Id.* at 163-64.

non-compensation reasons to work.[17] As it becomes clearer that law firms are mirroring each other's hiring practices with no real hiring threat around, it also becomes clearer that few law firms have any *unique* competitive advantage in terms of talent, which is the very prized possession of any professional service firm. In Biglaw's tight contest for talent, slightly tipping the scales by finding marginally better candidates can make an impressive difference.

Although a law firm may employ hundreds or even thousands of revenue-generating employees, law firms operate under a Pareto or power-law distribution,[18] where it quickly becomes apparent that having one rainmaking partner is many times more valuable than an army of entry-level associates. This is especially true in a profession where "the dollar value produced by each person can be precisely tracked"[19]—the billable hour for associates and fees collected for partners conveniently serving as the dominant measures of productivity in law firms. Power-law distribution is not unique to law firms, since research of over 600,000 professionals and 198 samples showed that "[r]esults are remarkably consistent across industries, types of jobs, types of performance measures, and time frames and indicate that individual performance . . . follows a [Pareto] distribution."[20] Thus, it is advantageous for law firms to improve at recognizing and acquiring talent with the highest productive potential.

While the efficacy of existing hiring practices in Biglaw is certainly debatable, moving a mountain is not accomplished by finding the biggest shovel possible. A complete overhaul will take decades, bring chaos and costs, and be subject to great resistance in an already resistant-to-change industry.[21] Blatant first-movers will be punished by clients and

---

[17] MacEwen, *supra* note 14 ("[F]irms raising comp at the fastest rate fare no better at all in retention than their lagging-behind peers.").

[18] *See* Michael Barrons, *Do the Math: How the 80/20 Rule Can Elevate Law Firm Productivity*, INFOWARE (Oct. 12, 2017), https://infowaregroup.com/blog/do-the-math-how-the-80-20-rule-can-elevate-law-firm-productivity. This is true not only between colleagues but also between firms. *See* Bruce MacEwen, *Is Your Firm Playing to Win, or Not to Lose?*, ADAM SMITH ESQ. (Oct. 7, 2018), https://adamsmithesq.com/2018/10/is-your-firm-playing-to-win-or-not-to-lose/ ("10% of the entire revenue of the [AmLaw] 100 firms is accounted for by the top three . . . and the top nine firms garnered as much revenue as the entire bottom half of the 100 firms.").

[19] TARKI, *supra* note 1, at 6.

[20] *Id.* at 7.

[21] *See Overcoming Lawyers' Resistance to Change*, THOMSON REUTERS, https://legal.thomsonreuters.com/en/insights/articles/overcoming-lawyers-resistance-to-change (last visited Mar. 26, 2022); *see also* Himesh Chavda, *Breaking the Resistance to Change – The Cultural Challenges Hindering Innovation in Law*, LAW.COM (Jan. 15, 2019, 12:00 AM), https://www.law.com/international-

competition alike and will risk the disease of over-uniqueness.[22] There is also no obvious argument that Biglaw needs a complete transformation either because law firms are already successful as is in developing rainmakers and pleasing clients with competent legal services.[23] Change must be slow and deliberate using tools that are certain to work and will bring meaningful impact. Here, the task is not to upend what law firms are looking for; we can put faith into the fact that law firms have been and are continuing to be profitable while using the Cravath System.[24] Instead, the task is to approach select imperfections and improve Biglaw's hiring process rather than its hiring criteria. This Note combines a number of perspectives and disciplines to proffer a unique suggestion toward recognizing better talent and acquiring a new intra-industry competitive edge.

First, the Cravath System will be described and stripped down to its recognizable components. Understanding what Biglaw seeks in talent, how they have been finding it, and why the hiring practices have not budged will be crucial. Biglaw falls under the category of elite professional services ("EPSs"),[25] where both the criteria and process can be collapsed into one term: the *sponsored contest*.[26] After describing the sponsored contest, which gives great insight into both the what and the why of the Cravath System, the criteria and process behind Biglaw hiring

---

edition/2019/01/15/breaking-the-resistance-to-change-the-cultural-challenges-hindering-innovation-in-law/.

[22] Zebras are an excellent example of the potential dangers of standing out. Although camouflage is a default explanation for many animals' fascinating physical appearances, the zebra stands out in the Savannah grass while its lion predators are camouflaged instead. When scientists marked a particular zebra in order to distinguish it from the herd for studies, the distinction would lead to the zebra getting eaten by lions the quickest. The stripes on a zebra appear to camouflage them with each other than with the environment. Like a red-painted zebra, an over-unique law firm whose competitors and clients alike can distinguish as diverging too far from an accepted practice places itself in a high-risk high-reward scenario that can resemble gambling more than smart business strategy. *See* Taylor Foreman, *This Weird Zebra Story Will Make You Understand Creativity*, ILLUMINATION (Aug. 16, 2020), https://medium.com/illumination/this-weird-zebra-story-will-make-you-understand-creativity-89c83fce6ce4; *see generally* Brad Shorr, *Being Unique Is a Bad Way to Sell*, LEAD GENERATION INSIGHTS (May 9, 2017), https://www.straightnorth.com/insights/being-unique-bad-way-sell/.

[23] Nicholas Bruch, *Law Firms Are More Profitable Than Ever. How are They Doing It?*, L. J. NEWSL. (Nov. 2018), https://www.lawjournalnewsletters.com/2018/11/01/law-firms-are-more-profitable-than-ever-how-are-they-doing-it/?slreturn=20220029024524 ("[T]he vast majority of firms within the Am Law 200 have reported increases in inflation adjusted [profit-per-equity-partner] over the past decade.").

[24] *Id.*

[25] RIVERA, *supra* note 15, at 16.

[26] *Id.* at 30.

will be examined to see if and where it is open to improvement. Much like a patient's visit to the doctor, positive change begins with diagnosis rather than jumping straight to prescription. I will argue that it is the process—namely grounded in an excess of subjective evaluation—and not the criteria that warrants a priority for change if the goal is to make adjustments that are both meaningful and implementable. Thus, a proposal for updating the process, in order to be palatable and practical, must either stay true to the existing criteria or only offer the slightest of tweaks. The proposed change is *quantitative fit*, or an objective and structured evaluation of personality using psychometrics. The central reason why personality measurement can provide a proper change in the hiring process without perverting criteria is that Biglaw *already* subjectively selects for personality.

Second, because the prescription for changing the hiring process will be to introduce personality psychometrics, the Five-Factor Model of personality will be described, its methodology analyzed, and its validity and reliability defended. Then the Five-Factor Model will be made applicable to identified issues in the Biglaw hiring process. Each personality attribute will be assessed for its capacity to help identify proper talent for Biglaw. I will argue that objective personality testing will both rein in and complement the hyper-subjective process that law firms currently rely on.

Third, this Note will examine the effective implementation of quantitative fit. One apparent challenge is ensuring that personality profiles are considered in light of the endless, dynamic, immeasurable, and complex variables that go into understanding a human. In addition, in seeking to complement and not replace existing Biglaw hiring practices, quantitative fit should not draw conclusive boundaries to define what proper talent looks like. Rather, quantitative fit should be used by law firms as a calibration tool to guard against the fallibility of human judgment. I will argue that artificial intelligence can help to make quantitative fit sustainably implementable. Although Biglaw's talent acquisition model has been stable for decades, the modern professional world changes daily, and "[t]here are no quick fixes, and nothing works *all the time*."[27] What law firms want may change as the market, economy, candidate pool, technology, clients, and who-knows-what change. Artificial intelligence can become the tool by which a candidate's personality and the ever-changing desires of law firms remain aligned. This Note argues that implementation via artificial intelligence can play

[27] TARKI, *supra* note 1, at 37.

two simultaneous roles: reinforcing existing criteria while transforming process.

"Top-tier talent used to be equally inaccessible to all companies, but now . . . is reachable by companies that embrace innovating technologies and practices."[28] With Biglaw being relatively non-innovative to date, there exists an undefined and untapped potential for competitive edge driven by better measuring personality. In fact, personality is perhaps the most important measurement that already exists in Biglaw. In all EPS firms, Biglaw especially, *fit* is an anchoring criterion at the interview stage, a necessary component to ultimately be selected for a position. "Interviews—which . . . carried great weight in final hiring decisions—were seen as highly subjective assessments *based on applicants' personalities* rather than their qualifications listed on paper."[29] Because fit is shorthand for a candidate's personality, the existing criteria simply become complemented with its structured version in the form of a statistically rigorous personality profile—the quantitative fit. It is prescribing an objective and structured perspective to the evaluation of personality, specifically using machine learning, from which Biglaw can benefit—a treatment that is relatively easy to administer, already graduated from the uncertain realms of hope or scientific non-rigor, and self-adapting for the future.

I. WHO GETS HIRED BY BIGLAW?

"[M]ost interviewers [use] their own 'One Big Idea' that they believe will help them predict on-the-job success for candidates."[30] Biglaw's One Big Idea is fit. A study of EPSs by Professor Lauren A. Rivera showed that evaluators "named fit as *the most important criterion* at the job interview stage."[31] Fit can be summarized as shared values, an applicant's stable personality traits, and "similarity in play styles."[32] Fit is "perceived to be a stable personality characteristic of applicants—they either had it or they did not"[33] and "[f]irms try to minimize attrition by using fit as a selection tool."[34] One law firm hiring manager boasted to Rivera that "you can tell we were all recruited to come to [this firm] because we all have the same personalities. It's clear

---

[28] *Id.* at 27.
[29] RIVERA, *supra* note 15, at 111 (emphasis added).
[30] TARKI, *supra* note 1, at 34.
[31] RIVERA, *supra* note 15, at 136 (emphasis in original).
[32] *Id.*
[33] *Id.* at 137.
[34] *Id.* at 139.

we're all the same kind of people."[35] As the former chief talent officer at Netflix, Patty McCord stated, "[m]aking great hires is about recognizing great matches—and often they're not what you'd expect."[36]

Thus, Biglaw's current One Big Idea, fit, is the *matching of personalities*. "We think that our One Big Idea is the best predictor of future success when assessing a candidate, but why do we do this? Because none of the available methods are entirely able to predict on-the-job success, we are tempted to think that nothing works."[37] Does fit predict anything effectively? Before fully analyzing fit, the foundation for the process that leads to the consideration of fit should be explored. Fit is not the sole criterion, and despite being the *most important*, it is not considered first either. Biglaw must be more than finding personality matches, and there must also be a compelling reason to allow a single-minded focus on fit. Prior to subjectively evaluating fit, Biglaw is oddly obsessed with the exact opposite: *structured evaluations*.

Practically speaking, "structure" is accomplished by ensuring that each incident can be measured with *reliability*.[38] "Test reliability shows how consistent a measure is"[39] across multiple measurements. If a different evaluator can get the same measurement of a candidate across multiple repetitions of that measurement, then the measurement is reliable. While subjective evaluations are not necessarily unreliable (i.e., if you know a person very well to begin with), hiring generally involves people who cannot be subjectively measured because most interviews are novel interactions between strangers. Structure, however, does not guarantee *validity*. "Test validity shows the probability that . . . a variable will accurately measure what it is supposed to measure, such as how successful a candidate will be in a job."[40] Ten evaluators can ask a fully-

---

[35] *Id.*
[36] TARKI, *supra* note 1, at 118.
[37] *Id.* at 34.
[38] Structure, quantitative, and objective are not entirely synonymous, but the terms, for purposes of this Note, share a degree of interchangeability because they all collapse into the one overall idea being presented. Structure refers to consistency across evaluators which in turn indicates reliability. Generally, structured interviews have different evaluators asking the same questions rather than creating space for evaluator discretion. Quantitative refers to being able to measure data in some numerical fashion. Objective refers to the validity of data not changing between evaluators. Objectivity often requires a lack of bias. The answer to what year candidate John Doe graduated high school should not change regardless of who asks it or answers it. Meanwhile, asking if basketball is more fun to watch than hockey can produce different answers from different people that are all correct. The contrasting terms are unstructured, qualitative, and subjective, and these terms are also somewhat interchangeable for purposes of this Note.
[39] TARKI, *supra* note 1, at 172.
[40] *Id.*

grown adult candidate's height ten different times and the answer will be very reliable because it won't change. But answering one's height has nearly non-existent validity unless height can predict job performance. Height is valid in basketball, but no compelling evidence exists that taller or shorter lawyers can draft better merger agreements. This would mean that measuring height is invalid for measuring a lawyer's job performance.

Biglaw interviews have nearly nonexistent structure, or very low reliability because the interviews are not standardized.[41] While subjective evaluations may indeed be valid, the lack of reliability indicates that the measurement of personality can be improved. It becomes necessary to describe the hiring process from beginning to end and recognize structure where it does appear in order to understand why the interview stage lacks structure and is instead dominated by subjectivity.

A. The Structure in Biglaw Talent Acquisition

First-year Biglaw associates are primarily selected from on-campus interviews ("OCIs"), or during a similar season of hiring in which Biglaw firms engage most of their recruiting efforts for summer associates.[42] Students entering their second year of law school partake in OCIs and spend the following summer with the law firm, usually hoping to receive an offer to return full time after graduation.[43] This process is central to the Cravath System but is also analogous to the other EPS firms' talent acquisition models. Investment banking and consulting firms also conduct OCIs and hire the vast majority of their entry-level employees straight out of school.[44] First, the Cravath System will be explored in greater detail to explain which students even get to play this hiring game. Second, the sponsored contest, a shared phenomenon amongst EPS hiring practices will be explored to find out which students eventually get to win the game.

---

[41] RIVERA, *supra* note 15, at 124-25.
[42] *Prelaw - What Is the Timetable for Legal Recruitment?*, NAT'L ASS'N FOR L. PLACEMENT, https://www.nalp.org/pre-law_timetable (last visited Mar. 19, 2022) ("Most large law firms hire their entry-level attorneys out of their summer associate class.").
[43] *Id.* ("Not every summer hire will receive a permanent offer, but most usually do.").
[44] RIVERA, *supra* note 15, at 17.

1. The Cravath System

Initially, the Cravath System was developed by Cravath, Swaine & Moore LLP ("Cravath") because the "emphasis on credentials had a clear business purpose designed to compensate for the limitations of legal education."[45] Early on, "most law schools required little or no college education,"[46] and the Law School Admission Test ("LSAT") did not even exist until 1948. "In contrast, Harvard, Columbia, and Yale [law] grads typically had a college degree before entering law school."[47] As of 1948, nearly 70% of Cravath's associates had graduated from one of these three law schools.[48] As the landscape of the legal education system changed, the Cravath System not only kept its initial rationale but also developed new justifications to remain the preferred model for talent acquisition.[49]

"Intellectual horsepower" may be the briefest summarization of what the Cravath System seeks to secure. With few available signals of legal aptitude or competency, Cravath determined that "the inputs themselves (i.e., qualified associates) had little value to clients. Rather, they needed to be trained by the investment of intensive training."[50] Cravath would instead find graduates with the most potential to handle complex legal matters, established work habits, and a desire for growth and longevity. Although Cravath states that "[b]rilliant intellectual powers are not essential," what a brand new Cravath hire would be expected to provide was a balanced intellectual mold worthy of being crafted internally. A sound education history being one of the few available signals of such worth, college graduates who then performed sufficiently at an elite law school became desired over non-college graduates who likely attended non-elite law schools.[51]

However, law firms then and now did little to screen their applicants. Taking as axiomatic that pursuing graduates from elite law

---

[45] Bill Henderson, *Part II: How Most Law Firms Misapply the "Cravath System"*, LEGAL PRO. BLOG (July 29, 2008), https://lawprofessors.typepad.com/legal_profession/2008/07/part-ii-how-mos.html.

[46] *Id.*

[47] *Id.*

[48] *Id.*

[49] The Cravath System's design does more than act as a model for talent acquisition. The system also seeks to drive attorney development, promote sustainability, ensure lockstep compensation, protect tenure with the "up and out" partner track, internal promotions, and relationships between colleagues. However, these goals are temporally separate enough from hiring practices where they need not be explored in this Note.

[50] Henderson, *supra* note 45.

[51] *Id.*

schools (in turn inferring an undergraduate education as well) would give access to the type of mold desired, Cravath had little need for developing its own detailed criteria. Even today, "[e]valuators [believe] that 'the best and the brightest' [are] concentrated in America's most elite universities . . . Admission to an elite school [is] seen as a sign of superior 'intellectual horsepower' and well-roundedness . . . Such beliefs [lead] firms to outsource the first round of candidate screening to admissions committees at elite universities."[52] Because elite universities have already emphasized a student's high school GPA, test scores, extracurriculars, and personal statements, law schools then applying a similar process again became sufficient for law firms like Cravath to draw their associates out of the best law schools by default.

While some may entirely believe that the strength of one's legal education is largely indicative of performance on the job, this foundational belief behind the Cravath System makes even more sense for administrative ease. As long as there are enough law students at elite law schools (or gradually higher-performing students at gradually less prestigious schools), "[t]here may be really good candidates out there, but it's not worth the investment on [the firm's] part to spend a lot of resources looking for them when [they] have a very good pool that's easy to reach."[53] The "Big-fish-little-pond" effect[54] is not a groundbreaking concept anymore, and many hiring partners in Biglaw today are "firm believer[s] that you could get really good candidates from the top 5 percent of most colleges."[55] Malcolm Gladwell, in *David and Goliath*, finds that

> [t]he more elite an educational institution is, the worse students feel about their own academic abilities . . . And that feeling—as subjective and ridiculous and irrational as it may be—*matters*. How you feel about your abilities—your academic 'self-concept'—in the context of your classroom shapes your willingness to tackle challenges and

[52] RIVERA, *supra* note 15, at 36.
[53] *Id.* at 37.
[54] Krysten Crawford, *Stanford Education Study Provides New Evidence of "Big-Fish-Little-Pond" Effect on Students Globally*, STAN. GRADUATE SCH. OF EDUC. (Nov. 30, 2018), https://ed.stanford.edu/news/stanford-education-study-provides-new-evidence-big-fish-little-pond-effect-students-globally.
[55] RIVERA, *supra* note 15, at 36-37.

> finish difficult tasks. It's a crucial element in your motivation and confidence.[56]

Thus, while hiring the top student from every law school would be more fruitful than hiring all of the students from just one top school, competition between firms and availability of recruiting resources prove to be substantial barriers. The mentality adopted by firms is that "[t]he focus is on places like Harvard because it's just easier. You can go lower down in a class and still get those smart, hardworking, well-rounded people."[57] The Cravath System, still very much in effect today, continues to trust the screening filters that students have passed just to be admitted to an accredited law school, all of which now require undergraduate degrees. Screening deference is prioritized towards the most elite law schools, since they are the most competitive to get into. Meanwhile, attendance at gradually less prestigious law schools will require a student to prove to a greater degree his or her academic competence post-admittance.[58] As one attorney stated to Rivera, "I want people from Yale Law to walk through our doors. They are highly unlikely to be failing at life."[59]

In addition to the reasons stated above, firms continue to employ the Cravath System—prioritizing school prestige and law school grades above all else—for a number of other reasons. First, "firms [view] selecting new hires with prestigious academic credentials as a means of attracting clients and heightening their confidence in the firms."[60] Stated simply, *marketing matters*. Even first-year associates who have little experience in the actual practice of law are billed out at hundreds of dollars per hour and will have web profiles on the firm's site. Clients want to know that they are getting the best so listing degrees from reputable schools alongside Latin distinctions and other impressive credentials is an important marketing tool. Second and relatedly, "[r]ecruiting students from elite schools was also a means of consolidating a firm's status by developing connections with graduates who were perceived to be the future 'movers and shakers' of the world."[61] Such connections are

---

[56] MALCOLM GLADWELL, DAVID AND GOLIATH: UNDERDOGS, MISFITS, AND THE ART OF BATTLING GIANTS 80 (2013).
[57] RIVERA, *supra* note 15, at 37.
[58] *Id.* at 103 ("At super-elite campuses, grade thresholds were lower, if present at all . . . Conversely, students at less selective institutions needed to be at the top of their classes.").
[59] *Id.* at 38.
[60] *Id.* at 37.
[61] *Id.* at 38.

tied to eventually generating further business over time. Third, firms would "restrict competition to elite schools because their competition also did so."[62] Firms do not want to "leave [themselves] up for some kind of negative differentiation before the clients."[63] When it comes to talent acquisition, Biglaw has collectively adopted the Cravath System and refuses to budge for better or worse. While the Cravath system initially rewarded the 'first mover' that could gobble up elite talent straight out of law schools, Biglaw now has reasons to not drastically move first when competing for talent because of the fear that aiming for anything less than the appearance of elite leads to consequences. Biglaw talent acquisition has become an arms race for school prestige and top grades, with actual job performance as an afterthought.

Thus, the Cravath System mainly dictates who gets to play the Biglaw game at all. Students either from the best schools or with the best grades (ideally both) get their tickets punched. However, the Cravath System is not a law of nature, it is an industry practice. Explanations for who ultimately wins and why outliers exist are found within the framework of the sponsored contest.

2. Sponsored Contests

"In a *contest system*, competition is open to all; success depends on demonstrated ability . . . By contrast, in a *sponsored system*, existing elites select the winners, either directly or through third parties."[64] Biglaw hiring involves many shades of both contest and sponsored systems. Like a contest system, anyone can apply through a firm's job posting as long as they have the requisite application materials. Like a sponsored system, firms will show greater interest and dedicate the most resources to applicants partaking in OCIs from the most prestigious law schools or with referrals. Like a contest system, the barriers to entering law school are relatively low: there are no required majors, no minimum LSAT score or undergraduate GPA, no requisite prior work experience, and the total seats available across U.S. law schools are plentiful to the extent that complaints of a saturated legal job market are now common. Like a sponsored system, law firm positions historically were upper-class jobs "restricted to white, Anglo-Saxon, Protestant men from families with 'good names.'"[65] Thus, the Biglaw talent acquisition game can be

---

[62] *Id.*
[63] *Id.* at 39.
[64] *Id.* at 29 (emphasis original).
[65] *Id.* at 30.

considered a *sponsored contest*, where "[a]nyone may apply, but in reality, employees considered only those applications sponsored by existing elites: either prestigious universities or industry insiders."[66]

The Cravath System is one piece of the whole picture, albeit a significant one. An elite law school acts as an *institutional sponsor* to a candidate's Juris Doctor degree (or one in progress), and also endorses the grades earned. Generally speaking, as the prestige of a law school decreases, so does the strength of the sponsorship. What the Cravath System does not naturally capture is something that was traditionally prevalent in the legal industry and still is today: *individual sponsorship*. Although individual sponsors can gradually sponsor an institution rather than a student (i.e., "new or less prestigious schools could be put on the list [of target schools] if the firm had high-ranking employees who were graduates and pushed the firm to recruit from their alma mater"),[67] much of individual sponsorship takes the form of a personal relationship. "In many firms . . . an application from a student at a [less prestigious] institution was discarded without review unless the applicant had an individual sponsor . . . ."[68] In order to be considered as an applicant without fitting the default criteria of the Cravath System, "[y]ou need to know someone, you need to have a connection, you need to get someone to raise their hand and say, 'Let's bring this candidate in.'"[69] An individual sponsor can then be understood as "a person in a firm who would vouch for [an applicant] and push their application into the consideration set."[70]

There are three dominant hypotheses for why individual sponsorships work. "Each of these theories presents the value of referrals as stemming from employers' rational calculations about what makes a more productive worker and workforce."[71] The *better match hypothesis* states that "because existing employees know important information about the formal and informal demands of jobs, they may bring forward applicants who are a better fit with job requirements than those acquired through less personalized sources."[72] This hypothesis seems at least plausible, since law firms seem to recognize that their default metric, the Cravath System, may not provide results reliable enough to capture exceptions to the rule. The second is a *richer pool hypothesis*, which

---

[66] *Id.* at 30.
[67] *Id.* at 32.
[68] *Id.* at 35.
[69] *Id.*
[70] *Id.* at 48.
[71] *Id.* at 49.
[72] *Id.*

states that applicants presented through referrals are more appropriate based on screening requirements. This hypothesis seems unlikely since (1) the Cravath System appears to capture those screening requirements quite well already, and (2) Rivera's studies showed that "[r]eferred applicants usually were atypical; referrals compensated for candidates' lack of desirable and easily observable qualifications."[73] Stated simply, good qualifications do not need referrals. The third hypothesis is the *social enrichment hypothesis*, that "preexisting ties . . . can enhance on-the-job training, satisfaction, or mentoring."[74] Social enrichment seems plausible in many instances but is far from the rule. The power of individual sponsorship is not limited to the hiring of associates that will directly work in the same team or office as the sponsor. Social enrichment may very well be the case in some sponsorships but not in others.

The forms of individual sponsorship commonly fall into a few categories. First, a "sponsoring employee would directly deliver the job seeker's application (in person or via email) and draw attention to it."[75] Since a firm's first line of evaluators often ignore resumes and applications that do not seem desirable according to the Cravath System, individually sponsored applicants would instead receive an express lane to review (i.e., consideration for interviews). Second and third, "[d]ue to internal and external power dynamics, the referrals of senior employees and clients carr[y] great weight."[76] "A senior employee . . . could push through an applicant to the interview stage for any reason, even a personal whim regardless of the quality of the candidate's resume," while *high-tough referrals* (referrals from clients or judges) "were widely seen as 'business development activity,'"[77] and would also secure a first-round interview though usually not more.

Here is where the "structure," or objective portion of Biglaw talent acquisition ends. Admittedly, there is a lot that has gone into it by now, but considerations that can be compared by numbers or answered in a reliable yes/no fashion do not systematically exist beyond this point. Although it is worth investigating the validity of a school's rank,[78] a rank

---

[73] *Id.*
[74] *Id.*
[75] *Id.* at 50.
[76] *Id.* at 52.
[77] *Id.*
[78] Perceptions of which schools are more or less prestigious is derived not from any one official ranking but general perceptions, historical relevance and longevity, ranking reports. RIVERA, *supra* note 15, at 32 ("Firms commonly made their school selections based on general perceptions of . . . institutions' prestige . . . . In addition, firms used the reports of external ranking organizations such as U.S. News and World

can be listed in a reliable manner. A school either is or is not ranked higher than another. GPA or class rank is also objective and reliable because a candidate's answer will not change from one evaluator to the next. Individual sponsorship is also objective despite its fluid and arbitrary appearance. Competing for individual sponsorships is the primary goal of professional networking for an active job-seeker and although luck, preexisting personal relationships, and subjective judgment may all affect whether a candidate has an individual sponsor, there is a structure for law firms because it is an objective and rather simple inquiry: "do you have a sponsor and who is it?"

The winners of the Cravath System compounded with individual sponsorships are not those who ultimately get the job. At the interview stage, a brand new process shows face, and it uses heavily unstructured criteria. Even for first-round interviews (which are shorter but not procedurally different than second/final round interviews), many law firms choose to shift gears to subjective evaluation and sometimes entirely ignore who the better candidate was at the structured level.

B. Unstructured Interviews: Evaluation Designed for Human Error

"[I]f something feels as if it should work, many of us convince ourselves that it does."[79] Biglaw talent acquisition is no exception. Evaluators in Biglaw interviews are often given no significant instructions other than for presentation's sake (e.g., don't cut interviews short, don't take notes so that it feels more like a conversation, and don't forget to respond to thank you emails).[80] Rivera's insider experience at an EPS firm's training for evaluators showed that subjectivity was not only acknowledged but even endorsed. Instructions were the likes of "[i]f someone bothers you, don't let them go forward,"[81] and "[w]e trust your judgment. You'll get a sense of the whole candidate."[82]

It sounds odd that even if the interview is unstructured, evaluators would suddenly remove objective criteria entirely. If a candidate came

---

Report and the Law Schools Admissions Council."). In addition, while the most prestigious schools are desired in just about any Biglaw firm, more local schools to a specific office tend to be given more consideration, in part because having a candidate remaining in the office for many years is desired and having local ties serves as evidence of it. Remaining in the same firm for many years to eventually become an internally-developed capable attorney is also part of the Cravath System.

[79] TARKI, *supra* note 1, at 31.

[80] RIVERA, *supra* note 15, at 115, 117.

[81] *Id.* at 116.

[82] *Id.* at 117.

from an elite school with top grades, shouldn't that candidate's pedigree be weighted during final consideration for the position? If a candidate was sponsored, shouldn't the sponsorship mean something beyond an invitation to interview? The lack of harmony between the two criteria sets at the resume and interview stages is a head-scratcher for sure. "Evaluators believed that merit was best assessed by evaluating 'the person' not 'the paper' and they did not trust resumes to reliably predict job performance."[83] If evaluators do not care for a school's prestige and the candidate's grades—despite heavily screening for them earlier—and do not believe that they serve as evidence of merit or job performance, then it must mean that both merit and job performance predictors can be observed in an interview, the only other stage before final decisions for job offers are made. However, Biglaw does not conduct case-based interviews, behavioral questions, or any particular kind of filter for competency at all. Biglaw believes that interviewing requires no formal training and instead relies on common sense to have "just a conversation."[84] Because there are no detailed guiding principles for the Biglaw interview, it lacks consistency across multiple evaluations and is unstructured to the point where human error in judgment appears to be invited rather than guarded against.[85] One law firm hiring manager told Rivera that "[o]ur attorneys bring their own styles to interviews. . . . We trust their instincts."[86] This means that the full arsenal of human biases is welcome in making final hiring decisions.[87] A biased result is one that

---

[83] *Id.* at 118.
[84] *Id.* at 123.
[85] *Job Interviews Don't Work*, FARNAM ST., https://fs.blog/job-interviews/ (Last visited Jan. 29, 2022).
[86] RIVERA, *supra* note 15, at 124.
[87] One rather obvious problem in addition to bias is the possibility for reinforcement of any discriminatory outcomes resulting from existing hiring practices. However, discrimination is an issue to be addressed separately from this Note. I assume that firms do not want to change what they want in their talent, and implicit in that assumption is that what law firms want in their talent is/should be legal (i.e., non-discriminatory). There is no argument underlying this Note that any existing or potential discriminatory outcomes should be permitted or reinforced as a consequence of psychometrics or artificial intelligence having greater presence in talent acquisition. Discriminatory outcomes attributable to the use of personality profiles or machine learning should be stress-tested for and addressed with great attention as any hiring process or criteria should be. If a discriminatory outcome is suspected, taking a step back from implementation in order to assess whether implementation was the cause or revealer of such outcomes should be one of the first questions asked. With that in mind, I will briefly mention that I expect that the FFM can be validated using criterion-related validation in a disparate impact suit. "Of the three methods of validation, criterion-related validation is the only one which correlates tests results with actual work performance and is thus considered preferable to methods based on less direct evidence." JOEL WM. FRIEDMAN, THE LAW OF EMPLOYMENT DISCRIMINATION: CASES AND MATERIALS 320, 321 (13th ed. 2020). Criterion-related validation is in fact

is "systematically off target,"[88] and, in talent acquisition, off target means that the best candidates are getting overlooked.

There is a myriad of potential biases that can enter the scene when making an evaluative judgment about another person. Keep in mind that a law firm interview's One Big Idea is fit. The claim in this Note is not that injecting structured psychometrics into evaluations will change the criteria, but rather that final determinations of the same fit that is currently assessed will be less biased—less systematically off target. "[F]or the purpose of evaluating the quality of an employer's judgments when selecting employees, it seems reasonable to use the judgments that the same employer makes when evaluating the employees thus hired."[89] If Biglaw wants to interview for the best fit, then we should be assessing whether or not bias in interviews affects fit.

> Interviews are also a minefield of psychological biases. In recent years, people have become well aware that interviewers tend, often unintentionally, to favor candidates who are culturally similar to them or with whom they have something in common, including gender, race, and educational background. Many companies now recognize the risks posed by biases and try to address them through specific training of recruiting professionals and other employees.[90]

With Biglaw providing little to no training and believing that effective interviewing requires little more than common sense and intuition, it is no surprise that many biases become fully expressed. This is already a pervasive issue for any one given pair of evaluator and interviewee. However, "[d]ifferent interviewers respond differently to the same candidate and reach different conclusions."[91] While there is some correction against biases when interviewing the same candidate multiple times (as is often the case in a second-round, or "callback" interview), first-round interviews, or "screeners," are often conducted by only one

---

one way of describing the very machine learning implementation process that I later introduce, Part III, *infra*, and fits well with the concepts of correlation and factor analysis that I later introduced, Part II, Section A.1, *infra*. Criterion-related validity "should consist of empirical data demonstration that the selection procedure is predictive of or *significantly correlated* with important elements of job performance." 29 C.F.R. § 1607.5 (emphasis added).

[88] DANIEL KAHNEMAN ET AL., NOISE: A FLAW IN HUMAN JUDGMENT 4 (2021).

[89] *Id.* at 302.

[90] *Id.* at 303.

[91] *Id.*

revenue generating employee or a human resources staff member. A little positive luck can go a long way if a candidate ends up being evaluated by someone who has biases that work favorably for said candidate. A little negative luck, however, may foreclose an otherwise excellent match between candidate and firm.

Surely, a law firm would not define an ideal candidate to be one who was fortunate enough not to bump into evaluators that had biases against them, but rather a true fit. A complete evaluation of fit is inevitably going to require a subjective component and the subjective component will inherently be riddled with biases. This is why the unstructured evaluation of fit should be complemented (rather than entirely replaced) with quantitative assessments of fit. All it would take for a purely objective talent acquisition model to fall apart is one instance of a new hire, who was entirely decided based on structured measurements of fit, to perform poorly. "The strengths of quantitative methods are that you can measure, standardize, and replicate many of the outcomes. The strengths of qualitative methods are the richness and depth of the insights . . . *both methods should be used as complementary tools when assessing candidates*."[92]

II. PERSONALITY PSYCHOLOGY IN THE WORKFORCE

A. Measurement: The Five-Factor Model

Personality, or "the general psychology of individual differences,"[93] is an admittedly strange subject with an obscure history. The *Five-Factor Model* ("FFM") of personality failed to launch in the mid-1930s despite coming from Louis Thurstone, a "U.S. pioneer in psychometrics."[94] In Thurstone's Presidential Address for a meeting of the American Psychological Association, Thurstone remarked that "[i]t is of considerable psychological interest to know that the whole list of sixty adjectives can be accounted for by postulating only five independently common factors."[95] Thurstone had subjects use sixty adjectives to describe close acquaintances. At this time, the statistical discovery of a personality factor was no eureka moment, since many factors or sub-factors had been discovered in earlier models: the General Factor of

---

[92] TARKI, *supra* note 1, at xv (emphasis added).
[93] THE FIVE-FACTOR MODEL OF PERSONALITY vii (Jerry S. Wiggins ed., Guilford Publications 1996).
[94] *See L. L. Thurstone*, NEW WORLD ENCYCLOPEDIA, https://www.newworldencyclopedia.org/entry/L._L._Thurstone.
[95] *Id.* at 1; *see also* note 127, *infra*.

Intelligence (g) by Spearman in 1904, "will" by Webb in 1915, and "cleverness" by Garnett in 1919.[96] These factors appear to be rough sketches of what eventually became Openness, Conscientiousness, and Extraversion respectively. Oddly enough, the chronology of discovery roughly aligns with the popular mnemonic for the FFM, 'O.C.E.A.N.' where Agreeableness and Neuroticism round out the model. The FFM is also popularly known as the Big Five.

The FFM did not maintain a singular form from the 1930s to modern-day. "Until recent times . . . the psychometric approach to the essential dimensionality of personality constructs had failed to produce a generally accepted model."[97] Having undergone transformations ranging from three to even ten or more factors, personality appeared to be outside the reach of precision for many decades. However, what is most important is that some variation of a multi-factor model persisted. "Personality psychology rediscovered the five-factor model in the 1980s"[98] when findings about the statistical model revealed, somewhat reluctantly, that "five-factor solutions were remarkably stable across studies, whereas more complex solutions were not."[99] Through the past few decades, the reliability of the FFM has been established with greater scientific rigor, empowered finally by a widespread acceptance within the clinical psychology field, at least for the number of primary factors in a personality model. The FFM, though far from a complete theory of personality, has shown robustness across cultures, media, age groups, and evolution.[100]

The methodology of the FMM is of particular importance for understanding its reliability and validity. "In his *Nicomachean Ethics*, Aristotle attempted to provide [a] map for human 'character' traits, and since his time, others have tried similar mappings."[101] However, mapping such characteristics requires solving two scientific problems: "(1) a procedure for sampling human attributes, and (2) a method for structuring that sample of attributes."[102] The lexical hypothesis and factor analysis, respectively, address those problems when it comes to personality. "[M]odern science aims to obtain new knowledge . . . by gathering observations and then using mathematical tools to connect

---

[96] FIVE-FACTOR MODEL OF PERSONALITY, *supra* note 93, at 2.
[97] *Id.* at 12.
[98] *Id.*
[99] *Id.* at 13.
[100] *Id.* at 16.
[101] *Id.* at 22.
[102] *Id.*

these observations into comprehensive theories."[103] Isaac Newton, in *The Mathematical Principles of Natural Philosophy*, "showed that the book of nature is written in the language of mathematics."[104] Statistics is used in fields of science, like psychology, that are too complex to speak of solely using the language of mathematics. Oftentimes in university psychology curricula, one of the first—if not the very first—required courses is an introductory course in statistics and methodology. A brief, but detailed look into both the lexical hypothesis and factor analysis will be useful in order to apply personality traits—each a product of phenotypical observations and statistics—to Biglaw talent acquisition. Doubt over theoretical perspectives or methodology should be properly addressed first, or else objective results post-implementation that are undesired will be easily cast aside, further propagating an imbalance in favor of subjective hiring.[105] Francis Bacon, in *The New Instrument*, argued that knowledge is power. "The real test of 'knowledge' is not whether it is true but whether it empowers us . . . . Consequently, truth is a poor test for knowledge. The real test is utility."[106] After describing its methodology sufficiently to cast aside common levels of doubt ("[m]ost people have a hard time digesting modern science because its mathematical language is difficult for our minds to grasp, and its findings often contradict common sense"[107]), the real power of the FFM—utility of understanding the attributes for Biglaw talent acquisition—will be explored.

---

[103] YUVAL NOAH HARARI, SAPIENS: A BRIEF HISTORY OF HUMANKIND 251 (2015).
[104] *Id.* at 256.
[105] One reason why it is critical to defend the reliability and validity of the FFM is because much of pop personality psychology is riddled with "appealing fictions." In science, particularly statistics-driven sciences like clinical psychology, measurement is what distinguishes "real" from "not yet real." For example, the four learning styles (visual, auditory, reading and writing, and kinesthetic), though seemingly plainly observed, have not yet been measured in social psychology and thus have not been made "real." This does not mean that learning styles do not exist, but rather that learning styles so far have failed to manifest on a scientific level and continue to be an unproven hypothesis if left as is. Entertaining the FFM with room to question its reliability or validity, like the learning styles, especially at the implementation stage is a quick path to ensuring that confirmation bias continues to run rampant in talent acquisition. *See e.g.*, Harold Pasher et al., *Learning Styles: Concepts and Evidence*, 9 PSYCH. SCI. IN PUB. INT. 105 (2009); Cedar R. Reiner & Daniel Willingham, *The Myth of Learning Styles*, 42 CHANGE: THE MAG. OF HIGHER LEARNING Issue 5, 34 (2010); Beth A. Rogowsky et al., *Matching Learning Style to Instructional Method: Effects on Comprehension*, 107 J. OF EDUC. PSYCH. 64, 65 (2015); William Furey, *The Stubborn Myth of "Learning Styles"*, 20 EDUC. NEXT Summer 2020 8, 9.
[106] HARARI, *supra* note 103, at 259.
[107] *Id.*

1. Methodology

"Over the years, a number of philosophers and linguists have remarked about the 'wisdom' embedded in natural languages."[108] The *lexical hypothesis* draws on this wisdom and, as language philosopher J.L. Austin notes, "our common stock of words embodies all the distinctions men have found worth drawing."[109] Under the lexical hypothesis, the "common stock of words" (i.e., everyday language) is a complete collection from which descriptions of individual differences can be acquired. Psychologist Raymond Cattell stated that "[t]he position we shall adopt is a very direct one . . . making only the assumption that all aspects of human personality which are or have been of importance, interest, or utility have already become recorded in the substance of language."[110] Further, although linguistic theory and everyday language will change,[111] the appearance of new terms is balanced out by the obsolescence of old terms, and this concern is further mitigated when considering that "[a]lthough faddish terms appear and disappear . . . within decades, the overall framework of language is comparatively conservative . . . and most personality terms have been used in a recognizably similar way for centuries."[112]

To better embrace the lexical hypothesis, a few axiomatic propositions should be elaborated on. First, "personality language refers to phenotypes and not genotypes."[113] Thus, personality consists of mere linguistic observations and are not intended to explain why individual differences are the way they are at a biological level. Second, personality traits, although they may be traits, are more cautiously described as "attributes." Although I will frequently refer to both traits and attributes interchangeably throughout this Note, the distinction takes presence over the nomenclature. Traits "are relatively stable over time and across situations. The lexical perspective itself does not require these assumptions."[114] Third, a combination of two propositions are maintained, that "[t]he more important is an individual difference in human transactions, the more languages will have a term for it,"

---

[108] FIVE-FACTOR MODEL OF PERSONALITY, *supra* note 93, at 22.
[109] *Id.* at 22.
[110] *Id.* at 23.
[111] ROGER BROWN, PSYCHOLINGUISTICS: SELECTED PAPERS ix (1972) ("The fact that linguistic theory changes, and does at a rapid clip, poses real difficulties for the psychologist who wants to use linguistic theory in his own work.").
[112] FIVE-FACTOR MODEL OF PERSONALITY, *supra* note 93, at 28.
[113] *Id.* at 24.
[114] *Id.* at 25.

alongside that "the more important is such an attribute, the more synonyms and subtly distinctive facets of the attribute will be found *within any one language*."[115] These two propositions are respectively the *across-language form* and the *within-language form*. Unsurprisingly, language also follows a power-law distribution, where the vast majority of linguistic communication is accomplished by a small percentage of existing words. In the English language, roughly 80% or more of speech is accomplished using less than a thousand of the most common words.[116] This linguistic phenomenon is strong evidence of both the across-language and within-language forms. Fourth, the *adjective function*, whether carried out by actual adjectives (e.g., he is unorthodox) or other words like nouns or verbs (e.g., she is a maverick), serves as "the central repositories of the sedimentation of important individual differences into the natural language."[117] Although the FFM using the English language relies primarily on adjectives (as most languages do), potential variations must be considered when comparing across languages. Fifth, the lexical hypothesis draws strength, not weakness, from the usage of single words instead of phrases and sentences. "[S]ingle terms often function holophrastically; that is, they can incorporate complex ideas that are normally expressed in sentences."[118] Describing oneself as courageous bypasses with little leakage of meaning the excess words in the sentence "I believe that I am courageous," all while dodging added ambiguities from attaching additional descriptive words (e.g., "willingly courageous"). Finally, the lexical hypothesis requires that "[t]he most important dimensions in . . . personality judgments are the most invariant and universal dimensions."[119] "A robust and replicable factor solution is one that is so clear and strong that the choice of analytic method becomes unimportant,"[120] and this becomes particularly important when recognizing consistent results despite variances in language, culture, and reporting environments. The FFM is grounded upon these axiomatic propositions that allow it to be considered a valid, reliable, and universal scientific approach to individual differences.

---

[115] *Id.* at 26 (emphasis in original).
[116] 1000 MOST COMMON WORDS, https://1000mostcommonwords.com/ (last visited Mar. 19, 2022) ("Language learning, like most things in life, follow the Pareto principle. It's been said that the top 1,000 most frequent words in a language make up over 80% of the speech.").
[117] FIVE-FACTOR MODEL OF PERSONALITY, *supra* note 103, at 30-31.
[118] *Id.* at 32.
[119] *Id.* at 35.
[120] *Id.*

Following the lexical hypothesis and just as crucial to the FFM is *factor analysis*. Uncharacteristic for most sciences (if that has not been established already), the FFM came about in an atheoretical manner. In a gross over-summarization, the scientific method hypothesizes a reality and then tests it with control and variable groups. The FFM was more or less "discovered" by utilizing the lexical hypothesis, but no model or even a preferred number of factors was hypothesized in advance. Psychologists did not test to see if "Agreeableness" or "Neuroticism" would be personality traits. The five factors (Extraversion, Openness, Agreeableness, Conscientiousness, and Neuroticism) could have easily been named I, II, III, IV, and V to point to the factors that were discovered. Psychologists simply pushed human language through a statistical process and accepted what came out of the other end.

"[F]actor analysis summarizes the relations between many variables by expressing each variable as some unique combination of a few basic dimensions, known as factors."[121] A deeply technical and mathematical understanding of factor analysis is unnecessary, and an illustration of factor analysis for the trait Neuroticism might suffice. Neuroticism, which will be fully detailed below, deals in part with anxiety.[122] If a set of one hundred questions regarding words with an adjective function were given to many self-reporting individuals, one can expect to find that after accumulating sufficient reports, patterns begin to surface. If many people answered affirmatively to three of the hundred words "fearful," "worrisome," and "nervous," then a *cluster* begins to form. A cluster indicates that if a person gives a particular response to a word, he or she is likely to also give a similar response to another word within a group to which that word belongs.[123] Clusters are not binary with strict boundaries but will instead be formed out of meaningful *correlations*. Using arbitrary numbers to illustrate, "fearful" and "worrisome" may have a correlation of 0.7, meaning that 70% of affirmative answers for one will also affirmatively answer for the other. "Fearful" and "confident" may have a correlation of 0.15, meaning that only 15% of responses for one match responses for the other. If a correlation is 1.0, it means that 100% of the responses for one word match the other word's responses and this indicates that *the two are the*

---

[121] HANDBOOK OF RESEARCH METHODS IN PERSONALITY PSYCHOLOGY 424 (Richard W. Robins et al. eds., Guilford Publications 2009).
[122] *See* II.A.2.e, *infra*.
[123] *See* Bernard S. Gorman, *The Complementary Use of Cluster and Factor Analysis Methods*, 51 J. OF EXPERIMENTAL EDUC. 165 (1983) ("[C]luster analysis aims primarily to provide relatively homogeneous groups of subjects and/or variables on the basis of one or more multivariate similarity criteria.").

same thing. In psychometrics, a correlation of 0.3 or greater is usually considered significant in that the two items are meaningfully related,[124] and that *something underlying about the two items must be the same*. Thus, if "fearful," "worrisome," "nervous," and a whole host of other words begin to cluster due to their meaningful correlation, a psychologist may conclude that all of these words point to some broader underlying idea and can then title that idea "Neuroticism."

By presenting the common stock of adjective-functioning common language, across languages, to a very large number of self-reporting participants, five clusters were extracted. Five broader ideas, dimensions, or factors of human personality appeared. The Law of Large Numbers is the principle that "while it might be difficult to predict with certainty a single event . . . it [is] possible to predict with great accuracy the average outcome of many similar events."[125] Each of these factors do not definitively speak to any one individual's proclivities in any particular circumstance, but it does speak both reliably and validly about how a person high in Neuroticism tends to behave across an aggregate of circumstances over time. In fact, one might move the goalpost and pick another level of what a meaningful correlation is, or even check for correlation among discovered clusters. By doing so, two broader factors of personality—Plasticity and Stability—were derived alongside two aspects (i.e., sub-traits) for each of the five main traits.[126] No trait exists in a bubble, free of any correlation from other traits, and it should not even have to be mentioned that a person is the product of all of his or her personality traits acting in unison, creating a harmony of individual differences that ceaselessly manifests and adapts itself in new circumstances.

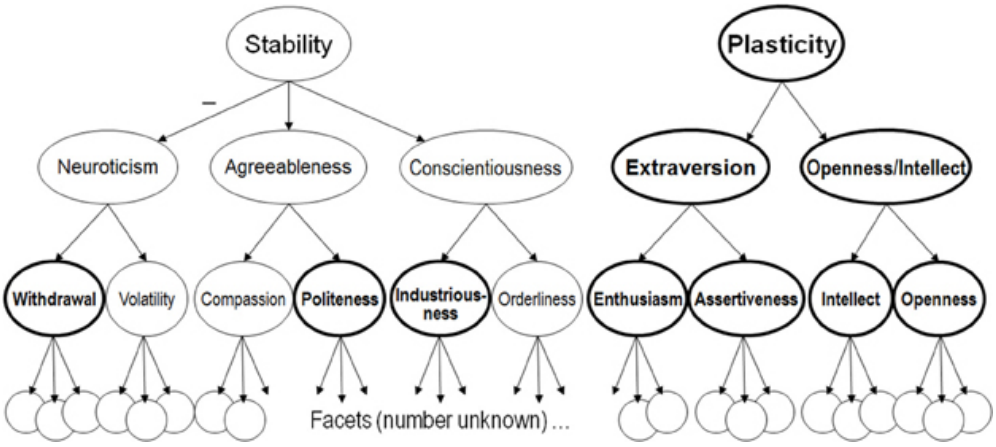
Although the field of statistics can offer even greater and more detailed insights, that is a job best left for the statisticians and computers at the stage of implementation. It is, for now, sufficient to provide the tools for a Biglaw firm to select their data to measure and understand the elementary insights provided by the FFM. For example, if meeting billable hours requirements and being in the upper quartile of Conscientiousness positively correlates by 0.8, then it brings into perspective that approximately 64% of the variance (correlation coefficient squared) for meeting billable hours can be explained by being

---

[124] TARKI, *supra* note 1, at 36.
[125] HARARI, *supra* note 103, at 256-57.
[126] Gregory J. Feist, *Creativity and the Big Two Model of Personality: Plasticity and Stability*, 27 CURRENT OP. IN BEHAV. SCI. 31, 31 (2019) ("[T]he five personality dimensions do not seem to be completely independent of each other and hence are not the highest level in the hierarchy of personality.").

in the upper quartile of Conscientiousness. Such statistical insights are far from becoming the Bible of talent acquisition, but it is unlikely that firms will be harmed by adopting them as calibrators to better recognize talent and ensure that the best candidates are not slipping through the cracks.

2. The Personality Attributes



The FFM, though nominally focused on the number five, has a hierarchy structure within itself that can lend itself to two, five, or ten depending on the level of analysis. All five attributes can be grouped into Plasticity (Extraversion and Openness) and Stability (Agreeableness, Conscientiousness, and Neuroticism).[127] Each of the five attributes further splits into two aspects (e.g., Conscientiousness consists of sub-clusters Orderliness and Industriousness). Thus, factor analysis allows for someone to endorse either a "Big Two" model or even a "Big Ten" model if desired. Starting with the two attributes that make up Plasticity and ending with the three attributes that make up Stability, each of the five main attributes (and their aspects) will be described. In doing so, implementation can be considered in light of multiple levels of analysis. For example, a personality profile that exhibits just high or low Plasticity and high or low Stability is a far less complex personality profile than one that balances five or even ten attributes. Going deeper is possible with even smaller factors, but even "factor analysis researchers are often

[127] Although Stability and Plasticity exist from the same atheoretical factor-analysis process that the five main traits were borne out of, psychologists speculate that the two largest clusters form because Stability is related to serotonin-producing experiences and Plasticity is related to dopamine-producing experiences. *See generally*, Feist, *supra* note 126.

plagued by the problem of choosing an adequate number of factors,"[128] and it may be imprudent to explore beyond what is commonly researched. Despite the temptation to systemize and break models into more definite components, effective implementation may come at the cost of simplicity, and providing a menu for levels of analysis allows both broad and detailed implementation as desired.

Emphasizing again that the FFM is a model of observation rather than explanation, how we choose to view personality attributes can shape how we frame their utility in the real world. Each attribute can be viewed as a sub-personality or a statistically meaningful cut-out from one overall personality. Each attribute can also be viewed as a frame of reference. To illustrate with a few gross oversimplifications, neurotic people frame the world as a place of possible threats, open people frame the world as a place to engage with abstraction, and conscientious people frame the world as a place to work and organize. If attributes serve as frames of reference, they mold and shape a person's perceptions. Attributes may also be viewed as value and goal setters. The extraverted person may value, and thus set goals to form relationships with new people. The agreeable person may value, and thus sets goals to alleviate conflict and encourage collaboration. If the purpose of a scientific endeavor is to comprehend and utilize, then it is both necessary and practical to think about the utility of attributes both while they are measured and implemented.

a. Extraversion

The first trait within Plasticity, Extraversion, is one that is well known to the point that it has pervaded popular culture and is a staple even in widespread personality models that lack statistical rigor (e.g., the Myers-Briggs Type Indicator). "Extraversion describes active people who are sociable, talkative, and assertive."[129] Although Extraversion is generally perceived as a social trait (i.e., manifests itself because of the presence or absence of other people), it has also been described as reward sensitivity, in which social situations tend to induce the kinds of rewards that people, inherently social creatures, are sensitive to.[130] Although

---

[128] Gorman, *supra* note 123, at 166.
[129] Kira O. McCabe & William Fleeson, *What is Extraversion For? Integrating Trait and Motivational Perspectives and Identifying the Purpose of Extraversion*, 23 PSYCHOL. SCI. 1498, 1500 (2012).
[130] *Sensitivity to Rewards May Distinguish Extraverts From Introverts Rather Than Higher Sociability, According to New Study*, AM. PSYCHOL ASS'N (2000), https://www.apa.org/news/press/releases/2000/09/extraverts.

"[t]he specific subcomponents of extraversion are debated."[131] Assertiveness and Enthusiasm are the two most established aspects of Extraversion. "One area of agreement among most researchers is that extraversion is related to positive affect . . . the relationship between extraversion and positive affect holds up even within individuals, such that people experience more positive affect when they act in an extraverted manner than when they act in an introverted manner."[132] What was notable for considering the Extraversion of a professional in the workplace was that "[p]ositive affect can be viewed as a proxy for goal achievement—people pursuing our hypothesized goals should show increases in state extraversion, and increases in state extraversion should lead to increased positive affect."[133] Goal achievement here refers to setting and pursuing new career-related goals, not necessarily career success or satisfaction.

b. Openness

The second trait in Plasticity is a rather interesting one, and often controversial. Openness to Experience ("Openness") splits into the aspects of Openness Proper ("Creativity") and Intellect. Here is where the atheoretical personality model begins to understandably raise doubt into the minds of non-statisticians. Creativity is not conventionally seen as a personality attribute, and neither is Intellect, which is essentially one's intelligence quotient ("IQ"). Keep in mind that all factors overlap to a degree, so sub-traits are sub-traits exactly for that reason—Creativity and Intellect significantly correlate into Openness. Openness can be roughly summarized as a facility with ideas and experiences. More broadly defined, "[o]penness to experience refers to the extent to which a person actively seeks and appreciates different experiences and tolerates and explores novel situations."[134] Although Openness is consistently associated with all measures of creativity and is thus reliable, it may not be causal. "Openness to experience might not directly cause creativity, but it serves as a 'catalyst' for the expression and exploration of creative ideas and activities."[135] Creativity can be measured as either a proclivity to engage in divergent thinking or by the accumulation of

---

[131] McCabe, *supra* note 130, at 1500.
[132] *Id.*
[133] *Id.* at 1501.
[134] Baoguo Shi et al., *Openness to Experience as a Moderator of the Relationship Between Intelligence and Creative Thinking: A Study of Chinese Children in Urban and Rural Areas*, 7 FRONTIERS IN PSYCHOL. 1, 1 (2016).
[135] *Id.* at 2.

creative achievements (e.g., musical compositions, publications, films acted in, etc.). In either case, something creative must simultaneously be both novel and useful. Creativity can also be broken down into *fluency* (how many ideas one produces) and *originality* (how improbable the ideas are to be produced by others).

Intellect, however, is where things begin to get dicey. "Openness to experience often shows positive associations with IQ test performance . . . intelligence and creativity are positively correlated to a point . . . but the correlation becomes trivial or non-existent above the threshold."[136] One tentative conclusion relating Creativity with Intellect is that creative endeavors will often require sufficient Intellect to play with the ideas and concepts, but more Intellect beyond that threshold does not always ensure greater Creativity. There is also a diminishing returns hypothesis instead of a threshold hypothesis. The most controversial part about Intellect, however, is not about its relation to Creativity. IQ, having come about a similar, if not same factor analysis process as the rest of personality and much of social sciences at large, is extremely reliable but heavily scrutinized and criticized for its validity.[137] Admittedly, intelligence of any kind as a component of personality is an uncomfortable finding to many. Clinical psychologists swear by its reliability, but the competition is thin if any other type of intelligence becomes overshadowed by IQ. If there were more than one type of intelligence, IQ merely being one of them, then there should be a range of correlations between IQ and the other intelligence type, but meaningfully diverse correlations have not yet been found.[138] In any event, the validity of IQ, like the validity of any scientific phenomenon, should continue to be investigated so that it is further strengthened or challenged in search for the truth.

c. Agreeableness

The first of three traits within Stability is Agreeableness, or the trait that describes "individuals [who] generally engage in less quarrelsome behavior and more cooperative behavior in daily life . . . agreeable individuals exhibit a preference for more socially adaptive

---

[136] *Id.*
[137] *See generally* Daphne Martschenko, *The IQ Test Wars: Why Screening for Intelligence is Still So Controversial*, THE CONVERSATION (Oct. 10, 2017), https://theconversationotecom/the-iq-test-wars-why-screening-for-intelligence-is-still-so-controversial-81428.
[138] *See* note 158, *infra*.

modes of conflict resolution."[139] Agreeableness has been hypothesized to have been evolutionarily selected for the proper care of infants or pair-bonding and disagreeableness for purposes of inter-community conflict and tribalism. Agreeableness further has been linked with effortful control (i.e., self-control against intrapsychic urges) which has an inverse relationship with anger and aggression.[140] Agreeableness has also been hypothesized to "describe [a] general tendency to be altruistic," and an unwillingness to be exploitative of others.[141]

Politeness and Compassion are the two aspects of Agreeableness. Politeness is the "tendency to be respectful of others and to suppress aggressive, norm-violating impulses," while Compassion is "the tendency to be emotionally concerned about others."[142] It is more obvious, at least compared to Openness, to see that the two aspects would belong together and correlate into one larger trait. While some may conflate the two aspects in their day-to-day lives (e.g., rude people do not appear to concern themselves with the emotions of others), the two are in fact distinguishable. Because Politeness and Compassion can be interpreted differently by people according to their culture, values, priorities, and even language, one person might see 'telling the hard truth' as both polite (lying is disrespectful) and compassionate (the truth is for your own good), while the recipient may perceive it as both impolite and uncompassionate.

Agreeableness should not be conflated with empathy. Empathy is a term thrown around liberally and has its own arena of common confusion in the social sciences, but it should be noted that while there are agreeable people who are empathetic, they are not necessarily so (e.g., a salesperson must be agreeable but need not be empathetic to customers). Empathy will not be further discussed in this Note below, but out of caution for those who perceive testing for Agreeableness as testing for empathy, it should be noted that the two are not the same. Sometimes, this conflation has served organizations well. Companies and professions like the medical field have begun incorporating empathy training when in reality they are commonly developing agreeable workforces because they value the appearance of Compassion and

---

[139] Scott Ode & Michael D. Robinson, *Agreeableness and the Self-Regulation of Negative Affect: Findings Involving the Neuroticism/Somatic Distress Relationship*, 43 PERSONALITY & INDIVIDUAL DIFFERENCES 2137, 2138 (2007).
[140] *Id.*
[141] Kun Zhao et. al., *Politeness and Compassion Differentially Predict Adherence to Fairness Norms and Interventions to Norm Violations in Economic Games*, 7 FRONTIERS IN SCI. May 2016, at 1, 2.
[142] *Id.*

Politeness. Empathy is not a personality component to increase, nor is it a skill that can be taught easily for a professional environment.

d. Conscientiousness

Conscientiousness is probably the most impactful trait for Biglaw talent acquisition to focus on. Conscientiousness is "the propensity to follow socially prescribed norms for impulse control, to be goal-directed, planful, able to delay gratification, and to follow norms and rules."[143] The number of potential aspects for Conscientiousness numbers up to seven, the two most recognized being Industriousness and Orderliness, but with room to further recognize impulse control, reliability, conventionality, virtue, and decisiveness.[144] Industriousness is "the tendency to stay focused and to pursue goals in a determined way" whereas Orderliness is "the preference for routines, deliberation, and detail-orientation."[145] Orderliness and Industriousness might cluster because orderly people need to put in work to keep their lives ordered and that automatically sets a temperamental goal that one pursues by working towards.

Conscientious people are better oriented toward long-term planning and delay gratification. Conscientiousness is also positively correlated with self-reported overall life satisfaction. Because there is always work to be done in life, industrious people especially enjoy working, and working usually improves one's life rather than destroys it (i.e., earning and saving money to build wealth), it makes sense that Conscientiousness is positively correlated with overall life satisfaction and serves serotonergic functions well by actively resisting chaos.

e. Neuroticism

Neuroticism is defined as "the tendency to experience frequent and intense negative emotions in response to various sources of stress . . . includ[ing] anxiety, fear, irritability, anger, sadness, and so forth."[146] Conscientiousness and Neuroticism share an inverse correlation, but the

---

[143] Joshua J. Jackson et al., *What Do Conscientious People Do? Development and Validation of the Behavioral Indicators of Conscientiousness (BIC)*, 44 J. OF RSCH. IN PERSONALITY 501, 501 (2010).
[144] *Id.* at 502.
[145] Mark Travers, *Two Hidden Personality Traits That High Achievers Have in Common*, FORBES (Nov. 6, 2020), https://www.forbes.com/sites/traversmark/2020/11/06/two-hidden-personality-traits-that-high-achievers-have-in-common/?sh=e6f5d646bda7.
[146] David H. Barlow et al., *The Nature, Diagnosis, and Treatment of Neuroticism: Back to the Future*, 2 CLINICAL PSYCHOL. SCI. 344, 344-345 (2013).

inverse correlation by no means indicates exclusivity. One can be both high in Conscientiousness and high in Neuroticism, as commonly seen in law schools where anxious high-achievers and those with imposter syndrome appear to congregate.[147] This relationship leads to Neuroticism possibly being the second most valuable trait to Biglaw.

Generally, Neuroticism is an "exaggerated negative emotionality" and is accompanied by "the pervasive perception that the world is a dangerous and threatening place, along with beliefs about one's inability to manage or cope with challenging events."[148] The two aspects of Neuroticism are Withdrawal and Volatility. Logically, this split makes sense. If Neuroticism is considered a sort of "threat sensitivity," then any time a threat appears, the two options would be to either hide from it or behave in a manner that will shake up one's reality to counteract the threat-induced volatility. Neuroticism is not simply a measure of sadness, nor is it necessary to push the slightest bit of Withdrawal or Volatility into the realm of psychological disorder.[149] Because many psychologists attempt to confirm that evolution would have selected out useless levels of high or low Neuroticism, some hypothesize that the utility of high Neuroticism—which appears to be exclusively detrimental at first glance—is to limit the consequences of human exploratory behavior, often driven by Plasticity. Predators, discovery of new foods, and outside tribes would come with risks, and a temperament that could not process risk but only opportunity would certainly lead to early death. A continuation of that hypothesis states that high Neuroticism is becoming increasingly obsolete in the modern world, where most threats at a biological level (e.g., disease or starvation) have been largely eliminated. This would imply that evolution has yet to catch up with the

---

[147] *Id.* at 345 ("These beliefs often are manifested in terms of heighted focus on criticism, either self-generated or from others, as confirming a general sense of inadequacy and perceptions of lack of control over salient events.").

[148] *Id.*

[149] However, disorders like depression may often times appear no different than an individual high in Withdrawal that has had a series of negative events in their lives, often one reinforcing the next. Practically speaking, excessive proclivities in line with Withdrawal or Volatility may not induce behavior all too different from depressive or manic disorders. *See generally* Chengwei Lui et al., *Influence of Neuroticism on Depressive Symptoms Among Chinese Adolescents: The Mediation Effects of Cognitive Emotion Regulation Strategies*, 11 FRONTIERS IN PSYCHIATRY May 2020, at 1, 2 (describing how "neuroticism is closely related to depressive symptoms and anxiety."); Gregg Henriques, *Trait Neuroticism and Depressive and Anxiety Disorders*, PSYCH. TODAY (Feb. 26, 2017), https://www.psychologytoday.com/us/blog/theory-knowledge/201702/trait-neuroticism-and-depressive-and-anxiety-disorders ("Given the very close association between anxiety and depression and the understanding of high [Neuroticism] . . . it is clear that high [Neuroticism] should be related to anxiety and depressive disorders.").

changing reality for humans and those who are exceptionally high in Neuroticism are too neurotic for their, or anyone else in society's own good.

B. Effects: Predicting Performance in the Workplace

Having described the five personality attributes, the implications of those attributes should be speculated as to how they relate to performance in Biglaw. It is vitally important to remember that this Note does not prescribe "better" personalities for Biglaw, and these predictions are primarily speculative in order to illustrate the kinds of observations one could make when seeing the attributes in action. The end goal is to improve Biglaw's ability to pursue candidates that it believes are best, not candidates that this Note determines to be best. In addition, although the attributes are generally analyzed one at a time, it is crucial to remember that all five are in action at any given moment. People are complex and personality cannot serve as the sole model to explain away everything about a candidate.

The analysis will be mainly divided into Plasticity and Stability because most of the meaningful considerations for Biglaw occur on the Stability side. While Plasticity is not a small or ignorable portion of personality and may have very desirable balances for the "ideal" candidate, Stability is where the wider range of possible performance predictions can be found. It should be noted that it is improbable that the Biglaw candidate pool expands to every reach of the spectrum for all five attributes. For example, individuals who are excessively low in Conscientiousness would have dropped out of college or never attended to begin with because a graduate degree is intensive in both work and long-term planning. The most disagreeable members of society (especially when combined with low Conscientiousness) are also unlikely candidates because the most disagreeable demographic has a high probability of being presently incarcerated and thus not in law school or applying for Biglaw.[150] While Biglaw talent acquisition has little to worry about for the most troublesome candidates compared to all of society, there is still a sufficient range of each personality attribute to be able to

[150] Scott A. McGreal, *The Paradox of Conscientious Prisoners*, PSYCH. TODAY (Dec. 27, 2016), https://www.psychologytoday.com/us/blog/unique-everybody-else/201612/the-paradox-conscientious-prisoners ("[C]riminals tend to be lower than most people in agreeableness (sympathy for others) and conscientiousness (self-control).").

locate more or less fitting personality profiles out of a large enough pool of candidates.

Across the board, it appears that pursuing high Plasticity is not very rewarding when pitched against existing Biglaw hiring criteria. Rather, it is the avoidance of extremely low Extraversion and extremely low Openness that seems to do the most work. With regards to Plasticity, the screening measures provided by law school admissions and law school grades, combined with the subjective social screening provided by the conversational interview or individual sponsor, seem to sufficiently screen out problematic candidates for Openness and Extraversion respectively. Extraversion, both Assertiveness and Enthusiasm, and Creativity are not very helpful predictors of workplace performance to start,[151] and Intellect currently carries a host of administrability issues that even if overcome, would prove to be relatively unhelpful.[152] First, learning appropriate social skills can mask surface-level problems for those that fall a little deep into the introverted side of the spectrum. Another reason why Extraversion might correlate so little with predictions of on-the-job success is that the wide variety of possible work for lawyers may allow extraverts to self-select into extraverted roles and for introverts to self-select into introverted roles.[153] Beyond the extremely low end of the spectrum, which could throttle colleague collaboration and client-facing interactions, Extraversion is otherwise not a personality trait to greatly worry about in terms of finding fit.[154]

---

[151] TARKI, *supra* note 1, at 56 ("Extraversion has a correlation with predicting on-the-job success of 0.09—almost meaningless for validity—while GPA has a correlation of 0.34.").

[152] This is not to downplay the amount of complexity and general cognitive aptitude necessary in order to engage in legal work at all. To say that measuring Intellect would not be helpful does not make for ignorance of the difficult work a lawyer must do in Biglaw, or any lawyer job at that. The pool of possible candidates (law school students and graduates) sets a very high absolute floor of Intellect relative to all of society. It is precisely because most, if not all of the candidate pool is already within a strata of high Intellect to begin with, that further measuring the trait would not be too helpful. If Biglaw were to hire regardless of education, prescribing measurements of Intellect (whether that takes the form of IQ or some other evaluation) would be much more forceful.

[153] Susan Cain, *How to Level the Playing Field for Introverts and Extroverts*, QUIET REVOLUTION ("Write comprehensive job descriptions that inform people how much interaction, networking, collaboration, and advocacy is required in positions before candidates take the jobs. This will enable introverts to self-select out of jobs that they might not thrive in.").

[154] Some firms emphasize seeking "entrepreneurial spirit" or candidates capable of one day building their own books of business. Such talents may measure high in overall exploratory behavior, and if a firm desires high exploratory behavior, Plasticity could be more valued. In addition, although Extraversion is described as including a proclivity towards goal-achievement, this is exploratory behavior (the tendency to set

Creative lawyers are a double-edged sword. Legal analysis is rewarding for those with high fluency and originality in order to argue novel arguments or preempt potential forthcoming issues. However, in Biglaw, excessive Creativity can also be a curse. In an interview, creative candidates will often express themselves in ways that fall outside of existing evaluating frameworks. Here is where an unstructured evaluation portion proves to remain valuable since EPS firms "rejected standardizing evaluation on the grounds that it was an approach that could lead to missing out on 'diamond[s] in the rough.'"[155] Biglaw already screens for sufficient cognitive aptitude using the Cravath System's criteria and measuring IQ enters uncertain legal territory due to its history with discriminatory outcomes.[156] While General Mental Ability ("GMA") and other cognitive tests such as working memory "games" (already utilized by accounting firms) seem to circumvent the variety of issues surrounding the use of IQ, they essentially test for the same factor.[157] Even though higher Intellect is one of the more reliable job performance indicators available,[158] Biglaw already has a process for attaining that criterion, and it would not be a small adaptation to begin outright testing IQ. It does not make sense for the Cravath System, which only seeks sufficient cognitive aptitude, to suddenly choose to seek maximal cognitive aptitude.

Various balances of the three Stability attributes can have a wide range of consequences for Biglaw professionals. In a collaborative environment like a law firm, a certain amount of Agreeableness is necessary in order to not be a detriment to teamwork.[159] The problem

---

goals) and should not be conflated with the proclivity to actually accomplish goals, which is more aptly associated with Conscientiousness.

[155] RIVERA, *supra* note 15, at 125.

[156] *See generally* Griggs v. Duke Power Co., 401 U.S. 424 (1971).

[157] G Factor, or general intelligence factor, is arrived at much like any personality factor using the same statistical method of factor analysis. While personality ends up with five factors that do not meaningfully correlate with each other except into two further categories of Plasticity and Stability, intelligence factors end up correlating with each other across the board. Thus, if five different types of intelligences (e.g., fluid, crystallized, spatial, quantitative, and working memory) correlate so that someone high in one intelligence ends up being high in all of the intelligences, then there is one underlying factor; a statistical conclusion that there is something that is the same about all five intelligence types and thus one intelligence factor. *See* Kendra Cherry, *What Is General Intelligence (G Factor)?*, VERYWELL (Apr. 25, 2021), https://www.verywellmind.com/what-is-general-intelligence-2795210 ("The idea is that this general intelligence influences performance on all cognitive tasks.").

[158] *See* note 152, *supra*.

[159] Bryan Robinson, Ph.D., *One Personality Trait Enhances Job Performance and Success The Most, New Study Finds*, FORBES (Apr. 3, 2022), https://www.forbes.com/sites/bryanrobinson/2022/04/03/one-personality-trait-enhances-job-performance-and-success-the-most-new-study-

faced by the most disagreeable people is that they do not like being told what to do and will speak their mind (especially those low in Politeness), often to the point of breaking professional workplace customs. Thus, the ability to collaborate and communicate with colleagues, supervisors, and clients all requires a base amount of Agreeableness. However, those who are the most agreeable suffer from other ailments. Perhaps obvious in the name, overly agreeable people may agree to do anything for the sake of conflict avoidance rather than healthy conflict resolution.[160] In Biglaw, the inability to set boundaries may allow the most agreeable lawyers to suffer from exploitation both personally and professionally. In a culture like South Korea, where the hierarchy grounded in age and seniority is taken seriously to the extent that it is encoded in the language itself, juniors culturally behave more agreeably towards seniors, stifling potentially important communications in the workplace that can result in disastrous consequences.[161] Conscientiousness is one of the best predictors of performance in the workplace available, at least among statistically measured factors.[162] It should surprise no one that high Industriousness, or the general proclivity to work, would be desired in Biglaw. The work environment is demanding, and all EPS firms share characteristics such as time-intensive work in excess of sixty-five hours per week.[163] Orderliness appears to have almost zero drawbacks and it is much more forgiving to be excessively orderly than to be excessively disorderly. Disorderly people will have a difficult time even getting their own lives in order, so it would be unreasonable to expect them to manage their work, the interests of supervisors, colleagues, and clients, or even their own office space and emails. Excessive Orderliness can be

---

finds/?sh=5d9078be2848 ("The key to creating a strong and healthy workplace is good communication. Agreeableness . . . among coworkers is mutual [and] flows freely.").

[160] *See* Tim Dahi, *The Personality Trait That Makes You Vulnerable To Exploitation*, ILLUMINATION (Nov. 10, 2021), https://medium.com/illumination/the-personality-trait-that-makes-you-vulnerable-to-exploitation-990c459f7148 ("[Y]ou feel that asserting your own needs/wants would lead to conflicts, and agreeable people always shy away from conflict.").

[161] Ashley Halsey III, *Lack of Cockpit Communication Recalls 1999 Korean Airlines Crash Near London*, WASH. POST (July 8, 2013), https://www.washingtonpost.com/local/trafficandcommuting/lack-of-cockpit-communication-recalls-1999-korean-airlines-crash-near-london/2013/07/08/0e61b3ca-e7f5-11e2-a301-ea5a8116d211_story.html ("[T]he first officer said nothing, even though the instrument in front of him indicated that the plane was turned almost sideways . . . Korean culture is hierarchical. You are obliged to be deferential toward your elders and superiors in a way that would be unimaginable in the U.S.").

[162] TARKI, *supra* note 1, at 61.

[163] RIVERA, *supra* note 15, at 17.

detrimental at the truly extreme end, echoing ADHD-like behavior, where the focus on organization takes priority over meeting the goals of the work itself. However, much of Orderliness can be channeled into work itself, and a lawyer who is very high in both Orderliness and Industriousness could even end up working hard enough to irreparably harm their own health and lifestyle. While it is tragic to see professionals obsess over their work—sometimes to the extent where they kill themselves with it—they are quite rare and pursuing candidates that are high in Conscientiousness is a generally effective strategy for a work environment like Biglaw. High turnover rates[164] in this sense may indicate that the demands of work are too high for the average Biglaw associate and that it is a job that is not only best suited for high Conscientiousness, but also severely ill-suited for low Conscientiousness. Law firms are often concerned with high turnover rates,[165] and the long-term, work and goal-oriented nature of conscientious candidates should be seen in most cases as an attractive trait that will help minimize turnover. Finally, low Neuroticism tends to be favorable for Biglaw firms. Neurotic people are less likely to focus under stress[166] and are more likely to burnout from work.[167] The extremely high end of Neuroticism may require medication and psychiatric treatment in order to function normally in the workplace.[168] Although Neuroticism at an individual level may be undesirable, across the board it is probably beneficial to have a meaningful level of Neuroticism within one's firm, office, or even smaller task team. Sufficient threat sensitivity in the aggregate is what balances opportunity with risk and having enough Neuroticism can sometimes serve as "voices of reason" in the midst of an otherwise very risk-tolerant team.

---

[164] *See* Debra Cassens Weiss, *Law firms came 'dangerously close' to losing almost a quarter of their associates in 2021*, new report says, ABA J. (Jan. 11, 2022), https://www.abajournal.com/news/article/law-firms-came-dangerously-close-to-losing-a-quarter-of-their-associates-in-2021 ("The associate turnover rate for law firms reached 23.2% through November 2021 on a rolling 12-month basis.").

[165] *See generally* Link Christin, *Confronting Lawyer Turnover in Law Firms*, ATT'Y AT WORK (Mar. 27, 2021), https://www.attorneyatwork.com/confronting-lawyer-turnover-in-law-firms/ ("44 percent of associates leave their firms after being there for three years, including entry-level and lateral hires.").

[166] Marissa Higgins, *How Neuroticism May Affect You At Work*, BUSTLE (Oct. 6, 2016), https://www.bustle.com/articles/188204-how-neuroticism-affects-you-at-work-according-to-science-might-explain-your-tendency-to-get-distracted ("[P]eople who displayed neurotic tendencies tended to have a lower ability to focus on tasks for an extended period of time.").

[167] Renzo Bianchi, *Burnout is more strongly linked to neuroticism than to work-contextualized factors*, 270 PSYCHIATRY RSCH 901, 904 (2018).

[168] *See* note 150, *supra*.

III. USING ARTIFICIAL INTELLIGENCE TO GROUND CRITERIA AND ADAPT PROCESS

"What if we could use many more predictors, gather much more data about each of them, spot relationship patterns that no human could detect, and model these patterns to achieve better prediction? This, in essence, is the promise of AI."[169] Revisiting the core argument for the injection of quantitative fit into Biglaw talent acquisition, this Note suggests an upgrade for the process surrounding the existing criteria. At the interview stage, evaluators are seeking fit and simultaneously gravitating positively toward candidates that are like themselves. Considering the flaws introduced by letting human judgment run rampant, it makes sense to adopt AI to help reinforce Biglaw's endeavors to find candidates that are similar to the existing revenue generating employees, who are the supervisors and colleagues that an eventual hire would work with.

"Machine learning is a subset of artificial intelligence which applies statistical techniques to 'enable machines to improve at tasks with experience.'"[170] Thankfully, personality is also the product of statistical techniques and lends itself extremely well to machine learning. What then is the task that we can assign to machine learning? The proposal is to not only have candidates present a personality profile but also to have existing employees, associates, and partners alike, submit their personality profiles consistently throughout their careers. Machine learning will take the data of personality profiles over time and provide clarification on what sort of temperamental proclivities are held by high-level performers. "For data mining and deep learning to work, programmers have to translate the problem or desired outcome 'into a question about the value of some target variable.'"[171] Each firm should investigate what its desired outcome is. The billable hour or fees collected as a measure of productivity is one possible metric for what a "good hire" is in Biglaw. Over time, AI would identify the personalities of the most productive lawyers, and candidates at the interview stage can have their personalities compared to that of the expected star performer. My gut-level prediction is that Conscientiousness, particularly Industriousness, will show to be a desirable personality trait. Partnership can be a measure of success in Biglaw, and perhaps a firm

---

[169] KAHNEMAN, *supra* note 88, at 128-29.
[170] McKenzie Raub, *Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices*, 71 ARK L. REV. 529, 531 (2018).
[171] *Id.* at 533.

wants to prioritize longevity and career development over a raw number of hours worked. AI would learn over time what personalities are most likely to become partners. My gut-level prediction is that here, high Extraversion and somewhat low Agreeableness may be the sweet spot. Another possible metric is self-reported satisfaction or turnover rates. Lawyers in Biglaw can be surveyed to see how satisfied they are with their careers at different stages in their career or whether they expect to move on from Biglaw entirely. If turnover rates are to be reduced, AI may be able to learn what kinds of proclivities those who resign tend to have. My gut-level prediction is that avoiding high Neuroticism may lead to lower turnover rates.

One apparent criticism of machine learning is that candidates may learn to lie (as they already do in qualitative evaluations, giving answers that they believe interviewers will want to hear) or that existing employees will ill-perceive their own personalities. Such flaws are inevitable in self-reported data sets such as personality. However, machine learning is a patient process, and such concerns should be alleviated over time. These criticisms further necessitate that machine learning is conducted over the course of many "generations" of data. In Biglaw, because there are sets of years for the expected partnership track and sets of years with higher turnover rates, a candidate's lies or an employee's faulty self-reporting will correct itself in due time, hopefully within just one generation's worth of firm-wide data. While it is unfortunate that any one particular individual may get away with exploiting the flaws of a self-reported machine learning hiring tool, such exploitation already exists and would not worsen because AI and self-reported personality fails to entirely prevent it. It would be better to have long-term safeguards of self-correction than to continue hiring practices that have no safeguards at all.

In any event, having the data of both existing and potential employees over the course of years and decades, combined with metrics of success that the firm chooses for itself (whether the firm believes it is an accurate representation of themselves or a goal moving forward instead), allows machine learning and personality to output statistical models that calibrate and clarify good hires from bad ones. Personality should be implemented with machine learning because it reinforces the statistical rigor that factor analysis already puts it through. If a firm's goals or business environment changes over time, machine learning will reflect that shift accurately. At the hiring stage, a personality profile does little to fight bias if only the candidate provides such a profile for an evaluator to subjectively analyze. The bias would then be transferred to

the personality profile, and nothing would have really been improved. Ultimately, machine learning will allow a matching of personalities—exactly what unstructured interviews for fit seek to do already—by allowing a firm to come up with its own metrics, which can even be a dynamic blend of considerations. What quantitative fit through machine learning allows *is for a firm to become more like itself*.

CONCLUSION

Within the legal industry, Biglaw is in the best—arguably unique—position to implement personality testing via AI to attain a competitive edge in talent. First, Biglaw may be the only type of organization in the legal industry to have the resources to implement such practices. Second, Biglaw probably crosses the minimal threshold of candidates and employees necessary to provide enough data to put data-intensive machine learning processes to use. Third, Biglaw with its profit-based motivations places a premium on having a competitive edge with talent in ways that other kinds of legal entities do not. Thus, the effort-to-reward ratio is sufficient to justify dedicating resources to develop AI for talent acquisition.

Personality testing is only going to reward a firm for its increased efforts if the testing is implemented in a manner that can adapt to changing circumstances both internally and externally. Personality profiles for candidates yet to be hired may be interesting and helpful, but they can also reinforce biases, discriminatory outcomes, or other existing issues if a firm has poor personalities to begin with. Law firms that come to conclusive decisions about what a "right" personality for an attorney will find themselves with a series of difficult problems to solve. Even if a specific balance of traits is determined to be desired in the most favored candidate, what benefits from other balances of traits are being left off the table? Will the candidate pool remain sustainable in light of new, more specific criteria? What happens when those hired and retained in a firm become *too* similar to each other?

Quantitative fit can be implemented using machine learning and provides a sustainable process that allows for a firm to not only find the best personalities to match its existing community of professionals, but also offers a fairer assessment to candidates. Wholesale objections to using personality profiles imply overthrowing Biglaw's existing criteria because it is precisely personality that is already being measured at the interview stage. Wholesale objections to using machine learning to accomplish personality matching are objections to the utility of statistics

and self-adapting solutions. Biglaw should not be criticized for continuing to seek what it believes is best for itself, but that cannot be an excuse for deciding against improved talent. Quantitative fit driven by machine learning will give Biglaw the talent it wants, but more accurately, consistently, and efficiently.

VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

NOTES

REGULATING ARTIFICIAL INTELLIGENCE: A CALL FOR A UNITED STATES ARTIFICIAL INTELLIGENCE AGENCY

Noah John Kahekili Rosenberg

REGULATING ARTIFICIAL INTELLIGENCE: A CALL FOR A UNITED STATES ARTIFICIAL INTELLIGENCE AGENCY

*Noah John Kahekili Rosenberg**

INTRODUCTION

"Sorry, I didn't quite get that," rudely interrupts Siri any time the word "seriously" or "series" is mentioned in a conversation.[1] Many Americans have become accustomed to hearing this voice coming from their pockets, but there was a time when artificial intelligence (AI) seemed like a distant dream, an unreachable fiction, a phenomenon that only existed in movies. For developers, the rapid growth of AI technologies is exciting—for others, it's frightening.[2] Today, AI is everywhere: it talks to us from our phones, it navigates our roadways, and it sends you those "perfectly" targeted advertisements on social media platforms.[3]

Despite the involvement of AI in our daily lives, the federal government has largely left the field unregulated.[4] AI has many advantages that include reducing human error and taking on risks that

* J.D. Candidate, Notre Dame Law School, Class of 2023. I would like to express my sincere gratitude to Professor Stephen Yelderman for his invaluable guidance and feedback on this Note. Mahalo nui loa to my family, friends, and loved ones for their consistent and endless support. Lastly, thank you to my colleagues at the Notre Dame Journal on Emerging Technologies for their diligent work and insight. Any errors are my own.

[1] Siri is a digital assistant built into Apple products that can be activated with the verbal command "hey Siri." *Siri*, APPLE, https://www.apple.com/siri/ (last visited May 1, 2022).

[2] Ron Schmelzer, *Should We Be Afraid of AI?*, FORBES (Oct. 31, 2019), https://www.forbes.com/sites/cognitiveworld/2019/10/31/should-we-be-afraid-of-ai/?sh=4e1799944331 ("One of the most widespread fears of AI is just general anxiety about it and what it's potentially capable of. A recurring theme in movies and science fiction is AI systems that go rogue . . . .").

[3] Mike Kaput, *AI in Advertising: Everything You Need to Know*, MARKETING AI INSTITUTE (Mar. 10, 2022), https://www.marketingaiinstitute.com/blog/ai-in-advertising.

[4] Heather Sussman et al., *U.S. Artificial Intelligence Regulation Takes Shape*, ORRICK (Nov. 18, 2021), https://www.orrick.com/en/Insights/2021/11/US-Artificial-Intelligence-Regulation-Takes-Shape#:~:text=Next%20Steps,regulation%20is%20on%20the%20horizon (addressing generally that there is no artificial intelligence regulation in the U.S.).

would ordinarily burden humans.[5] Additionally, AI systems are available at all times of the day, every day of the week, compared to the eight hours most humans work.[6] They can help expedite the process of tedious and repetitive jobs, and they can make decisions much quicker than humans.[7] Notwithstanding these benefits, there are many concerns about AI, including human unemployment, its potential to make humans lazy, high costs of innovation, its inability to feel emotions, and a lack of creative thinking.[8] More significantly, AI has the potential—if left unregulated—to be dangerous to public safety and equality.

For example, a widely used risk-prediction program in the U.S. healthcare system was found to favor white patients over black patients in determining who would be likely to need extra medical care.[9] Similarly, an Amazon facial recognition technology, Rekognition, wrongly identified a number of professional athletes as criminals, including Duron Harmon, a professional football player and safety for the New England Patriots.[10] Since federal agencies and their regulations are often designed to promote equality and safety, these incidents make it clear that there are significant risks with leaving AI technology unregulated.[11]

Proceeding in three parts, this Note draws upon two examples of emerging AI technologies that demonstrate the need for federal regulation: autonomous vehicles (i.e., self-driving cars) and algorithm-based hiring software. Part I illustrates the public safety concerns associated with AI technologies by outlining the inadequacy of existing laws and regulations on autonomous vehicles. Part II addresses the

---

[5] Sunil Kumar, *Advantages and Disadvantages of Artificial Intelligence*, TOWARDS DATA SCI. (Nov. 25, 2019), https://towardsdatascience.com/advantages-and-disadvantages-of-artificial-intelligence-182a5ef6588c.
[6] *Id.*
[7] *Id.*
[8] *Id.*
[9] Cem Dilmegani, *Bias in AI: What It Is, Types, Examples & 6 Ways to Fix It in 2022*, AIMULTIPLE (Jan. 12, 2022), https://research.aimultiple.com/ai-bias/ (finding that the AI program associated past medical spending with medical needs which inadvertently created racial bias since race and income are heavily correlated).
[10] Priya Dialani, *Famous AI Gone Wrong Examples in the Real World We Need to Know*, ANALYTICS INSIGHT (Mar. 9, 2021), https://www.analyticsinsight.net/famous-ai-gone-wrong-examples-in-the-real-world-we-need-to-know/.
[11] *See generally About Us*, U.S. DEPT. LABOR, https://www.dol.gov/general/aboutdol#:~:text=To%20foster%2C%20promote%2C%20and%20develop,work%2Drelated%20benefits%20and%20rights (last visited Apr. 26, 2022); *About NHTSA*, U.S. DEPT. OF TRANSPORTATION, https://www.nhtsa.gov/#:~:text=About%20NHTSA,%2C%20safety%20standards%2C%20and%20enforcement (last visited Apr. 26, 2022) ("Our mission is to save lives, prevent injuries, and reduce economic costs due to road traffic crashes, through education, research, safety standards, and enforcement.").

shortcomings of current regulations on algorithm-based hiring software and the issue of discrimination and inherent bias in AI. Part III recommends the creation of a new federal agency to guide AI regulation and enforcement.

I. AUTONOMOUS VEHICLES AND THE IMMINENT THREAT TO PUBLIC SAFETY

Every year in the United States, more than 38,000 people die as a result of car accidents.[12] This means that over one hundred people die in the U.S. each day due to vehicle collisions, making road crashes the leading cause of death in the nation for people under the age of fifty-four.[13] In addition to fatalities, approximately 4.4 million people are injured in car accidents and require medical treatment.[14] A study conducted by the National Highway and Traffic Safety Administration ("NHTSA") found that the economic costs of motor vehicle crashes totaled $242 billion in 2010.[15] The numbers become more horrifying after factoring in lost quality of life valuations,[16] which bring the total economic societal loss in America due to car crashes to $836 billion.[17] With nearly $1 trillion in costs to American taxpayers, it is no wonder the government is attracted to the idea of autonomous vehicles and is worried about stifling innovation by creating regulations.

Autonomous vehicles are expected to increase the safety of American roadways because we largely attribute motor vehicle collisions

---

[12] *Road Safety Facts*, ASIRT, https://www.asirt.org/safe-travel/road-safety-facts/#:~:text=More%20than%2038%2C000%20people%20die,for%20people%20aged%201%2D54 (last visited Jan. 28, 2022) ("Road crashes are the leading cause of death in the U.S. for people aged 1-54.").

[13] *Id.*; NHTSA, *Fatal Motor Vehicle Crashes: Overview*, NHTSA, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812456.

[14] *Id.*; *see also* Melanie Musson & Sara Routhier, *Which States Allow Self-Driving Cars? (2021 Update)*, AUTO INS. (Nov. 17, 2021), https://www.autoinsurance.org/which-states-allow-automated-vehicles-to-drive-on-the-road/.

[15] MILLER BLINCOE ET AL., NHTSA, THE ECONOMIC AND SOCIETAL IMPACT OF MOTOR VEHICLE CRASHES, 1 (revised ed. 2010), https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013 (finding that the total economic loss of motor vehicle crashes in the U.S. is $242 billion when considering factors such as "lost productivity, medical costs, legal and court costs, emergency service costs (EMS), insurance administration costs, congestion costs, property damage, and workplace losses").

[16] The term "lost quality of life" or "diminished quality of life" refers to the reduction of a person's ability to enjoy normal areas of life and overall health because of the injures or disabilities resulting from an accident. *Diminished Quality of Life in a Personal Injury Lawsuit*, LEGAL MATCH, https://www.legalmatch.com/law-library/article/diminished-quality-of-life-in-a-personal-injury-lawsuit.html (last visited May 1, 2022).

[17] BLINCOE ET AL., *supra* note 15, at 1.

to human error.[18] A study on autonomous vehicle collisions in California conducted by IDTechEx revealed that out of 187 reported autonomous vehicle accidents, only two were the fault of the performance of the autonomous system—roughly one percent of the accidents.[19] This seems to lend support for a quick rollout of self-driving vehicles because "computer drivers are in principle fundamentally safer drivers. They never text, do their makeup, or fall asleep at the wheel."[20] With widespread deployment and use of autonomous vehicles on our roadways, self-driving vehicles would be able to communicate with each other and warn nearby cars of its planned maneuver before changing lanes, coming to a stop, or similar actions.[21] Computers also react faster at about 0.5 seconds compared to humans who typically have a reaction speed of approximately 1.6 seconds.[22] Theoretically then, releasing autonomous vehicles into the public should prove to be a positive development that reduces fatalities, accidents, and injuries.

Unfortunately, there is good reason to be skeptical of any study that claims to show autonomous vehicle collisions are too infrequent to be important. For one, companies self-report their own collision statistics.[23] Second, even if the IDTechEx study is accurate, a small percent of crashes being caused by system failure becomes more significant when millions of these cars enter the roadways. Third, even if the lead developers of autonomous vehicles are releasing safe technology, that does not guarantee that competitors will not rush the

---

[18] *See* Ben Wodecki, *Human Error Causes 99% of Autonomous Vehicle Accidents: Study*, IOT WORLD TODAY (Oct. 20, 2021), https://www.iotworldtoday.com/2021/10/20/blame-the-humans-idtechex-finds-99-percent-of-autonomous-vehicle-accidents-caused-by-human-error/ (finding that only one percent of autonomous vehicle collisions were the result of actual malfunction or poor performance by the vehicle's autonomous system).

[19] *Id.* California requires companies testing autonomous vehicles to report all collisions to the California DMV which allowed IDTechEx to conduct its study seen in its report "Autonomous Cars, Robotaxis & Sensors 2022-2042." IDTechEx, *The Biggest Challenge for Autonomous Vehicles, Discussed by IDTechEx*, PR NEWSWIRE (Oct. 19, 2021), https://www.prnewswire.com/news-releases/the-biggest-challenge-for-autonomous-vehicles-discussed-by-idtechex-301403437.html.

[20] Nathan A. Greenblatt, *Self Driving Cars Will be Ready Before Our Laws Are*, IEEE SPECTRUM (Jan. 19, 2016), https://spectrum.ieee.org/selfdriving-cars-will-be-ready-before-our-laws-are.

[21] *Id.*

[22] *Id.*; Aarian Marshall, *Puny Humans Still See the World Better than Self-Driving Cars*, WIRED (Aug. 5, 2017), https://www.wired.com/story/self-driving-cars-perception-humans/#:~:text=Machines%20can%20react%20faster%20than,autonomous%20vehicles%20do%20even%20better.

[23] U.S. DEP'T OF TRANSP., STANDING GENERAL ORDER 2021-01: INCIDENT REPORTING FOR AUTOMATED DRIVING SYSTEMS (ADS) AND LEVEL 2 ADVANCED DRIVER ASSISTANCE SYSTEMS (ADAS) (2021).

same process. Moreover, "human error" in the IDTechEx study refers to the error of human drivers of other vehicles or the human error of pedestrians.[24] Since it is unlikely and even improbable that *every* vehicle on American roadways will be replaced by autonomous vehicles in the near future, human drivers will remain and accidents will continue to occur, putting the public at risk. So, while it's expected that the implementation of autonomous vehicles on our highways will eventually decrease motor vehicle accidents, current evidence suggests that we should be hesitant to allow companies to release vehicles before the federal government deems them safe.[25] Indeed, ever since companies began testing vehicles with varying degrees of autonomous driving features on public roadways, there have been disturbing reports of system failure, some of which resulted in fatalities.[26]

Taking into account that this technology has been released without federal safety regulations, it is unsurprising that self-driving cars are involved in more automobile collisions per miles driven than conventional cars.[27] Although the injuries sustained in these crashes are often less severe than those in human-driven cars,[28] this does not justify the lack of safety standards, regulations, or testing on self-driving vehicles *before* they are used on public roadways. As demonstrated by the following examples, extreme system failure in autonomous vehicles have resulted in tragedy.

---

[24] *Id.*

[25] *See* Steven Palermo, *Self-Driving Car Manufacturers May be Safe from Lawsuits Even if Their Cars Cause Accidents*, PALERMO L., https://thesuffolkpersonalinjurylawyer.com/self-driving-car-defects-manufacturer-may-never-face-lawsuit/ (last visited Nov. 22, 2021) (claiming that self-driving cars could prevent tens of millions of traffic fatalities, but acknowledging that dangerous mistakes occasionally occur in technology); *see also* Rachel Abrams & Analynn Kurtz, *Joshua Brown, Who Died in Self-Driving Accident, Tested Limited of His Tesla*, N.Y. TIMES (July 1, 2016), https://www.nytimes.com/2016/07/02/business/joshua-brown-technology-enthusiast-tested-the-limits-of-his-tesla.html (reporting on Joshua Brown's death that occurred as a result of his Tesla's autopilot failing to apply the brakes after a tractor-trailer made a left turn in front of his vehicle).

[26] *See* Abrams & Kurtz, *supra* note 25; Ray Stern, *Trial Delayed for Backup Driver in Fatal Crash of Uber Autonomous Vehicle*, PHX. NEW TIMES (May 12, 2021), https://www.phoenixnewtimes.com/news/uber-crash-arizona-vasquez-herzberg-trial-negligent-homicide-charge-11553424 (explaining that an Uber autonomous vehicle failed to brake as a pedestrian walked her bike across the road resulting in the death of Elaine Herzberg, the pedestrian).

[27] *Autonomous Vehicles Statistics*, GERBER INJ. L. (June 25, 2015), https://gerberinjurylaw.com/autonomous-vehicle-statistics/; *The Dangers of Self-Driving Cars*, NAT'L L. REV. (May 5, 2021), https://www.natlawreview.com/article/dangers-driverless-cars.

[28] *Id.*

A. Documented System Failure

The most well-known accident involving the system failure of an autonomous vehicle occurred in 2016 in Florida and was the first fatal Tesla autopilot crash.[29] Forty-five-year-old Joshua Brown died tragically after his Tesla Model S crashed into the side of a semi-truck while traveling on autopilot.[30] According to Tesla and Elon Musk, the white side of the tractor against a brightly lit sky caused the front-facing sensors of the autopilot system—a camera, a radar, and ultrasonic sensors—to fail to detect the semi-truck.[31] Additionally, since the semi-truck was higher off the ground than typical vehicles, the radars tuned it out, believing it to be an overhead road sign, and thus the autopilot chose not to apply the brakes.[32] More perplexing however, is that the NHTSA conducted an investigation into the crash, and ultimately decided that there was no defect on the Tesla sensor system and did not issue a recall.[33]

Similar outcomes came of a 2018 Uber self-driving crash. A pedestrian named Elaine Herzberg was struck by one of Uber's autonomous vehicles while walking across the street in Arizona.[34] It was determined that the vehicle turned off its automatic braking system in order to avoid unsafe driving conditions, and that the driver, Rafaela Vazquez was watching "The Voice" in the Hulu app on her phone in the minutes leading up to the crash.[35] While it seems like both the vehicle and driver may be at fault, criminal prosecutors only pursued charges against the driver.[36] These examples show the imperfection of autonomous vehicle technology, the drastic consequences of public

---

[29] Fred Lambert, *Tesla Is Under Scrutiny from Feds Again Over Crash with Semi Truck*, ELEKTREK (Mar. 16, 2021) [hereinafter *Tesla Crash*], https://electrek.co/2021/03/16/tesla-under-scrutiny-feds-again-over-crash-semi-truck/.

[30] *Id.*

[31] Fred Lambert, *Understanding the Fatal Tesla Accident on Autopilot and the NHTSA Probe*, ELEKTREK (July 1, 2016) [hereinafter *Elon Musk*], https://electrek.co/2016/07/01/understanding-fatal-tesla-accident-autopilot-nhtsa-probe/.

[32] *Id.*

[33] *Tesla Crash*, *supra* note 29.

[34] Jim Gill, *How 3 Cases Involving Self-Driving Cars Highlight eDiscovery and the IOT*, JD SUPRA (Sept. 3, 2019), https://www.jdsupra.com/legalnews/how-3-cases-involving-self-driving-cars-76886/.

[35] *Id.*; Ray Stern, *Trial Delayed for Backup Driver in Fatal Crash of Uber Autonomous Vehicle*, PHX. NEW TIMES (May 12, 2021), https://www.phoenixnewtimes.com/news/uber-crash-arizona-vasquez-herzberg-trial-negligent-homicide-charge-11553424 (explaining that the charges were filed against Rafael, Rafaela's name prior to her transition as a transgender woman).

[36] *Id.*

testing, and the necessity of proactive federal regulations to prevent more needless accidents like these from occurring.

B. Defining Autonomous Vehicles and the State of Current Technology

The U.S. Department of Transportation has defined six levels of automation to categorize autonomous vehicles based on how much control the human operator maintains:

> At SAE Level 0, the human driver does everything; [a]t SAE Level 1, an automated system on the vehicle can *sometimes assist* the human driver conduct *some parts of* the driving task; [a]t SAE Level 2, an automated system on the vehicle can *actually conduct* some parts of the driving task, while the human continues to monitor the driving environment and performs the rest of the driving task; [a]t SAE Level 3, an automated system can both actually conduct some parts of the driving task and monitor the driving environment in *some instances*, but the human driver must be ready to take back control when the automated system requests; [a]t SAE Level 4, an automated system can conduct the driving task and monitor the driving environment, and the human need not take back control, but the automated system can operate only in certain environments and under certain conditions; and [a]t SAE Level 5, the automated system can perform all driving tasks, under all conditions that a human driver could perform them.[37]

For the purposes of this article, the assumption will be that any vehicles referred to as "autonomous" will fall between SAE levels 3-5 in which the human operator is not required to perform *any* driving tasks for at least some period of time. For these levels of automation, the human operator of the vehicle can be considered a passenger rather than a driver while the vehicle maintains control of itself.

---

[37] U.S. DEPARTMENT OF TRANSPORTATION, FEDERAL AUTOMATED VEHICLES POLICY: ACCELERATING THE NEXT REVOLUTION IN ROADWAY SAFETY 9 (2016) [hereinafter FEDERAL AV POLICY].

C. The Inadequacy of Existing State Law and Federal Regulation

The role of states is commonly perceived to be regulating *drivers* while the regulation of *cars* is often left to the federal government.[38] Since the federal government has remained largely silent on the issue of autonomous-vehicle regulation, states have been conflicted between the choice of hindering innovation and trying to protect drivers and other individuals on their roads.[39] Most states' laws either assume a human being will be in control (or ready to retake control) of the vehicle or require that a human being with a valid driver's license remain in the driver seat at all times.[40] Due to the vagueness, lack of clarity, or lax laws and regulations, even fully autonomous vehicles can probably be legally deployed in any state as long as a licensed human is behind the wheel.[41] However, the presence of a driver alone is an insufficient safeguard, because research shows that drivers of autonomous vehicles will often be unprepared or unable to regain control in the event of a system failure.[42]

1. Current State Law on Autonomous Vehicles

As of 2017, twenty-eight states had already introduced legislation concerning autonomous vehicles, but these focused primarily on development and testing rather than actual safety standards and protections for public consumers and road users.[43] For example, in 2017, the New York legislature passed a law regulating autonomous vehicles on

---

[38] *See* Marielle Segarra & Sasha Fernandez, *The Road Ahead: What About Regulation for Self-Driving Cars?*, MARKETPLACE TECH. (Oct. 1, 2021), https://www.marketplace.org/shows/marketplace-tech/what-about-regulation-for-self-driving-cars/.

[39] *See id.*

[40] HG Legal Resources, *Are Self-Driving Cars Legal?*, HG.ORG, https://www.hg.org/legal-articles/are-self-driving-cars-legal-31687 (last visited Nov. 22, 2021); *see also* Press Release, New York City Department of Transportation, Notice of Adoption Relating to the Demonstration or Testing of Autonomous Vehicles (Sept. 7, 2021) (on file with author) [hereinafter Notice of Adoption].

[41] HG Legal Resources, *supra* note 40 (stating that "the laws of most states assume a human being will be in control, but this legal vagueness means that autonomous vehicles may technically be allowed to operate over the roads provided a human being sits behind the wheel").

[42] Nancy Grugle, *Human Factors in Autonomous Vehicles*, ABA (Nov. 20, 2019), https://www.americanbar.org/groups/tort_trial_insurance_practice/publications/tortsource/2019/fall/human-factors-autonomous-vehicles/.

[43] *See* Ben Husch & Anne Teigen, *Regulating Autonomous Vehicles*, NAT'L CONF. STATE LEGS. (Apr. 2017), https://www.ncsl.org/research/transportation/regulating-autonomous-vehicles.aspx.

the roadway, requiring, *inter alia*, that there be a natural person with a valid driver's license present within the vehicle during the duration of the trip.[44] However, these regulations applied only to "demonstrations and tests."[45]

In 2017, The National Conference of State Legislatures acknowledged that states needed to implement further regulations in areas including traffic enforcement, insurance, registration, and licensing, but theorized that the creation of these regulations was not a pressing concern because it "will likely be many years before fully autonomous vehicles see widespread deployment."[46] Although it is true that fully autonomous vehicles (i.e., Level 5 vehicles) have not yet been widely deployed in the United States, Level 3 vehicles have already infiltrated American roadways.[47]

Tesla's recently released "Full Self-Driving Capability" package includes the ability for the vehicle to navigate on autopilot, auto lane change, auto park, summon itself, and have traffic light and stop sign control.[48] In addition, Tesla advertises that new features, such as the ability to autosteer on city streets are "coming soon."[49] Some states require companies to obtain permits before testing or deploying autonomous vehicles on public roadways,[50] but the permit application requirements are often insufficient to ensure public safety. For example, California requires that applicants for its Autonomous Vehicle Tester program have tested their vehicles and have "reasonably determined" that they are safe to operate.[51] No further information is provided that defines what constitutes "reasonable." Given this ambiguity, one should be skeptical of a claim that a company has met sufficient safety guidelines or standards merely because they hold a permit.

---

[44] Notice of Adoption, *supra* note 40; RULES OF CITY OF NY DEP'T OF TRANSP, 34 RCNY § 4-17 (2021).
[45] Notice of Adoption, *supra* note 40.
[46] Husch & Teigen, *supra* note 43.
[47] Fred Lambert, *Tesla Launches its Full Self-Driving Subscription Package for $199 Per Month*, ELECTREK (July 16, 2021, 8:33 PM), https://electrek.co/2021/07/16/tesla-launches-full-self-driving-subscription-package-199-per-month/ [hereinafter *Tesla Package*]. For a description of the automation levels, *see supra* Part I(B).
[48] *Tesla Package*, *supra* note 47.
[49] *Id.*
[50] Segarra & Fernandez, *supra* note 38.
[51] STATE OF CALIFORNIA DEPARTMENT OF MOTOR VEHICLES, AUTONOMOUS VEHICLE TESTER (ATV) PROGRAM FOR MANUFACTURER'S TESTING PERMIT 4 (2020).

2. Current Federal Law and Regulations on Autonomous Vehicles

The lack of adequate state law and regulation on autonomous vehicles may be explained by the federal government's declaration that the federal government alone is responsible for "setting safety standards for new motor vehicles" and "enforcing compliance with the established safety standards."[52] Yet, even at the federal level, there are no laws or mandatory standards specifically geared toward self-driving vehicles. The NHTSA released its first set of guidelines in September 2016, *Federal Automated Vehicle Policy: Accelerating the Next Revolution in Roadway Safety*.[53] While these appeared to be federal regulations on autonomous vehicles, they were merely non-mandatory guidelines and mostly impractical or based on misunderstandings of the technology.[54] For example, the policy asks manufacturers to ensure that ethical decisions are made "consciously and intentionally," which is improbable for an AI system.[55] The National Conference of State Legislatures outlined the policy as follows:

> Section 2 of the guidance, the Model State Policy (MSP) delineates federal versus state authority. While the federal government is responsible for setting motor vehicle safety standards, states remain the lead regulator when it comes to licensing, registration, traffic law enforcement, safety inspections, infrastructure, and insurance and liability.
>
> The MSP outlines a road map for states wanting to move ahead with testing and eventually deploying autonomous vehicles. It offers steps a state could consider rather than a detailed set of legislative language. Specifically, it notes that "this guidance is not mandatory," though the agency may make "some elements of the guidance mandatory and binding through future rulemakings." Further, it identifies

---

[52] *See* Musson & Routhier, supra note 14 (summarizing the federal and state responsibilities regarding self-driving cars).
[53] *See generally* FEDERAL AV POLICY, *supra* note 37.
[54] *See* Jeremy Laukkonen, *Are Self-Driving Cars Legal in Your State?*, LIFEWIRE (July 13, 2021), https://www.lifewire.com/are-self-driving-cars-legal-4587765.
[55] *See* Srikanth Saripalli, *Before Hitting the Road, Self-Driving Cars Should Have to Pass a Driving Test*, SCI. AM. (Feb. 22, 2018), https://www.scientificamerican.com/article/before-hitting-the-road-self-driving-cars-should-have-to-pass-a-driving-test/; FEDERAL AV POLICY, *supra* note 37.

> several areas of state law that might require updating to accommodate a world full of automated vehicles. These include law enforcement and emergency response, vehicle registrations, liability and insurance, education and training, vehicle inspections and maintenance, and environmental impacts.[56]

This presents two irreconcilable ideas. First, the Federal Automated Vehicle Policy is an express set of non-mandatory guidelines. Thus, if states want autonomous vehicles to be subjected to mandatory safety standards, they must implement those standards alone. However, if the federal government lacks the knowledge and resources to regulate self-driving technology, individual states are likely to find themselves similarly situated. Second, according to the MSP, "setting safety standards" is the responsibility of the federal government.[57] States are explicitly encouraged not to regulate safety standards in order to "ensure the establishment of a consistent national framework rather than a patchwork of incompatible laws."[58] In summary, the U.S. Department of Transportation gave itself the responsibility to provide states with motor safety standards for self-driving vehicles, and then failed to provide adequate protections through the implementation of mandatory regulations.[59] The MSP goes further to clarify that:

> Under current law, manufacturers bear the responsibility to self-certify that all of the vehicles they manufacture for use on public roadways comply with all applicable Federal Motor Vehicle Safety Standards (FMVSS). Therefore, if a vehicle is compliant within the existing FMVSS regulatory framework and maintains a conventional vehicle design, there is currently no specific federal legal barrier to an HAV being offered for sale.[60]

The 2016 Automated Vehicle Policy provided a performance guide that asked manufacturers to voluntarily provide a safety assessment that covered: data recording and sharing, privacy, system safety, vehicle cybersecurity, human machine interface, crashworthiness, consumer

---

[56] NAT'L CONF. STATE LEGS., *Regulating Autonomous Vehicles*, https://www.ncsl.org/research/transportation/regulating-autonomous-vehicles.aspx.
[57] Musson & Routhier, supra note 14.
[58] FEDERAL AV POLICY, *supra* note 37, at 7.
[59] *See generally id.*
[60] *Id.*

education and training, registration and certification, post-crash behavior, federal, state, and local laws, ethical considerations, operational design domain, object and event detection and response, fall back, and validation methods.[61] The system safety guidelines suggested that the goal should be designing systems "free of unreasonable safety risks," but failed to provide any meaningful standards by which manufacturers should measure such safety risks.[62] This was still a step in the right direction, but the voluntary nature of the guidelines rendered them less effective and inhibited public confidence. Additionally, although the 2016 policy predicted possible mandatory guidelines in the future, the three subsequent reports have followed the voluntary framework of their predecessors.[63]

As of July 13, 2021, "nowhere in the United States is it strictly illegal to own or operate a self-driving car."[64] This absence of regulations means manufacturers are able to release their newly developed self-driving features to the public without meeting any federal safety standards specific to autonomous vehicles.[65] In large part, the reason for Congress's absence in self-driving car regulation is due to the difficulty of writing performance standards in an unfamiliar emerging technology such as autonomous vehicle software.[66] Further, technology companies who have not been traditionally subject to such regulations have significantly opposed any proposed legislation attempting to fill this void.[67] The consequences are drastic. In the absence of substantive regulation, manufacturing companies have been using the general public as "guinea pigs." Jason Levine, executive director of the Center for Auto Safety, stated that tech and car companies are testing the safety of their self-driving modes by "using you and me and everyone in your neighborhood as part of their experiment . . . just putting vehicles out on public roads, public highways, neighborhood streets, across the country, and collecting data and seeing how it goes."[68] As discussed above, the NHTSA—instead of imposing proactive safety restrictions, standards, or

---

[61] *Id.* at 15.
[62] FEDERAL AV POLICY, *supra* note 37.
[63] NATIONAL SCIENCE & TECHNOLOGY COUNCIL & U.S. DEPARTMENT OF TRANSPORTATION, AUTOMATED VEHICLES 4.0: ENSURING AMERICAN LEADERSHIP IN AUTOMATED VEHICLE TECHNOLOGIES 29-30 (2020) ("The U.S. Government will promote voluntary consensus standards as a mechanism to encourage increased investment and bring cost-effective innovation to the market more quickly.").
[64] FEDERAL AV POLICY, *supra* note 37.
[65] Segarra & Fernandez, *supra* note 38.
[66] *Id.*
[67] *See id.*
[68] *See id.*

regulations on manufacturing companies—has required manufacturers to report when their vehicles crash.[69] While that is important data to collect, the NHTSA is essentially trying to determine whether the cars are safe by seeing how many people get into accidents rather than making sure the cars are safe *prior* to releasing them into the public.[70]

Despite these criticisms on the NHTSA's automated vehicle guidelines, the real issue is that the NHTSA is ill-equipped to develop anything more substantial. In fact, the NHTSA has expressed its desire to create a safety framework with objective standards to define and measure the safety of autonomous vehicles,[71] but also acknowledged that it lacks the necessary funding and expertise to accomplish this goal.[72] This illustrates the inadequacy of current federal regulations and supports the conclusion that the creation of a federal AI agency may be a workable solution by providing an increase in expertise, funding, and rulemaking authority.

II. AI HIRING ALGORITHMS AND DISCRIMINATION PROTECTION

Employers in the United States are increasingly using AI programs in their hiring practices.[73] In 2019, a Mercer report found that 40% of U.S. companies used AI programs to assist their hiring processes.[74] There are a variety of programs that recruiters may use throughout the different stages of the hiring process.[75] At the earliest stage, companies use AI programs to selectively advertise certain job openings to candidates based on information submitted by the candidates and their job application history on the site.[76]

---

[69] *Id.*; U.S. DEP'T OF TRANSP., STANDING GENERAL ORDER 2021-01: INCIDENT REPORTING FOR AUTOMATED DRIVING SYSTEMS (ADS) AND LEVEL 2 ADVANCED DRIVER ASSISTANCE SYSTEMS (ADAS) (2021).
[70] *See id.*
[71] U.S. DEP'T OF TRANSP., AUTOMATED VEHICLES: COMPREHENSIVE PLAN 12 (2021).
[72] NAT'L CONF. STATE LEGS., *supra* note 56 ("Finally, the guidance lays out some possible policy changes that NHTSA believes could help it better respond to this new technology. These include additional funding to support more research, a larger network of experts, premarket approval authority for vehicles and software upgrades after vehicles sell.").
[73] Rebecca Heilweil, *Artificial Intelligence Will Help Determine if You Get Your Next Job*, VOX (Dec. 12, 2019), https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen.
[74] Tim Kulp, *AI and Hiring Bias: Why You Need to Teach Your Robots Well*, HUM. RES. EXEC. (Apr. 14, 2021), https://hrexecutive.com/ai-and-hiring-bias-why-you-need-to-teach-your-robots-well/.
[75] Heilweil, *supra* note 73.
[76] *Id.*

Other companies offer a recruiting tool that goes beyond self-submitted information by potential candidates, and find top candidates based on information on the open web.[77] This recruiting tool is sometimes even able to identify candidates most likely to leave their current job.[78] At the next stage of the recruiting process, companies may use AI tools to filter through resumes and present the employer with a list compiling the top candidates to interview.[79] One of the companies that offers this tool, HireVue, takes this practice a step further and uses AI to analyze and conduct actual interviews, during which candidates are prompted with structured questions and asked to record themselves responding.[80] The AI program uses proprietary machine learning algorithms to analyze data points from the interview—including, for example, non-verbal cues such as "facial expressions, eye-movements, body movements, details of clothes, and nuances of voice"—to predict future job performance.[81]

Like self-driving cars, the use of AI in hiring practices has the potential for many societal benefits.[82] These programs are often used by companies and recruiters to greatly reduce the time and effort needed to sift through and evaluate candidates.[83] Proponents of AI hiring boast its potential to remove human biases from the recruiting process and its ability to be more predictive of job success than traditional interviews.[84]

Yet, the use of this emerging technology has revealed disturbing discrepancies between its goal of removing racial bias and its unintended result of racial and gender discrimination.[85] For example, Amazon, the world's largest online retailer, abandoned its 2014 project to create an AI program to automate its recruitment process after discovering that it filtered out female candidates.[86] The initial goal of Amazon's AI program

---

[77] *Id.*; *Products: Arya Quantum*, ARYA LEOFORCE, [hereinafter *Arya*] https://goarya.com/arya-quantum/ (last visited Jan. 30, 2022).
[78] *See Arya*, *supra* note 77.
[79] *See* Heilweil, *supra* note 73.
[80] *Hiring Experience Platform*, HIREVUE, https://www.hirevue.com/ (last visited Jan. 30, 2022).
[81] *HireVue Interview Guide: How to Prepare for a HireVue Interview*, CORP. FIN. INST., https://corporatefinanceinstitute.com/resources/careers/interviews/about-hirevue-interview/ (last visited Jan. 30, 2022).
[82] McKenzie Raub, *Bots, Bias, and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices*, 71 ARK. L. REV. 529, 530 (2018).
[83] Heilweil, *supra* note 73.
[84] *Id.*
[85] *See generally id.*
[86] Isobel Asher Hamilton, *Amazon Built an AI Tool to Hire People but Had to Shut It Down Because It Was Discriminating Against Women*, INSIDER (Oct. 10, 2018), https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-

was to create a system capable of analyzing resumes and producing a list of top candidates.[87] After a year, developers realized that the AI software used statistics about the company's past male-dominated employment and resume collection, which led the AI program to conclude that male candidates were preferred.[88] Thus, the AI engine scored resumes lower or filtered out the candidate altogether if their resume contained the word "women's" or the candidate had attended an all-women's college.[89] Similarly, the Electronic Privacy Information Center ("EPIC") filed a complaint against HireVue with the Federal Trade Commission alleging "unfair and deceptive trade practices."[90] Although it is unclear whether the program actually displayed racial biases because the biometric data was analyzed secretly, HireVue reportedly stopped using facial expressions as a factor in its algorithmic analysis of video interviews after the complaint was filed.[91]

These companies are not alone in their struggle to develop a non-discriminatory AI hiring program, and the issue is not limited to the context of employment discrimination.[92] The same issue was found in a 2016 ProPublica study on AI software that aided in making parole judgments by predicting which criminals were likely to reoffend.[93] This software was found to display racial biases against Black defendants, finding them more likely to reoffend based only on their skin color.[94] EPIC criticizes the use of AI in similar practices alleging that it has caused substantial harm to the American public whom are subjected to "opaque and un-provable decision-making in employment, credit, healthcare,

---

against-women-2018-10; Troy Segal, *Who Are Amazon's (AMZN) Main Competitors?*, INVESTOPEDIA (July 17, 2021), https://www.investopedia.com/ask/answers/120314/who-are-amazons-amzn-main-competitors.asp#:~:text=Amazon%20is%20the%20world's%20largest,subscription%20services%2C%20and%20web%20services.

[87] *Id.* (quoting an unnamed source: "They literally wanted it to be an engine where I'm going to give you 100 résumés, it will spit out the top five, and we'll hire those"); Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[88] *See* Dastin, *supra* note 87.

[89] *Id.*

[90] Jeremy Kahn, *HireVue Drops Facial Monitoring Amid A.I. Algorithm Audit*, FORTUNE (Jan. 19, 2021), https://fortune.com/2021/01/19/hirevue-drops-facial-monitoring-amid-a-i-algorithm-audit/; Complaint, In re HireVue, Inc. (F.T.C. Nov. 6, 2019), https://epic.org/wp-content/uploads/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf.

[91] *Id.*

[92] Dastin, *supra* note 87.

[93] *Id.*

[94] *Id.*

housing, and criminal justice."[95] According to EPIC, other commercial uses of this technology include ranking sports players and evaluating potential Airbnb guests.[96]

A. What Causes an Algorithm to be Discriminatory or Biased?

Algorithm developers and recruiters hope that AI hiring systems can provide a way to evaluate candidates objectively and eliminate human prejudice and subjectivity, but the current reality is that human biases unexpectedly infiltrate decisions made by AI.[97] One source of bias in AI programming originates from the creation of the algorithms themselves and those designing them. Accordingly, many argue that AI algorithms are biased due to the "lack of meaningful diversity in Silicon Valley."[98] The fundamental problem is that algorithms are thought to embed the authors' opinions into the code.[99] Since there is a lack of diversity in the tech industry—and thus, a lack of diversity in the creators of these algorithms—the algorithms reproduce the authors' implicit biases as well as existing societal biases.[100] When human resource managers work together with data scientists to create these algorithms, they decide which factors are important and how the AI coding can account for them.[101] In doing so, they design AI systems to consider certain factors without accounting for many of the unconscious judgments that would normally help inform the human recruiter.[102] For example, while a human recruiter may value proximity of the address on an applicant's resume to the firm's location, an AI tool designed to value the same factor may inadvertently discriminate on race in a segregated city.[103] Additionally, non-minority white developers may not ensure (or even be aware that they should be ensuring) the programs they design

---

[95] Kahn, *supra* note 90.
[96] *Id.*
[97] Miranda Bogen, *All the Ways Hiring Algorithms Can Introduce Bias*, HARV. BUS. REV. (May 6, 2019), https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias.
[98] Raub, *supra* note 82, at 540.
[99] *Id.* at 542.
[100] *Id.* at 542. Implicit bias is loosely defined as "preconceived notions or stereotypes that—[we all have and that are] beyond our control—affect our understanding, actions, and decisions about others." Stacy Cantu-Pawlik, *What Is Implicit Bias and Why Should You Care?*, SALUD AMERICA (Apr. 1, 2019), https://salud-america.org/what-is-implicit-bias-and-why-should-you-care/.
[101] Was Rahman, *AI-Powered Recruitment Can Be Racist or Sexist – and Here's Why*, DIVERSITYQ (Jan. 13, 2021), https://diversityq.com/ai-powered-recruitment-can-be-racist-or-sexist-and-heres-why-1511217/.
[102] *Id.*
[103] *Id.*

sufficiently distinguish non-white faces and fairly assess non-verbal cues for minority users.[104] A lack of diversity also creates gender bias. For example, recruiters designing an AI program to analyze resumes may want to weed out those that have career gaps.[105] However, if those designers are men, they may not account for the fact that many women will have gaps in their employment due to maternity leave and other childcare obligations, and effectively use gender as an eliminating or downgrading criteria.[106]

Machine learning, defined as "a class of methods for automatically creating models from data,"[107] is another source of bias and discrimination in AI hiring. The Amazon program referenced earlier is a great example of machine learning and illustrates how an algorithm can unintentionally create discriminatory preferences through data analysis. The data in that case was a ten-year collection of resumes submitted to Amazon, most of which came from male candidates.[108] The male dominance in the industry led the program to infer that male candidates were better suited for the job and thus, began recommending men over women.[109] Further, the system analyzed the text on the resumes for commonalities and began to assign little value to skills that were common to all applicants, and placed higher value on verbs found mostly on men's resumes such as "executed."[110] Although Amazon was able to revise the algorithm to be gender-neutral in these contexts, the unpredictability of the program making future discriminatory inferences was so great that the developers ultimately abandoned the project.[111] The root of the problem in machine learning is that it acts to perpetuate existing biases and underrepresentation in historical data. When your data set lacks a representative amount of diversity, a program modeled after that data has no way of determining how those groups have performed in the

---

[104] Michael Li, *To Build Less-Biased AI, Hire a More Diverse-Team*, HARV. BUS. REV., (Oct. 26, 2020), https://hbr.org/2020/10/to-build-less-biased-ai-hire-a-more-diverse-team.
[105] *See* Parmy Olson, *Employers Beware: Hiring Software Could Weed Out Future Stars*, WASH. POST (Sept. 14, 2021), https://www.washingtonpost.com/business/employers-beware-hiring-software-could-weed-out-future-stars/2021/09/14/e2fcb574-152a-11ec-a019-cb193b28aa73_story.html.
[106] Rahman, *supra* note 101.
[107] Martin Heller, *Machine Learning Algorithms Explained*, INFOWORLD (May 9, 2019), https://www.infoworld.com/article/3394399/machine-learning-algorithms-explained.html.
[108] Dastin, *supra* note 87.
[109] *Id.*
[110] *Id.*
[111] *Id.*

past—and therefore, no way of knowing how they will perform in the future.[112] Thus, because minorities have historically been drastically underrepresented in many industries, AI programs are modeled to determine that they are less preferable than white candidates with tons of collected historical data to analyze.[113]

B. Existing Laws and Regulations on AI Employment Discrimination

Many states have introduced bills targeting AI, but few have actually enacted any AI legislation. Many of these bills contain loopholes or regulate only certain entities, and are therefore, insufficient to provide employment protections to consumers from discriminatory AI hiring tools.[114] Since there is also a lack of federal AI regulation, there is hardly any oversight on AI hiring programs in the United States.[115] Still, we can find examples of promising state legislation in Illinois, New York, and Maryland.

Illinois passed one of the first laws that targeted AI hiring practices: the Artificial Intelligence Video Interview Act (the "AIVIA"), which took effect in January 2020.[116] The AIVIA took the first step towards regulating a largely uncertain, nontransparent technology by focusing on privacy, disclosure, and consent.[117] Essentially, the law requires companies that use AI video interviewing programs to disclose to applicants that their applications will be reviewed by AI rather than a human recruiter.[118] Additionally, the law requires that such applicants consent to an AI interview before employers may subject them to one.[119] At face value, this law appears to provide consumers with protections, but in reality it falls short—far too short. For starters, the law fails to address any concerns for bias, and thus, fails to provide any protections

---

[112] Sarah K. White, *AI in Hiring Might Do More Harm Than Good*, CIO (Sept. 17, 2021), https://www.cio.com/article/189212/ai-in-hiring-might-do-more-harm-than-good.html.
[113] *See generally id.*
[114] *Id.*
[115] *Id.*
[116] Artificial Intelligence Video Interview Act § 1, 820 ILL. COMP. STAT. 42/1 (2020); Rebecca Heilweil, *Illinois Says You Should Know if AI Is Grading Your Online Job Interview*, VOX (Jan. 1, 2020) [hereinafter *Illinois AI Video Interview Act*], https://www.vox.com/recode/2020/1/1/21043000/artificial-intelligence-job-applications-illinios-video-interivew-act.
[117] Artificial Intelligence Video Interview Act § 5; *Illinois AI Video Interview Act*, *supra* note 116.
[118] *Illinois AI Video Interview Act*, *supra* note 116.
[119] *Id.*

for discrimination.[120] Additionally, the law reaches only video interviewing technology, which makes up a relatively small portion of the AI hiring tools.[121] Lastly, although the law requires consent by the interviewee, it does not offer any alternative remedies to those who do not wish to consent—thus, potential applicants are left with a choice between: (1) consenting to the AI interview program despite their reservations; or (2) withdrawing their application and not being considered for the job at all.[122] Notwithstanding these shortcomings, the AIVIA did increase interview transparency to some degree. Transparency is important to protect candidates against discrimination because candidates are often unaware that they were even eliminated by a program rather than a human.

The New York City Council passed a bill in early November 2021, which prohibits employers from using AI hiring tools unless the program undergoes a "bias audit" one year prior to its use and can demonstrate that the program will not discriminate based on an applicant's race or gender.[123] Additionally, the bill follows the AIVIA's strides towards transparency, and requires that employees and candidates be notified if an AI tool is used to make the hiring decision.[124] The penalty for failure to disclose is a fine of $500 to $1500.[125] Although the requirement of an audit is a promising start, many critics argue that the law sets too weak of a standard to effectively protect against bias.[126] One issue is that the audit requirement is too vague and only requires companies to show that they comply with basic requirements that are "very easy to meet."[127] The ineffectiveness of audits can be seen through a third-party audit of HireVue, which despite the problems in the system, commended the company for its efforts to eliminate potential bias.[128] The auditors went on to *recommend* that the company take further steps to investigate

---

[120] *Id.*
[121] *Id.*
[122] *Id.*
[123] Automated Employment Decision Tools, N.Y. COMP. CODES R. & REGS. tit. 20, § 870-74 (2023); Matt O'Brien, *A New Bill Would Limit Employers' Use of A.I. Hiring Tools to Recruit New York City Applicants*, FORTUNE (Nov. 19, 2021), https://fortune.com/2021/11/19/new-york-city-bill-employers-ai-hiring-tools-applicants/.
[124] Erin Mulvaney, *NYC Targets Artificial Intelligence Bias in Hiring Under New Law*, BLOOMBERG L. (Dec. 10, 2021), https://news.bloomberglaw.com/daily-labor-report/nyc-targets-artificial-intelligence-bias-in-hiring-under-new-law.
[125] *Id.*
[126] O'Brien, *supra* note 123; *see also* Kate Kaye, *New York City Passed a Bill Requiring 'Bias Audits' of AI Hiring Tech*, PROTOCOL (Nov. 12, 2021), https://www.protocol.com/bulletins/nyc-ai-hiring-tools.
[127] O'Brien, *supra* note 123.
[128] Kahn, *supra* note 90.

potential biases.[129] Another issue is that the law only protects against racial and gender bias and fails to address other protected classes, such as disability or age.[130] The measure will go into effect on January 1, 2023.[131]

Maryland passed legislation similar to the AIVIA. The new law simply requires employers to get applicant consent before they can use a facial recognition service (essentially Maryland's coined phrase to refer to AI video interviewing programs that analyze facial expression, word choice, and voice).[132] Because the Maryland law is very similar to its counterpart in Illinois, it likewise faces similar challenges. Thus, consent does not adequately protect applicants from biased AI hiring tools. Additionally, as Maryland employment attorneys have noted, the law does not specify any penalties or fines for companies that fail to comply.[133]

To date, there are no existing federal regulations that address AI discrimination in employment.[134] However, Title VII of the Civil Rights Act of 1964, amended by the Americans with Disabilities Act and the Pregnancy Discrimination Act of 1978, prohibits employment discrimination on the basis of race, color, religion, sex, disability, pregnancy and national origin.[135] Although these laws do not address the use of AI in hiring, their protections may extend to such situations.

Title VII liability falls into two separate categories of claims: (1) disparate treatment, and (2) disparate impact claims.[136] Disparate treatment claims require intentional discrimination, and thus, aside from being extremely difficult to prove, would theoretically not apply to unintentionally created bias from AI hiring tools.[137] Consequently, job

---

[129] *Id.*
[130] O'Brien, *supra* note 123.
[131] Jason C. Gavejian, *NYC Places Groundbreaking Restrictions on AI Use in Hiring Practices*, JACKSONLEWIS (Dec. 20, 2021), https://www.workplaceprivacyreport.com/2021/12/articles/artificial-intelligence/nyc-places-groundbreaking-restrictions-on-ai-use-in-hiring-practices/.
[132] MD. CODE ANN., LAB. & EMPL. § 3-717 (West 2020); Charles R. Bacharach & James D. Handley, *Maryland Passes Consent Requirement for Employment Interview Use of Facial Recognition Services*, GORDON FEINBLATT (June 16, 2020), https://www.gfrlaw.com/what-we-do/insights/maryland-passes-consent-requirement-employment-interview-use-facial-recognition.
[133] *Id.*
[134] Sussman et al., *supra* note 4.
[135] Civil Rights Act of 1964, 42 U.S.C. § 2000e-2 (1964); Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101-12213 (2008); *Title VII of the Civil Rights Act of 1964*, SHRM, https://www.shrm.org/hr-today/public-policy/hr-public-policy-issues/pages/titleviiofthecivilrightsactof1964.aspx (last visited Jan. 31, 2022).
[136] Raub, *supra* note 82, at 544.
[137] *Id.*; Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395, 405 (2018).

applicants are likely limited to disparate impact claims under Title VII as their legal avenue for protection from discriminatory AI hiring tools. These claims arise when an employer uses a policy that is neutral on its face (appears to be non-discriminatory at face value), but has a discriminatory effect on the basis of one of the protected classes.[138] As the United States Supreme Court held in *Griggs v. Duke Power Co.*, Title VII covers practices that are "fair in form, but discriminatory in operation."[139] This language seems to indicate that Title VII should extend to AI hiring, but many scholars are skeptical of its application. Some argue that AI hiring discrimination cases being brought as disparate impact claims will likely fail due to the business necessity defense, or because the algorithm in question would be facially discriminatory as it "classifies on a prohibited ground."[140] If the practice is not facially neutral, it would need to be brought as a disparate treatment claim, and therefore, would fail due to the difficulty in proving intent.[141]

Even if Title VII encompasses AI hiring discrimination, discriminatory impact claims face the problem of "intentional reverse discrimination." This term is used to describe the situation in which an employer actively tries to account for disparate impact on a protected group by actively making that group more likely to get a job—thereby, intentionally discriminating against those not in the protected group. The conception of this term arises from the Supreme Court case, *Ricci v. DeStefano*, in which the city of New Haven, Connecticut discovered that white candidates consistently outperformed minority candidates on an examination they used to evaluate potential firefighters.[142] When city officials found the racial disparity, they threw out the results of the exam in order to make the hiring criteria more equitable for Black candidates.[143] The Court held that the city's intentional discrimination was impermissible under Title VII, absent a strong showing that the city would have been liable under a disparate impact claim if no action was taken.[144] This case creates an uncertainty of how and whether companies can account for racial disparities they discover in their AI hiring programs. The holding also makes clear that the standard for disparate impact is not such a low bar, and it remains to be decided whether racial

---

[138] Raub, *supra* note 82, at 544.
[139] 401 U.S. 424, 431 (1971).
[140] Sullivan, *supra* note 137, at 410-11.
[141] *Id.* at 411.
[142] 557 U.S. 557, 562 (2009).
[143] *Id.*
[144] *Id.* at 563.

disparities present in AI hiring programs would meet that bar. Yet, even if disparate impact claims could succeed, the current law almost encourages companies to forgo taking corrective measures since doing so may open them up to the same sort of liability found in *Ricci*.[145]

The lack of transparency in AI hiring makes it extremely difficult for candidates to learn why they were eliminated, particularly where state laws do not require consent. Reactive solutions like Title VII claims are insufficient to protect those that rarely know they were victims of discrimination. The American public should not be forced to rely on companies to self-regulate their AI hiring tools. It is unrealistic to hope that every company will strictly scrutinize its AI software data, find discriminatory results, and correct or abandon the programs. Therefore, it is vital that this technology is proactively regulated. Until effective regulations are created, companies will continue to use and test their programs at the expense of candidates who are serving as guinea pigs in this nationwide experiment.

III. RECOMMENDATION: THE UNITED STATES ARTIFICIAL INTELLIGENCE AGENCY

Before arguing that a United States Artificial Intelligence Agency is necessary, it is fundamental to explain why federal regulations are necessary. AI is everywhere—driving on our roads, scouring our social media, and sitting behind a desk reading our resumes—and it affects everyone. Any single flaw in AI could affect millions of people in the U.S. [146] Without federal regulations, consumers are left without protections and are often unaware of the effects that AI may be having on them. Disclosure regulations are extremely important. In the AI hiring context, companies are not required to provide any proof that their programs actually detect factors relevant to job performance.[147] Many AI scholars suggest that future regulations on AI should require controls on the application of AI technologies, data collection, limits on how long data can be retained, the use of the AI technologies, the use of independent third-party testing, and significant transparency.[148]

---

[145] *See* Raub, *supra* note 82, at 555.
[146] François Candelon et al., *AI Regulation is Coming: How to Prepare for the Inevitable*, HARV. BUS. REV., Sept.-Oct. 2021, https://hbr.org/2021/09/ai-regulation-is-coming.
[147] Mark MacCarthy, *AI Needs More Regulation, Not Less*, BROOKINGS (Mar. 9, 2020), https://www.brookings.edu/research/ai-needs-more-regulation-not-less/.
[148] White, *supra* note 112.

This is a national problem, and national problems require national solutions. As one commentator notes, the creation of a federal agency is a proven solution when "an entire field begins to set a broad set of challenges for the public, demanding thoughtful regulation."[149] A federal agency was created to help alleviate a new national concern in: (1) 1906, when the Food and Drug Administration ("FDA") was created in response to a national concern of unsanitary and shocking conditions in U.S. meat-packing plants; (2) 1934, when the Securities and Exchange Commission ("SEC") was created in response to the national concern of the worst stock market crash in history; and (3) 1970, when the Environmental Protection Agency ("EPA") was created in response to the national concern for pollution.[150] The list goes on and on, and the theme is consistent—when the nation is faced with a broad issue, the federal government has successfully responded by creating federal agencies to make and enforce effective regulations. Although the government has recently created the National Artificial Intelligence Initiative ("NAII"), it does not carry the same authority as an agency and the NAII's mission is geared towards winning the international race on AI.[151] Further, the NAII simply works between the existing agencies that are not equipped with AI expertise or focused on AI regulation.[152] In contrast, federal agencies have a significant amount of expertise in specialized areas, and they are required to allow public participation through public comments.[153] Most significantly, agencies have rulemaking authority to "write and enforce regulations that have the force and effect of law."[154]

A. Addressing Existing Criticism Toward an AI Agency

There are two main concerns in the literature that have created skepticism about the idea of creating an AI agency: (1) the complexity of the technology, and (2) impeding innovation.[155] The first concern expresses the fear that regulators will be unable to understand complex

---

[149] Rob Toews, *Here Is How the United States Should Regulate Artificial Intelligence*, FORBES (June 28, 2020), https://www.forbes.com/sites/robtoews/2020/06/28/here-is-how-the-united-states-should-regulate-artificial-intelligence/?sh=402d86377821.
[150] *Id.*
[151] *National Artificial Intelligence Initiative: Overseeing and Implementing the United States National AI Strategy*, NAT'L A.I. INITIATIVE, https://www.ai.gov/ (last visited Jan. 31, 2022).
[152] *Id.*
[153] *See* MAEVE P. CAREY, CONG. RSCH. SERV., IF10003, AN OVERVIEW OF FEDERAL REGULATIONS AND THE RULEMAKING PROCESS (2021), https://sgp.fas.org/crs/misc/IF10003.pdf.
[154] *Id.*
[155] *See generally* Raub, *supra* note 82, at 566-67.

coding, and thus, unable to create meaningful regulations on AI programs.[156] This concern is rooted in the false assumption that regulators need to regulate *inputs* rather than *outputs*.

The distinction between inputs and outputs can be better illustrated in the context of medicine.[157] One issue with regulating autonomous vehicles is that it is not obvious how to test the effectiveness of an algorithm. Likewise, it would be difficult to effectively regulate the vast amount of data going into and being analyzed by an AI hiring program. One way to solve this is to test algorithms the same way we test new medications.[158] In both cases, it is difficult for researchers to always know exactly why something works, but it is still possible to evaluate *what it does* (i.e., evaluate the outcome). In the case of medicine, the outcome tested for is whether a sick person gets better after taking the medication. In the case of algorithms, the outcome tested for could be whether a vehicle is able to detect and slow down for pedestrians walking against a red light or whether an AI hiring program displays racial disparities. Some state legislatures are already taking this output-focused approach. The New York City Council introduced a bill that aimed to increase transparency by disclosure of algorithms, but after backlash, amended the bill to focus on evaluating the outputs of AI to "figure out if and when there is harm done."[159]

The second concern, that strict AI regulations will impede innovation, is greatly contested by scholars.[160] In fact, many argue that regulations would actually increase innovation, because among other things, they encourage greater public trust.[161] This phenomenon was observed after Congress passed the 1974 Fair Credit Billing Act ("FCBA") to regulate credit card companies.[162] The protections from the FCBA increased public trust in the new technology and stimulated growth in the industry and an increase in innovation.[163] The key to is to be proactive. Proactive regulation gets out in front of the new technology to

---

[156] Tristan Greene, *US Government Is Clueless About AI and Shouldn't Be Allowed to Regulate It*, TNW (Oct. 24, 2017), https://thenextweb.com/news/us-government-is-clueless-about-ai-and-shouldnt-be-allowed-to-regulate-it.
[157] *See* Srikanth Saripalli, *Before Hitting the Road, Self-Driving Cars Should Have to Pass a Driving Test*, SCI. AM. (Feb. 22, 2018), https://www.scientificamerican.com/article/before-hitting-the-road-self-driving-cars-should-have-to-pass-a-driving-test/.
[158] *See id.*
[159] Raub, *supra* note 82, at 567.
[160] *See, e.g.*, MacCarthy, *supra* note 147.
[161] *Id.*
[162] The FCBA amended the Truth in Lending Act. *See* Truth in Lending Act of 1968, 15 U.S.C. § 1666 (2010).
[163] MacCarthy, *supra* note 147.

protect consumers who will then trust and support further innovation.[164] Multiple research studies support the conclusion that well-designed regulations increase innovation, particularly when coupled with incentives for adoption of the technology.[165] However, even if innovation is stunted by regulations, the cost of regulative restraint falls largely on minority groups—a consequence that should be enough in itself to outweigh any potential loss of innovation.[166]

B. Setting the Floor for States to Build Upon

1. The Seat Belt Example

The need for automobile safety was a concern well before the development of autonomous vehicles. The federal government has continuously struggled to combat the horrific number of annual fatalities attributed to automobiles collisions. In 1966, Congress responded by passing the National Traffic and Motor Vehicle Safety Act of 1966 ("NTMVSA"), whose purpose was "to provide for a coordinated national safety program and [the] establishment of safety standards for motor vehicles in interstate commerce to reduce accidents involving motor vehicles and *to reduce the deaths and injuries occurring in such accidents*."[167] The NTMVSA further stipulated that the Secretary of Commerce shall establish appropriate standards to protect the public against "unreasonable risk of accidents occurring as a result of the design, construction or performance of motor vehicles" and against unreasonable risks to persons in the events of accidents.[168]

Seat belt legislation is one example of federal motor vehicle safety regulation. The first seat belt law took effect in 1968 and required car manufacturers to install seat belts in every vehicle.[169] While this new law required vehicles to *have* seat belts, it did not require drivers or

---

[164] *Id.*
[165] Will Upington, *Driving AI Innovation in Tandem with Regulation*, TECH. CRUNCH (Oct. 6, 2021), https://techcrunch.com/2021/10/06/driving-ai-innovation-in-tandem-with-regulation/.
[166] Louise Russell-Prywata, *Book Review: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor by Virginia Eubanks*, LSE REV BOOKS (July 2, 2018), https://blogs.lse.ac.uk/lsereviewofbooks/2018/07/02/book-review-automating-inequality-how-high-tech-tools-profile-police-and-punish-the-poor-by-virginia-eubanks/.
[167] National Traffic and Motor Vehicle Safety Act of 1966, Pub. L. No. 89-563, pmbl., 80 Stat. 718, 718 (1966) (emphasis added).
[168] *Id.* at § 102(1).
[169] Jennifer L. Donaldson, *History of Seat Belts: Effective for Men and Women?*, SAFE RIDE4KIDS, https://saferide4kids.com/blog/history-of-seat-belts-effective/.

passengers to *use* the seat belts. Notwithstanding this shortcoming, the federal government's action still gave state legislatures the means to implement seat belt *use* laws. Essentially, without a federal law requiring vehicles to have seat belts, states would be incapable of passing—and definitely incapable of enforcing—seat belt *use* laws.[170] Thus, the federal government opened the door for states to regulate seat belt use and increase motor vehicle safety by legislating on seat belts.

The 1968 seat belt law was very successful. New York passed the first seat belt use law in 1984,[171] which required drivers, front-seat passengers, and back-seat occupants under the age of 10 to wear a seat belt at all times.[172] Many states soon followed New York's lead and passed similar laws. Today, every state except New Hampshire has some variation of a seat belt law.[173] They often vary between primary and secondary enforcement and front-seat-only or all-seat requirements.[174] Despite these enforcement differences, seat belt laws have increased seat belt use, which in turn has reduced automobile collision deaths and injuries.[175] In fact, according to the National Highway Traffic Safety Administration, the national use rate of seat belts was 90.3 percent in 2020.[176] Furthermore, seat belt use in vehicles saved approximately 14,955 lives in 2017.[177] This increase in seat belt usage—and therefore the increase in survivability of occupants involved in car crashes—owes its thanks to the federal government for setting the floor (the minimum standard) for states to build upon.

---

[170] It's important to consider that state laws requiring manufacturers to install seat belts in vehicles would likely be ineffective since they would only reach those manufacturing companies incorporated in that state.

[171] Dennis Hartman, *When Did Seat Belts Become Mandatory*, ITSTILLRUNS, https://itstillruns.com/did-seat-belts-become-mandatory-5506603.html (last visited Jan. 31, 2022).

[172] *Id.*

[173] Samantha Bloch, NAT'L CONF. STATE LEGS., *State and Federal Efforts to Increase Adult Seat Belt Use*, 28 LEGISBRIEF, no. 16, May 2020, https://www.ncsl.org/research/transportation/state-and-federal-efforts-to-increase-adult-seat-belt-use.aspx.

[174] Primary enforcement is when police officers can stop vehicles based solely on a seat belt violation, while secondary enforcement allows police officers to enforce seat belt violations only when the stop was made pursuant to a different crime or traffic violation. *See* Riccola Voigt, *Primary and Secondary Traffic Violations*, NOLO, https://www.drivinglaws.org/resources/primary-and-secondary-traffic-violations.html (last visited Apr. 27, 2022)

[175] *Seat Belt Laws*, U.S. DEP'T TRANSP., https://www.transportation.gov/mission/health/seat-belt-laws (Aug. 24, 2015).

[176] *Seat Belts*, NHTSA, https://www.nhtsa.gov/risky-driving/seat-belts (last visited Mar. 13, 2022, 6:45 PM).

[177] *Id.*

2. The Employment Discrimination Example

Although the federal government has not specifically regulated the use of AI as it relates to employment discrimination, it has already set the employment discrimination floor that states have built upon. That floor is Title VII of the Civil Rights Act of 1964.[178] This federal law protects workers from discriminatory employment practices based on race, color, religion, sex, and national origin.[179] The federal government added protections for people with disabilities in 1990 with the passage of the Americans with Disabilities Act ("ADA").[180] These are the basic protections that all states have to comply with, or better phrased: the bare minimum list of categories states must protect against employment discrimination.

Title VII allowed states to build upon these mandatory protections and add additional protected classes to state laws. Some states impose fewer protections while other states go further in their protections and have passed anti-discrimination laws to provide equal employment regardless of sexual orientation, marital status, or weight.[181] For example, Alabama does not have a law protecting against racial discrimination, and therefore, leaves the issue in the realm of federal law.[182] In contrast, California expands Title VII to protect workers from discrimination based on gender identity, marital status, and sexual orientation.[183] Similar to California, New York is generally seen as "employee-friendly" in its employment discrimination laws and often

---

[178] Lisa Nagele-Piazza, *Not All State Employment Discrimination Laws Are Created Equal*, SHRM (Sept. 15, 2017), https://www.shrm.org/resourcesandtools/legal-and-compliance/state-and-local-updates/pages/state-employment-discrimination-laws.aspx.

[179] Civil Rights Act of 1964, 42 U.S.C. § 2000e-2 (1964); *Know Your Civil Rights*, END HATE, https://www.endasianhate.org/your-civil-rights?gclid=Cj0KCQiArt6PBhCoARIsAMF5wagfOI4MwyjzKKZhew9tnrzj7mEAYCg1dRUouJnpz8nvqxPdP7aEFmcaAu3CEALw_wcB (last visited Jan. 31, 2022).

[180] Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101-12213 (2008); *Questions and Answers: The Application of Title VII and the ADA to Applicants or Employees Who Experience Domestic or Dating Violence, Sexual Assault, or Stalking*, EQUAL EMPLOYMENT OPPORTUNITY COMM'N, https://www.eeoc.gov/laws/guidance/questions-and-answers-application-title-vii-and-ada-applicants-or-employees-who#:~:text=Title%20VII%20of%20the%20Civil,on%20the%20basis%20of%20disability (last visited Jan. 31, 2022) [hereinafter *Title VII & ADA*].

[181] *Employment Discrimination in Your State*, NOLO, https://www.nolo.com/legal-encyclopedia/employment-discrimination-in-your-state-31017.html (last visited Jan. 31, 2022).

[182] Nagele-Piazza, *supra* note 178.

[183] *Id.*

includes protected classes outside the scope of Title VII.[184] Additionally, Title VII only applies to businesses with a minimum of fifteen or twenty employees (depending on the state) and many states decrease that number to be more employee-friendly.[185] These state law additions built upon the groundwork laid by Title VII of the Civil Rights Act of 1964.

3. Application to AI

An effective U.S. Artificial Intelligence Agency ("USAIA") would focus on regulating the outputs of the algorithms, rather than inputs. This way, instead of struggling to analyze the data going into the coding, regulators could avoid the complexities of the technology by requiring that companies reach reasonable and acceptable results. By following the lead of the FDA and New York City Council, the USAIA could regulate even the most complex codes. The burden would shift away from lawmakers, and onto AI developers to obtain results within a tolerable range. Similar to the federal law requiring vehicles to have seat belts and Title VII of the Civil Rights Act of 1964, the USAIA needs to set a "floor" for AI regulations. These regulations may look like minimum safety standards or tests that autonomous vehicles need to pass, such as requiring them to be able to maneuver through unpredictable environments. In the context of AI hiring, possible regulations could be a requirement that any racial disparities in the system be negligible, and mandatory statistical studies on the outputs of the AI programs along with public reporting on the companies' findings. The USAIA may also choose to implement broader regulations such as prohibitions on technologies that violate fundamental human rights (e.g., predictive policing systems), clear public disclosure rules, accountability rules, remedies for consumers, and enforcement rules.[186]

In addition, the USAIA must ensure it is practical and safe for programmers to fix disparities based on race, gender, and other protected classes. AI developers will be hesitant to follow regulations that require them to correct discrimination in their programs unless they have confidence that doing so will not expose them to liability. Therefore, in creating regulations, it would be wise for the USAIA to consider the

---

[184] *Id.*
[185] *Id.* ("California's Fair Employment and Housing Act generally applies to businesses with five or more employees.").
[186] Jascha Galaski, *AI Regulation: Present Situation and Future Possibilities*, LIBERTIES (Sept. 8, 2021), https://www.liberties.eu/en/stories/ai-regulation/43740.

Supreme Court's holding in *Ricci* and other applicable law that risk subjecting companies to "intentional reverse discrimination."[187]

The USAIA's jurisdiction would encompass all forms of AI, but should be limited to those that interact with the general public. Essentially, the USAIA would regulate the readiness of AI products to be released to consumers. This would prevent the agency from being overburdened while also allowing it to ensure companies are not using human guinea pigs to test the safety and fairness of their AI products.

CONCLUSION

Artificial intelligence will continue to spark rigorous debate and concern in the United States as new and uncertain technologies continue to emerge. Although the future societal benefits may be great, we cannot ignore the immediate threats to cybersecurity, privacy, public safety, discrimination, biases, and civil and criminal liability. If left unregulated, artificial intelligence has the potential to cause severe societal harm. Autonomous vehicles are just one example of a public safety risk that artificial intelligence technologies create. Through this illustration, it becomes clear that our state and federal governments lack effective regulations to protect the public from these new dangers. With no mandatory federal regulations in place, car manufacturers will continue to use American public roadways as testing sites for unregulated and dangerous technologies. Likewise, artificial intelligence hiring tools highlight the lack of accountability and transparency of artificial intelligence technologies. Without effective and proactive regulation, the public will continue to serve as guinea pigs and minorities will continue to suffer disproportionately. The creation of the USAIA would increase funding and expertise in the regulation of artificial intelligence, thereby fostering targeted, meaningful, proactive regulations. These regulations will increase public safety, public trust, and innovation, allowing artificial intelligence technologies to flourish, and encouraging reluctant users to embrace the technology with confidence—leading to more fulfilling and happier lives.

---

[187] Ricci v. DeStefano, 557 U.S. 557, 562 (2009).