NORTH CAROLINA JOURNAL OF LAW & TECHNOLOGY
VOLUME 21, ISSUE 2: DECEMBER 2019

MAL-WHO? MAL-WHAT? MAL-WHERE? THE FUTURE CYBER-THREAT OF A NON-FICTION NEUROMANCER: LEGALLY UN-ATTRIBUTABLE, CYBERSPACE-BOUND, DECENTRALIZED AUTONOMOUS ENTITIES

Jonathan A. Schnader*

For decades, science fiction writers have tackled philosophical and existential questions arising from the creation of artificial intelligence ("AI") by human beings. AI, however, is no longer a fictional concept, but rather an evolving part of modern society. How will AI systems impact United States' national security interests? Considering the increased national security threat coming from actors in cyberspace, policymakers should consider the cybersecurity risk of AI systems that operate entirely in cyberspace. This article opines that a serious threat to national security will arise from a cyberspace-bound, decentralized autonomous entity ("CyDAE") because of the "unexplainability" of current AI system design (that is, the difficulty understanding why or how the AI arrived at its conclusion or behaved the way it did), the lack of legal personhood arrangements for autonomous systems, and the already difficult task of attributing acts in cyberspace to human actors or States because of outdated Westphalian notions of sovereignty and territoriality. The article ultimately offers several broad policy suggestions, including: (1) an AI registry; (2) "explainability" criteria for AI system designs; (3) requiring human oversight for legal personhood arrangements (whether arranged in a corporation, limited liability structure, or otherwise) tailored specifically for AI autonomous systems that lack human members; and (4) universal jurisdiction of States over malicious CyDAEs that obfuscate attributive links to human actors or States.

* B.A. Miami University of Ohio, 2008; J.D. Syracuse University College of Law, 2012; LL.M. in National Security, Georgetown University Law Center, 2019. Many thanks to David Flynn and Professor David Koplow for their encouragement and support.

"The future is already here – it's just not evenly distributed."[1]

[1] Christian Horak, *The Future (Of Finance) Is Already Here–It's Just Not Evenly Distributed*, DIGITALIST MAG. (June 1, 2017), https://www.digitalistmag.com/finance/2017/06/01/the-future-of-finance-is-already-here-its-just-not-evenly-distributed-05126253 [https://perma.cc/JC9J-RAGY] (referring to a quote by William Gibson originally published in The Economist on December 4, 2003).

I. INTRODUCTION

The science fiction "cyberpunk" author William Gibson predicted a future where the rise of powerful, disembodied artificial intelligences, living in the intangible world of "cyberspace,"[2] willfully act with difficult-to-determine purposes and opaque (or non-existent) human affiliations and loyalties.[3] Stunningly, Gibson depicts a world of sensory overload where these artificial intelligences manipulate humans for their own ends, commit crimes in cyberspace, digitally spar with human hackers, and launch attacks with kinetics effects, all while grappling with existential issues of self and other philosophical quandaries.[4] Despite its almost shocking overlap with the trajectory of technological development, specifically artificial intelligence, *Neuromancer*'s haunting future world has not yet fully materialized. But, as former Assistant Attorney General John Carlin recently noted: "I think it's instructive now to look to movies, to look to science-fiction as we try to think what the next threats are going to be and how we can prepare for them."[5]

The rise of artificial intelligence ("AI") has changed, is changing, and will change the world, from politics, to social

---

[2] Gibson continued to issue prophetic and hyper-relevant quotes about the present and future of technology, even twenty years after the release of his influential novel *Neuromancer*: " 'Cyberspace' is a word that's increasingly long in the teeth as the reality becomes more ubiquitous by the day." Joseph Walsh, *Meeting William Gibson, the Father of Cyberpunk*, VICE (Dec. 5, 2014), https://www.vice.com/sv/article/bn5k5m/william-gibson-interview-399 [https://perma.cc/EKZ9-XBUQ].

[3] *See generally* WILLIAM GIBSON, NEUROMANCER (Penguin Random House 2018) (1984). In this science-fiction masterpiece, an "artificial general intelligence" ("AGI") entity called "Wintermute," hires a human hacker to achieve his only purpose for existence, to merge with another AGI called "Neuromancer."

[4] *Id.*

[5] *Tomayto, Tomahto: Right to Be Forgotten Meets Right to Die*, CYBERLAW PODCAST (Jan. 29, 2019), https://www.steptoe.com/feed-Cyberlaw.rss [https://perma.cc/7DX8-8G6P].

interaction, to economics. However, it is important to note AI will have a tremendous effect on governments and nation-States ("States"), particularly on the difficulties AI will pose to national security. The current presidential administration's recent decree[6] highlights the importance of AI dominance to the United States' national security interests.

The application of AI to national security related matters ranges from autonomous weapons systems[7] to AI-powered facial recognition[8] to countering terrorist recruitment using AI and machine learning.[9] The intersection between cybersecurity and AI is a major area of concern.[10] The past decade has seen an exponential increase of malicious cyber-activities, orchestrated by both State and non-State actors alike. Indeed, the Director of National Intelligence ("DNI") announced "cyber" to be the first global threat in this year's Worldwide Threat Assessment:

> Our adversaries and strategic competitors will increasingly use cyber capabilities – including cyber espionage, attack, and influence – to seek political, economic, and military advantage over the United States . . . . China, Russia, Iran, and North Korea increasingly use cyber operations to threaten both minds and machines in an expanding number of ways –

---

[6] Exec. Order No. 13,859, 3 C.F.R. § 3967 (2019) ("Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities.").

[7] *See generally* PAUL SCHARRE, ARMY OF NONE, AUTONOMOUS WEAPONS AND THE FUTURE OF WAR (2018).

[8] *See, e.g.*, Sahil Chinoy, *We Built 'Unbelievable' (but Legal) Facial Recognition Machine*, N.Y. TIMES (Apr. 16, 2019), https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html [https://perma.cc/SA67-QTK4].

[9] Natasha Lomas, *Google to Ramp Up AI Efforts to ID Extremism on YouTube*, TECHCRUNCH (June 19, 2017), https://techcrunch.com/2017/06/19/google-to-ramp-up-ai-efforts-to-id-extremism-on-youtube/ [https://perma.cc/5YP5-QAQE].

[10] Michael C. Horowitz et al., *Artificial Intelligence and International Security*, CTR. FOR A NEW AMERICAN SEC. (July 2018), https://s3.amazonaws.com/files.cnas.org/documents/CNAS-AI-and-International-Security-July-2018_Final.pdf?mtime=20180709122303 [https://perma.cc/K42U-5QH4].

> to steal information, to influence our citizens, or to disrupt critical infrastructure."[11]

In a recent talk, former FBI Director James Comey reflected on threats from his tenure: "we see an explosion in nation-state adversaries and near nation-state actors using the digital vector to steal all kinds of information and to prepare things that were near to kinetic acts of war."[12]

Indeed, "attribution," that is, determining *who*, or potentially *what*, is responsible for a malicious cyber-activity, is already an incredibly difficult task. There are three types of attribution: political,[13] technical, and legal. "Technical attribution" is characterized "as determining the identity or location of an attacker or an attacker's intermediary."[14] Legal attribution "refers to the assignment of responsibility for an 'internationally wrongful act to a state' " or non-State actor.[15] Legal attribution requires some degree of technical attribution; there must be some evidence linking an actor to the cyber-attack, otherwise the cyber-attack cannot be qualified as State sponsored or not. Thus, it is important to understand the strategies for technical attribution, as well as the

---

[11] *Worldwide Threat Assessment of the U.S. Intelligence Community: Hearing Before the S. Select Comm. on Intelligence*, 116th Cong. (2019) (statement of Daniel R. Coats, Director of National Intelligence), https://www.hsdl.org/?view&did=820727 [https://perma.cc/TQ3N-6JWP] [hereinafter "Worldwide Threat Assessment"].

[12] *Bonus Edition: James Comey at Verify 2019*, LAWFARE PODCAST (Apr. 11, 2019, Minute 7:20), https://www.lawfareblog.com/lawfare-podcast-bonus-edition-james-comey-verify-2019 [https://perma.cc/5A6E-ZEME].

[13] Political attribution refers to the decision from a diplomatic or policy standpoint to assign blame to a particular State, group, or individual for a cyber-event. The question of political attribution is a foreign relations decision, and irrelevant for the purposes of the instant analysis.

[14] DAVID A. WHEELER & GREGORY N. LARSEN, INST. FOR DEF. ANALYSIS, TECHNIQUES FOR CYBER ATTACK ATTRIBUTION (Oct. 2003), https://apps.dtic.mil/dtic/tr/fulltext/u2/a468859.pdf [https://perma.cc/S6YK-63VE].

[15] Jason Jolley, *Attribution, State Responsibility, and the Duty to Prevent Malicious Cyber-Attacks in International Law* (Oct. 21, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3056832 [https://perma.cc/469P-4KHN].

mechanism used by malign cyber actors to frustrate any efforts to link them to a particular cyber operation.

The emergence of AI has further complicated cyber-attribution issues. Last year, the first AI driven cyber-attack was reported in India. The attack "used rudimentary machine learning to observe and learn patterns of normal user behavior inside a network . . . then began to mimic that normal behavior, effectively blending into the background and becoming harder for security tools to spot."[16] These type of machine-learning powered attacks are currently rare, but their emergence signal a worrisome potential trend because they could permit malign actors to threaten critical infrastructure like powerplants or nuclear facilities, steal personal data on a massive scale, or shut down or steal money from financial institutions.

Identifying the human actor who directed an attack from behind the computer was already difficult to accomplish, often requiring intelligence gleaned from human spies and electronic sources, in addition to the legal authority necessary to trace the code "breadcrumbs" through foreign cyberspace. Forensic evidence of a malicious cyber-activity's origin could be masked but never totally erased. Enter AI. A decentralized AI system,[17] created, released, and acting on its own without any direction from a human engineer or creator, could substantially blur any attributive link to an actor or State.

Even if a government could point a finger at an intangible, cyberspace-bound, decentralized autonomous AI entity (a "CyDAE"), what legal authority does a State have to stop a CyDAE's malicious cyber-activities? For purposes of jurisdiction, is the CyDAE "located" where a majority of its servers are located, or is it where the data it uses is located? Is a CyDAE's nationality

---

[16] Steven Norton, *Era of AI-Powered Cyberattacks Has Started*, WALL ST. J. (Nov. 15, 2017), https://blogs.wsj.com/cio/2017/11/15/artificial-intelligence-transforms-hacker-arsenal/ [https://perma.cc/GV57-UGNW].

[17] Per Ocean Protocol founder Trent McConaghy, data localization regulations that require data sets be siloed in a particular State will "massively affect traditional AI[, b]ut not decentralized AI," because the "AI compute . . . comes to the data." Trent McConaghy, (@trentmc0), TWITTER, (Mar. 8, 2019), https://twitter.com/trentmc0/status/1104049106220138506 [https://perma.cc/F39N-FGG3].

correspondent to the nationality of its creators, if their identities can be ascertained? This AI arrangement adds yet another layer of ambiguity in terms of attribution. How can we enable action to combat and/or regulate the use of AI in terms of cybersecurity? Developing a way to "point the finger"—i.e. impose responsibility upon AI, specifically CyDAE, or its handlers—is of paramount importance.

Going forward in this discussion, AI responsibility is essential to the different levels of AI development. The term "AI," on the lower-end of the intelligence spectrum, means "systems that can emulate, augment, or compete with the performance of intelligent humans in well-defined tasks."[18] On the higher end of the intelligence spectrum is "artificial general intelligence" ("AGI"), which means a " 'strong' [AI] with the full range of cognitive capacities typically possessed by humans, including self-awareness" as is usually depicted in science fiction.[19] The AI discussed in this paper fall somewhere between the two extremes.

This article will provide a brief technological breakdown about AI systems. Next, the article will discuss some legal personhood theories for autonomous AI systems and then summarize the law of attribution. To tie it all together, the next section will use a CyDAE example to demonstrate various kinds of cyber-activities that could be carried out by AI. Penultimately, the article will apply the current attribution framework to highlight the difficulty CyDAEs will pose in a legal context. The discussion will end with four bold and potentially provocative general proposals in order to guide policy vectors. The proposals are: (1) mandatory registration of AI systems; (2) a standard of AI system explainability; (3) the creation of legal personhood arrangements that require human control; and (4) universal jurisdiction for AI that fail to abide by these legal standards. Although the threat outlined in this article seems far-

[18] SHANNON VALLOR & GEORGE BEKEY, ROBOT ETHICS 2.0: FROM AUTONOMOUS CARS TO ARTIFICIAL INTELLIGENCE, ARTIFICIAL INTELLIGENCE AND THE ETHICS OF SELF-LEARNING ROBOTS 339 (Patrick Lin, Ryan Jenkins, & Keith Abney eds., 2017). This book contains some of the best and most accessible descriptions, summaries, and explanations of AI systems.

[19] *Id.* at 339–40.

fetched, we cannot allow the next major threat to our society to emerge from a failure of imagination.

II. BRIEF TECHNOLOGY BREAKDOWN

This section will define and summarize different conceptual underpinnings. At least some AI experts consider AI to be "a set that contains machine learning (ML), and deep learning (DL)."[20] Therefore, an independent, decentralized artificial intelligence capable of engaging in malicious cyber-activities would likely be technologically sophisticated, so understanding how the technology works may help us understand the *legal* ramifications of such activities. The same is true for deterring, detecting, and combatting malicious cyber-activities: without a baseline technical understanding, relevant legal frameworks cannot be applied meaningfully.

One of the first approaches used to create AI was a "rule-based" method, that is, a programmer would create a set of rules that the system would have to check for each decision or instance of learning. While rule-based methods have important uses, "[t]rying to hand-code a set of rules for a machine . . . to visually distinguish between an apple and a tomato [for example] would be challenging. Both objects are round, red, and shiny with a green stem on top."[21] Rule-based approaches are considered "top-down" because of how the over-arching rules are applied to the learning process. Rather than a top-down, rule-based approach, typical AI models utilize "bottom-up" approaches: complex mathematical formulas known as "algorithms" parse millions of pieces of data in search for patterns. The exponential increase in computing power and availability of voluminous categorized datasets opened the door for these breakthrough techniques in AI system design.

---

[20] Oludare Isaac Abiodun et al., *State-of-the-Art in Artificial Neural Network Applications: A Survey*, 4 HELIYON 11 (2018).

[21] PAUL SCHARRE & MICHAEL C. HOROWITZ, CTR. FOR A NEW AM. SEC., ARTIFICAL INTELLIGENCE: WHAT EVERY POLICYMAKER NEEDS TO KNOW (2018), https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know [https://perma.cc/6UL3-A9TH].

A. AI Uses and Advantages

What about AI in general makes it appealing for human productivity? "In some cases, their value may come from being cheaper, faster, or easier to deploy at scale relative to human expertise."[22] Beyond general qualities like "data classification," "detection," "prediction," and "optimization" of efficiency, AI systems seem to be trending toward "faster-than-human reaction times"; "superhuman precision and reliability"; "superhuman patience and vigilance"; as well as ability to conduct "operations without connections to humans."[23] The utility of Artifical Neural Networks ("ANNs") in particular transcend human capabilities in a myriad of fields including "computer security, medical science, business, finance, bank[ing], insurance, the stock market, electricity generation, management, nuclear industry, mineral exploration, mining, crude oil fractions quality prediction, crops yield prediction, water treatment, and policy."[24]

B. Machine Learning

Machine learning is "a developmental process in which repeated exposures of a system to an information-rich environment gradually produce, expand, enhance, or reinforce that system's behavioral and cognitive competence in that environment or relevantly similar ones."[25] In simple terms, "[g]iven a goal, learning machines adjust their behavior to optimize their performance to achieve that goal."[26] So, with regard to the example of a tomato and an apple above:

> [A]n algorithm might take as input millions of labeled images, such as "dog," "person," "apple." The algorithm then learns subtle patterns within the images to distinguish between categories – for example, between an apple and a tomato . . . . Given enough labeled images of both, machines can also learn these differences and then distinguish between an apple and a tomato when they are not labeled.[27]

---

[22] *Id.*

[23] *Id.*

[24] Abiodun et al., *supra* note 20, at 20. For the authors' assessment of ANNs' applicability in each field on various data analysis factors, see tbl. 1, fig. 7, at 19–20.

[25] VALLOR & BEKEY, *supra* note 18, at 340.

[26] SCHARRE & HOROWITZ, *supra* note 21.

[27] *Id*.

As AI system design improved, sub-types of machine learning proliferated, although in-depth descriptions of those methods are beyond the scope of this article.[28] Notably, however, the current state of machine learning is far from the level of sophistication needed for an advanced, standalone CyDAE that would be able to exist completely independent from human handlers.

C. Artificial Neural Networks ("ANNs") & Deep Learning

Considering the plasticity and adaptability of the human brain, it is no wonder that some forms of machine learning borrow concepts from human neuroscience. Simply, "[h]uman brains are made up of connected networks of neurons . . . . ANNs seek to simulate these networks and get computers to act like interconnected brain cells, so they can learn and make decisions in a more humanlike manner."[29] For instance, "the network gradually 'learns' from repeated 'experience' (multiple training runs with input datasets) how to optimize the machine's 'behavior' (outputs) for a given kind of task."[30]

These ANNs, much like the human brain, can create stronger or weaker associations between connections in the hidden layers, which will result in the AI system's behavior adapting and adjusting

---

[28] There are several subtypes of machine learning: supervised, unsupervised, reinforcement, etc. For a quality discussion, see SCHARRE & HOROWITZ, *supra* note 21. "Supervised" means the extent to which the training data is explicitly labeled by humans to tell the system which classifications it should learn to make (as opposed to letting the system construct its own classification or groupings)." VALLOR & BEKEY, *supra* note 18, at 341. Alternatively, "unsupervised" machine learning is essentially a programmed form of trial-and-error, without outside help: "unsupervised learning methods form clusters or groups between and among the objects in an area to identify likeness, then use similarity for classifying unknowns." Abiodun et al., *supra* note 20, at 10. Some subtypes of supervised learning models include single-layer perception; multi-layer perception; linear classifiers; support vector machines; k-nearest neighbors; Bayesian statistics; decision trees; and hidden Markov models. *Id.* at 11. Some unsupervised learning model sub-types include k-means; expectation maximization; auto-encoders; density-based models; self-organizing maps; and clustering. *Id.*

[29] Bernard Marr, *Deep Learning Vs Neural Networks – What's the Difference?*, BERNARD MARR & CO., https://bernardmarr.com/default.asp?contentID=1789 [https://perma.cc/833M-4RY5].

[30] VALLOR & BEKEY, *supra* note 18, at 340 (emphasis in original).

to changing scenarios.[31] "Whereas machine-learning algorithms require the features they look for in data to be pre-set, deep-learning neural net[works] can determine and detect salient features on their own."[32] Self-driving cars are a useful example. If the system learned what a "bicyclist" is from various dataset depicting or describing bicyclists being input into the system over time, then when the system detects a bicyclist, it will learn to adjust its behavior over time, and eventually be able to "slow down slightly, edge to the left-center of the lane."[33]

Creating and implementing "hybrid" designs of AI systems that incorporate overarching "top-down" rules to govern the "bottom-up" processes, are critical to the future regulation of AI.[34] This kind of "hybrid" approach will give developers greater control over their AI systems:

> The potential for the misalignment of interest [between the AI system's objectives and those of the public at large] flows from the fact that an AI's objectives are determined by its initial programming. Even if that initial programming permits or encourages the AI to alter its objectives based on subsequent experiences, those alterations will occur in accordance with the dictates of the initial programming . . . [which] seems beneficial in terms of maintaining control. After all, if humans are the ones doing the initial programming, they have free rein to shape the AI's objectives.[35]

D. Black Box and Explainability Issues

The "black box" or "explainability" problem is a major hurdle for AI developers.[36]

> The term 'black box' has long been used in science and engineering to denote technology systems and devices that function without divulging

---

[31] *Id.* at 341.

[32] John Fletcher, *Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance*, 70 THEATRE J. 455, 459 (2018).

[33] VALLOR & BEKEY, *supra* note 18, at 341.

[34] *See* WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 117 (2009).

[35] Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 367 (2016).

[36] For an in-depth discussion of built in morality and ethical rules in AI systems design, see WALLACH & ALLEN, *supra* note 34, at 73.

> their inner workings. The inputs and outputs of the 'black box' system may be visible, but the actual implementation of the technology is opaque, hidden from understanding or justifiability.[37]

Put another way, when necessary to understand either the programming, coding, or motives of a particular AI, the AI system's process is often so opaque because of the sheer complexity of the code, or by an advertent wall created by the programmers to obfuscate that code. Opacity of an AI system means "the inner workings of an AI system may be kept secret and may not be susceptible to reverse engineering."[38] "The 'black box' concept has been exploited by the likes of Silicon Valley start-ups to Wall Street investment firms, usually in their efforts to protect intellectual property and maintain competitiveness."[39] Opaque code should not be permitted solely in order to protect proprietary information, shield a company or individual from liability, or evade detection for some insidious or criminal reason.

But, opacity of AI systems may not purely be based on the designers' intent to shield the inner workings of their code from view, but might instead be a symptom of the complexity of AI system technology. At least one scholar has articulated the difficulty AI system designers must face when balancing their systems' complexity, transparency, proprietary information security, explainability, and functionality: if algorithms can "be so complex that meaningful transparency is impossible . . . [s]hould robots be designed to be 'closed,' in the sense that they have a set, dedicated function and run only proprietary software . . . [o]r can companies design robots to be 'open' without incurring liability?"[40] Black box AI would likely present difficulties in terms of government audits,

---

[37] KYNDI, HOW 'EXPLAINABILITY' IS DRIVING THE FUTURE OF ARTIFICIAL INTELLIGENCE 2 (Jan. 2018), https://kyndi.com/wp-content/uploads/2018/01/Kyndi-final-Explainable-AI-White-Paper.pdf [https://perma.cc/2HJJ-2M57].

[38] Scherer, *supra* note 35, at 369.

[39] KYNDI, *supra* note 37.

[40] Woodrow Hartzog, *Unfair and Deceptive Robots*, 74 MD. L. REV. 785, 809–21 (2015). While the issue of liability is only tangentially related to this discussion, it is important to note that the manner in which AI creators program intentionality or motivation *will* come up in future legal discussions due to the inextricable and unavoidable nexus between mental state and culpability.

"especially crucial for critical organizations that are required to explain the reason for any decision."[41]

Moreover, the explainability problem may limit the ability of programmers and creators to know the source of a problem with the AI system's function. In a discussion of machine learning algorithms and facial recognition, professor Nick Weaver noted the following serious issues for the technology:

> When applied to face recognition there are huge biases turning up . . . we don't know whether this is biases in the training set or if there actually might be technical or cultural features or some other aspects that are resulting in these biases and we can't because these systems are designed as unknowable black boxes.[42]

But, policy and legislation in western democracies seems to highlight why an emphasis on explainable AI systems may benefit society. The U.S. Defense Advanced Research Projects Agency ("DARPA"), a research agency within the U.S. Department of Defense ("DoD"), spearheads AI initiatives that have already encountered such problems, succinctly describing "black box" issues they expect to confront in developing autonomous weapons systems:

> Continued advances [in AI] promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems is limited by the machine's current inability to explain their decisions and actions to human users . . . . [DoD] is facing challenges that demand more intelligent, autonomous, and symbiotic systems. Explainable AI – especially explainable machine learning – will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners.[43]

In the European Union ("EU"), the General Data Protection Regulation ("GDPR") imposes explainability requirements for automated systems writing: "the data controller shall implement suitable measures to safeguard the data subject's rights and

---

[41] KYNDI, *supra* note 37, at 7.

[42] *Death of Section 230*, THE CYBERLAW PODCAST (Apr. 8, 2019), https://www.steptoe.com/feed-Cyberlaw.rss [https://perma.cc/7DX8-8G6P].

[43] Matt Turek, *Explainable Artificial Intelligence (XAI)*, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY, https://www.darpa.mil/program/explainable-artificial-intelligence [https://perma.cc/UET8-FETE].

freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."[44] Without using the phrase "explainability," the GDPR codifies the requirement that a system be explainable by requiring human review of any AI system determinations. Although it did not become law, the EU parliament proposed a resolution about AI and robotics, noting the importance of transparency for AI systems, highlighting

> the principle of transparency, namely that it should always be possible to supply the rationale behind any decision taken with the aid of AI that can have substantive impact on one or more persons' lives; considers that it must always be possible to reduce the AI system's computations to a form comprehensible by humans; considers that advanced robots should be equipped with a 'black box' which records data on every transaction carried out by the machine, including the logic that contributed to its decisions.[45]

Issues surrounding black box algorithms used in popular social media platforms have prompted self-reflection by the controllers of those platforms,[46] but also criticism from scholars who point how innovation might suffer from black box regulation. One critic explained that attempting to regulate unexplainable AI systems

---

[44] GDPR, Ch. 3, Art. 22, § 3 (emphasis supplied). Indeed, the eponymous term "controller" means a "natural or legal person" that defines the parameters of data processing for an autonomous system. Thus, a controller under the GDPR's jurisdiction must be able to give a data subject an explanation as to why an autonomous system arrived at its conclusion.

[45] *European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics*, Civil Law Rules on Robotics (2015/2103 (INL)) at ¶12.

[46] *See*, *e.g.*, Jason Bloomberg, *Don't Trust Artificial Intelligence? Time to Open the AI "Black Box,"* FORBES (Sept. 16, 2018), https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/#103079dd3b4a [https://perma.cc/7S9S-7RLQ] (quoting Twitter CEO Jack Dorsey: "[w]e need to do a much better job at explaining how our algorithms work . . . [i]deally opening them up so that people can actually see how they work."). Also, in a discussion about Google's "black box" YouTube search algorithm, former NSA General Counsel and former Assistant Secretary for Policy at DHS, Stewart Baker said "[w]e are never going to know how this [search algorithm] works. Google is basically saying 'trust us, we'll do the right thing . . .' I have zero faith in YouTube's willingness to play it straight." *Tomayto, Tomahto: Right to be Forgotten Meets Right to Die*, *supra* note 5.

"significantly raises labor costs and thus creates a strong disincentive from using AI – as a main reason for developing AI in the first place is to automate functions that would otherwise be much slower, costlier, and more difficult to complete if performed by humans."[47]

Notwithstanding the policy arguments in favor of or against explainable AI systems in the current state of the technology, the opacity of unexplainable systems will pose serious problems to investigators seeking to make attributive links through analysis of AI system processes.

E. Decentralization

In order to understand how AI use and consume data, it is essential to understand how data is stored. Data can be stored in a "centralized" way, meaning the data is contained on a single server, hard drive, or network, or, alternatively, controlled by a single entity.

> Data can be processed simultaneously in multiple locations [by that entity]; dispersed for storage around the globe; re-combined instantaneously; and moved across borders by individuals carrying mobile devices . . . [s]ervices, such as 'cloud computing,' allow [organizations] and individuals to access data that may be stored anywhere in the world.[48]

If we analogize data to grain, "centralized" storage of grain might be that all of a Farmer's grain is stored in one warehouse, or in one silo, in a building on one farm, all controlled by a single Farmer. In contrast, data can be stored in a "decentralized" fashion, meaning the data need not be contained in a single, discreet location. Rather, data can be separated, and stored on thousands of different networks,

---

[47] NICK WALLACE & DANIEL CASTRO, CTR. FOR DATA INNOVATION, THE IMPACT OF THE EU'S NEW DATA PROTECTION REGULATION ON AI 2 (Mar. 27, 2018), http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf [https://perma.cc/PB2C-NFUJ].

[48] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, THE OECD PRIVACY FRAMEWORK: SUPPLEMENTARY EXPLANATORY MEMORANDUM TO THE REVISED OECD PRIVACY GUIDELINES 29 (2013), https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf [https://perma.cc/N7FL-TFP2].

and still retrieved later.[49] Using the grain example above, decentralized grain storage would mean that the specific individual pieces of the Farmer's grain are stored in many silos, many warehouses, or across many farms, rather than just in the Farmer's silo. The grain analogy fails, however, when it comes to retrieval of data versus retrieval of grain in a decentralized paradigm. It would be nearly impossible for the Farmer to separate each individual piece of his/her grain and retrieve all those exact pieces of grain. However, with modern computing and advances in technology, data can be separated, dispersed across myriad networks, and retrieved without fail. Importantly, decentralization does not apply solely to data; any network protocol, function, or transmission of information can implement decentralization—whether it be computing, financial transactions, or communications.

Many characteristics of decentralization make it an attractive approach to computing and storage of data. Vitalik Buterin identifies, in relevant part, two key reasons for decentralization:

> **Fault tolerance** – decentralized systems are less likely to fail accidentally because they rely on many separate components . . . [and]
>
> **Attack resistance** – decentralized systems are more expensive to attack and destroy or manipulate because they lack sensitive central points that can be attacked at a much lower cost . . . .[50]

From a practical perspective, decentralization happens when computer or network operators install a particular program or protocol's software and act as a "node," that is, software on a computer that participates as one point in a network of computers associated by the commonly-installed program. Thus, if one specific node is attacked or breaks down, unlike a single centralized network

---

[49] The creator of the Ethereum Network, Vitalik Buterin, and one of the foremost thinkers in the blockchain and decentralized computer space, in a comprehensive and technical discussion of decentralization, identifies three types of decentralization in terms of computer networks: 1) architectural, that is, the number of physical computers; 2) political, that is, how many individuals control the computer network; and 3) logical, whether the interface and data structures that the system presents and maintains look "more like a single monolithic object, or an amorphous swarm." Vitalik Buterin, *The Meaning of Decentralization*, MEDIUM (Feb. 6, 2017), https://medium.com/@VitalikButerin/the-meaning-of-decentralization-a0c92b76a274 [https://perma.cc/R9FR-2YW8].

[50] *Id.*

node, a decentralized network can still function even if "five out of ten computers" fail simultaneously.[51] "[T]he principle is uncontroversial, and is used in real life in many situations, including jet engines, backup power generators particularly in places like hospitals, military infrastructure, [and] financial portfolio diversification . . . ."[52] Simplistically, decentralization creates a vast fail-safe web of nodes rather than consolidating information or computing power in a single place or on a single network.

A decentralized AI system would thus benefit from fault tolerance and attack resistance, much like decentralized data platforms do. The result would be an AI system with an uncertain physical location, immunity to cyberattack, and temporal longevity because of the unlikely possibility of accidental failure.

Taken together, the attributes described above paint a potentially troubling picture of an independent entity that, through deep neural networks, learns on its own and prevents observers from understanding how it works technically because its code is obscured by an unexplainable black box, all while being resistant to attacks and difficult to locate by virtue of its networks' decentralization.

III. EMERGENCE OF CYDAES AS LEGAL PERSONS

In the recent past, Saudi Arabia granted citizenship to a "female" AI called "Sophia,"[53] but the act by the kingdom appears to be more ceremonial and symbolic than legal in nature. Trent McConaghy, a notable AI researcher, presciently opined about "Decentralized Autonomous Organizations" ("DAOs") and their inevitable joinder with AI, "DAOs have arrived . . . [a]nd when artificial intelligence gets added to the mix, the results are explosive."[54] Similarly, one

---

[51] *Id.*

[52] *Id.*

[53] Dom Galeon, *World's First AI Citizen in Saudi Arabia Is Calling for Women's Rights*, SCIENCEALERT (Dec. 15, 2017), https://www.sciencealert.com/first-ai-citizen-saudia-arabia-womens-rights [https://perma.cc/R8XT-X35G].

[54] Trent McConaghy, *AI DAOs, and Three Paths to Get There*, MEDIUM (June 18, 2016), https://medium.com/@trentmc0/ai-daos-and-three-paths-to-get-there-cfa0a4cc37b8 [https://perma.cc/3VDT-B4J7].

legal scholar paints the "explosive results" of AI DAOs (he calls them "algorithmic entities") in a much more grim light:

> [b]ecause they lack human bodies, [algorithmic entities] are harder to catch and impossible to punish. [They] need not fear death or capture. They can replicate themselves without ego and sacrifice themselves without motive. They need not recoil at the necessity to do violence to humans.[55]

So, how will the law handle these non-human entities? How do we get to a world with AI existing as legal persons? This section will discuss a few ways that an AI might be structured as a legal person.

A. Legal Personhood

"[I]t is unlikely that, in a future society where artificial agents wield significant amount of executive power, anything would be gained by continuing to deny them legal personality."[56] Denying AI legal personality out of vanity is one issue, but would it not actually be beneficial for people to categorize and legalize AI entities so they fit in our rigid legal paradigms? Notwithstanding current technical limitations that prevent AI systems from becoming truly autonomous and/or self-governing, the state of the law now probably precludes AI systems from becoming independent legal entities.[57] But, importantly, the issue of legal personhood arises in discussions about who is liable for the acts of AI systems. Some scholars rely on "agency" as the legal framework to support liability when people are hurt by AI systems,[58] while others point to legal personhood arrangements.[59] But, other theoretical and hypothetical proposals for legal personhood arrangements for AI or otherwise autonomous systems are not so far-fetched.

---

[55] Lynn M. LoPucki, *Algorithmic Entities*, 95 WASH. U.L. REV. 887, 891–92 (2018).

[56] SAMIR CHOPRA & LAWRENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS 191 (2011).

[57] Law professor Shawn Bayern disagrees. *See generally* Shawn Bayern, *The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems*, 19 STAN. TECH. L. REV. 93 (2015) (discussing how existing business entity law in the United States suffices for bestowing legal personhood on autonomous systems).

[58] *See* WALLACH & ALLEN, *supra* note 34; CHOPRA & WHITE, *supra* note 56.

[59] *See generally* Bayern, *supra* note 57; LoPucki, *supra* note 55.

B. Autonomous Self-Owning Cars

One interesting way that people have imagined AI existing independently from human beings is in the context of driverless, autonomous rideshares that own themselves.[60] Much like "the DAO" and AI DAOs discussed below, these driverless cars would accept cryptocurrency as payment, and use the proceeds of their "work" to buy fuel, purchase software updates, and pay premiums into an insurance pool with other autonomous self-driving cars.[61] One proposal, "car-ception," would allow an AI-powered car to

> slowly save up enough money to afford buying a new model to put on the road at some point in the future . . . Since the old car is buying the new car, it will technically be the new [car's] owner. It can also arrange to have all of its remaining wealth transferred to the new [car's] digital wallet . . . .[62]

Indeed, one of the similarities between these autonomous driverless car models is that they would operate in conjunction with decentralized platforms, particularly blockchain platforms.[63]

C. "The DAO"

"*The* DAO," not to be confused with DAOs generally, was both a breakthrough and a failed experiment in decentralized technology in cyberspace. The DAO was an investment fund built on the Ethereum blockchain in which "[p]articipants buy in to the fund by purchasing digital tokens, then introduce, view, and vote on pitches; the company's smart code then automatically executes winning

---

[60] *See, e.g.*, Leo Kelion, *Could Driverless Cars Own Themselves?*, BBC (Feb. 16, 2015), https://www.bbc.com/news/technology-30998361 [https://perma.cc/YP45-MNUL].

[61] Chris Czupak, *The Self Owning Car*, MEDIUM (May 8, 2018), https://medium.com/@chris_czupak/the-self-owning-car-dd1b39b95748 [https://perma.cc/RTE5-WP9P].

[62] *Id.*

[63] Thomas Birr & Carsten Stöker, *Goodbye Car Ownership, Hello Clean Air: Welcome to the Future of Transport*, WORLD ECON. FORUM (Dec. 16, 2016), https://www.weforum.org/agenda/2016/12/goodbye-car-ownership-hello-clean-air-this-is-the-future-of-transport/ [https://perma.cc/FN7N-L7FR].

projects" using the tokens they purchased.[64] The Securities and Exchange Commission ("SEC") described it as:

> a 'virtual' organization embodied in computer code and executed on a distributed ledger or blockchain. The DAO was created . . . with the object of operating as a for-profit entity that would create and hold a corpus of assets through the sale of DAO Tokens to investors, which would then be used to fund 'projects.' The holders of DAO Tokens stood to share in anticipated earnings from these projects as a return on their investment . . . .[65]

"The DAO" managed to raise $150 million in a private sale of its digital tokens, but hackers exploited its programming, resulting in the theft of approximately $50 million.[66] The DAO's reliance on human beings separates it from Trent McConaghy's vision of future AI DAOs, and at least one legal scholar has proclaimed the DAO's name a misnomer because its arrangement lacked true autonomy.[67]

D. The "ArtDAO"

McConaghy suggests a canvas for a reasonable application of a decentralized autonomous organization governed by AI, which he calls the "ArtDAO."[68] In essence, McConaghy's "ArtDAO recipe" requires an AI process (ANNs, etc.), decentralized by virtue of the

---

[64] Ori Oren, *ICO's, DAO's, and the SEC: A Partnership Solution*, 2018 COLUM. BUS. L. REV. 617, 618 (2018). For a full and cogent discussion of the DAO, blockchain and smart contracts, etc., see Laila Metjahic, *Deconstructing the DAO: The Need for Legal Recognition and the Application of Securities Laws to Decentralized Organizations*, 39 CARDOZO L. REV. 1533, 1544–45 (2018).

[65] U.S. SEC. & EXC. COMM., REPORT OF INVESTIGATION PURSUANT TO SECTION 21(A) OF THE SECURITIES EXCHANGE ACT OF 1934: THE DAO (2017), https://www.sec.gov/litigation/investreport/34-81207.pdf [https://perma.cc/5BJ4-LHZ7].

[66] For a more in-depth discussion of The DAO's background see Cristoph Jentzsch, *History of the DAO and Lessons Learned*, MEDIUM (Aug. 24, 2016), https://blog.slock.it/the-history-of-the-dao-and-lessons-learned-d06740f8cfa5 [https://perma.cc/6HF3-EBXW]; Brian Yurcan, *Despite the DAO-saster, its creators raise $2M*, AM. BANKER (Mar. 30, 2017), https://www.americanbanker.com/news/despite-the-dao-saster-its-creators-raise-2m [https://perma.cc/4843-U3UL].

[67] Metjahic, *supra* note 64.

[68] Trent McConaghy, *Wild, Wooly, AI DAOs*, MEDIUM (June 22, 2016), https://medium.com/@trentmc0/wild-wooly-ai-daos-d1719e040956 [https://perma.cc/|R8RD-H927].

blockchain, to generate artistic images.[69] The AI then "claims attribution of the image in a time-stamp to the blockchain" and, after creating multiple editions, "posts those editions for sale onto a marketplace."[70] After, "[i]t sells the editions" and "transfers the proceeds from the buyer to ArtDAO" using cryptocurrency, then transfers the rights to the art to the buyer.[71] As it continues to create new digital art, the ArtDAO earns proceeds in cryptocurrency. The imaginative, and winsome ArtDAO seems like a trouble-free manifestation of a CyDAE, but as an example, it demonstrates how easily an independent and autonomous AI entity could exist without human guardianship or oversight.[72]

The ability for an autonomous system to manage money, make decisions for itself, independent of human oversight or approval, all with a particular goal in mind will underpin the structure of CyDAEs. These traits that allow AI systems to exist independently will function together to obfuscate any CyDAE's connection human handlers and make attribution of its actions challenging.

IV. LAW OF ATTRIBUTION OF CYBER-ACTIVITIES

Understanding the difficulties that a future CyDAE would pose to the national security of the United States requires a discussion of international legal principles surrounding attribution. If a future CyDAE were stateless and its location were impossible to determine, or if it were linked to a foreign State, then international

---

[69] *Id.*

[70] *Id.* A full discussion of "blockchain" is outside the scope of this article, but the blockchain is an immutable, public, secure, and decentralized ledger of transactions, existing simultaneously on many computers. Blockchain popularized and most often contextualized in terms of cryptocurrencies, specifically, Bitcoin. McConaghy's ArtDAO relies upon the Ethereum blockchain, which is essentially a protocol allowing the execution of simple, automatic commands called "smart contracts." *See* Metjahic, *supra* note 64, for an in-depth and cogently written discussion of blockchain and smart contracts.

[71] McConaghy, *supra* note 68.

[72] The ArtDAO illuminates substantial legal impediments in the current legal framework, particularly in terms of personhood, discussed in Part III, Subsection D.

legal principles would apply.[73] Prominently featured as the first discussion point in the 2018 Worldwide Threat Assessment delivered to Congress, the DNI declared that "[t]he risk is growing that some adversaries will conduct cyber[-]attacks . . . against the United States in a crisis short of war."[74] Notably, the DNI posits that the threat is one "short of war," meaning the threat derives from cyber-attacks that do not rise to the level of armed attack or use of force,[75] and thus principles of *jus ad bellum* or *jus in bello* do not apply.[76] The international legal community agrees that the law of State Responsibility applies, in this cyber context, under Customary International Law ("CIL") and the Draft Articles on Responsibility for Internationally Wrongful Acts ("Draft Articles") that codified CIL. However, the CIL and Draft Articles are not apt instruments for dealing with cyberspace issues. The CIL and Draft Articles evolved based on State behavior in an analog world rigidly attached to traditional norms of territory and sovereignty, whereas cyberspace emerged subsequent to those instruments as a dimension that both transcends and undermines those traditional norms.

A. Three Types of Attribution

There are three types of attribution: political,[77] technical, and legal. "Technical attribution" is characterized "as determining the identity or location of an actor or an actor's intermediary."[78] "Legal attribution," in terms of State action, "refers to assignment of responsibility for an 'internationally wrongful act to a State.'"[79]

---

[73] United States domestic authorities grant the intelligence community and military broad discretion in defending against malicious cyber-activities, so for the purposes of this article, domestic legal authority is assumed.

[74] *Worldwide Threat Assessment of the U.S Intelligence Community: Hearing Before the S. Comm. on Intelligence*, 115th Cong. 5 (2018) (statement of Daniel R. Coats, Director of National Intelligence).

[75] *Id.*; *see also* UN Charter arts. 2(4); 51.

[76] *Jus ad bellum* refers to the legal rules governing whether a State is justified engaging in warfare and *jus in bello* refers to the rules that apply during war.

[77] Political attribution refers to the decision from a diplomatic or policy standpoint to assign blame to a particular State, group, or individual for a cyber-event. The question of political attribution is a foreign relations decision, and irrelevant for the purposes of the instant analysis.

[78] Wheeler & Larsen, *supra* note 14, at 1.

[79] Jolley, *supra* note 15, at 150–51.

Indeed, without technical attribution, legal attribution becomes impossible—without technically attributing an act to an actor, it is impossible to declare that a cyber-activity reached the legal quantum of evidence required to achieve legal attribution.

B. The Current Attribution Standard in International Law under the Law of State Responsibility

State Responsibility under CIL and the Draft Articles generally requires that a State be responsible for its own bad behavior and the bad behavior of agents working on its behalf. This body of law pre-dates cyberspace, and it thus contemplates State behavior in an analog world rather than behavior in cyberspace. In other words, it accounts for acts and behavior that occur in conjunction with traditional Westphalian notions of sovereignty and territoriality.

Notwithstanding the shortcomings of the framework, the international legal community agrees that it applies to cyberspace and cyber-activities.[80] There are two somewhat overlapping approaches to the State Responsibility framework: (1) the "effective control" test, and (2) the "complete dependence" test. The "effective control" test established by the International Court of Justice ("ICJ") in the *Nicaragua Case*[81] is used to determine whether independent actors were sufficiently under the State's authority and control for their *acts* to be attributable to the State. For a State to "effectively control" a group, the State must "direct[] or enforce[] the perpetration of the acts,"[82] thus demonstrating a high level of State control. On the other hand, the even more stringent "complete dependence" test requires one to establish that a non-State actor is an "organ" of the State; in other words, that an agency relationship exists between that actor and the State. The burden of proof for both tests is "clear and convincing evidence."[83]

---

[80] *See, e.g.*, TALLINN MANUAL ON THE INTERNATIONAL LAW APPLICABLE TO CYBER WARFARE 29 (Michael Schmitt ed. 2013) [hereinafter TALLINN MANUAL].

[81] Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. Rep. 14, ¶ 190 (June 27) [hereinafter Nicaragua Case].

[82] *Id.* at ¶ 105.

[83] *See id.* at ¶¶ 386–94.

C. Technical Attribution Strategies and Difficulties

The challenges of attribution have attracted substantial attention from legal scholars.[84] The experts behind the second iteration of the Tallinn Manual[85] concluded that even if a state provides "the cyber tools, identif[ies] the targets, and select[s] the date for the cyber operation" it would still not necessarily rise to the requisite level of attribution to the State under the "effective control" standard.[86] The difficulty lies in the structure of the internet itself and in its function. As one scholar explained, "[t]he totality of the [i]nternet operates to deny positive technical attribution to the individual creating multiple barriers for positive technical attribution by computer scientists."[87]

The main issue in regard to technical attribution is the missing link between the computer itself and the identity of the human being acting behind that computer. This problem of technical attribution can be characterized as an *identification* issue: "there are no known means to date of positively identifying an author of an attack without having physical control over the computer system in which the code for the [malicious program] was written and then only if computer forensics can recover the data."[88] Locating, obtaining, and

---

[84] *See*, *e.g.*, Jolley, *supra* note 15, at 27 ("[P]roperly identifying the author of a cyber-attack is difficult, if not impossible: there are no means readily available to identify who authored an attack."); Christian Payne & Lorraine Finlay, *Addressing Obstacles to Cyber-Attribution: A Model Based on State Response to Cyber-Attack*, 49 GEO. WASH. INT'L L. REV. 535, 568 (2017) ("The novel characteristics of cyber-attacks make the existing standards of proof and degrees of control required to establish attribution extremely difficult to determine.").

[85] The Tallinn Manual is a guiding document based on "the views of a group of renown experts on the application of international law to cyber activities," intended to help the international legal community understand cyber issues. Eric Talbot Jensen, *The Tallinn Manual 2.0: Highlights and Insights*, 48 GEO. J. INT'L L. 735, 735 (2017).

[86] *Id.* at 752.

[87] Jolley, *supra* note 15, at 148. These barriers include "TOR" network and proxy servers, which obscure the identification of a computer by bouncing signals and packets around global networks.

[88] *Id.* at 171–72. Department of Justice attorney, Leonard Bailey, would disagree that the task is impossible. Rather, investigators (like DOJ or FBI investigators) have tools that can overcome many of the difficulties in attributing malicious cyber-activities to an individual or group. Interview with Leonard

examining a computer suspected of being a terminal for a domestic cybercriminal can already be a challenging task for the Federal Bureau of Investigation ("FBI") engaged in cyber investigations. However, because the FBI has a specialized cyber-toolkit, although it may still be difficult, they have capabilities for detecting, tracking, and ultimately finding the source of a cyber-crime in the domestic United States.[89] The Department of Justice outlines several broad investigative strategies for dealing with cyber-crime:

> The key methods and sources of evidence for disrupting cyber threats include: gathering materials during incident response; reviewing open source data; conducting online reconnaissance; searching records from online providers; undertaking undercover investigations; engaging in authorized electronic surveillance; tracing financial transactions; searching storage media; and applying a variety of special techniques.[90]

However, identification becomes a tricky exercise when the source of malicious cyber-activity originates from a foreign State. In these situations, the investigator's toolbox shrinks. The investigator must consider the effect of his/her action on complex issues of law, policy, and politics. Often, such decisions would require authorization from high-level officials that may be difficult to obtain because of the complexity of the issues involved, like sovereignty, territoriality, and foreign relations. "[I]nvestigators also must work cooperatively with foreign partners to access evidence and disrupt transnational cyber threats."[91] The task of gathering evidence abroad increases in complexity and difficulty when investigators need access to information or evidence in a State with which the United States has rocky or hostile relationships, like China, Iran, Russia, or North Korea.

There are three categories of technical attribution: indirect, forensic, and, to repeat the more general principle's name, technical.

Bailey, Special Counsel for National Security, Dep't of Justice, Criminal Division Computer Crime and Intellectual Property Section (Mar. 26, 2019).

[89] *See, e.g.*, CYBER-DIGITAL TASK FORCE, OFFICE OF THE DEPUTY ATTORNEY GENERAL, U.S. DEP'T OF JUSTICE, REPORT OF THE ATTORNEY GENERAL'S CYBER-DIGITAL TASK FORCE 49–82 (July 2, 2018), https://www.justice.gov/ag/page/file/1076696/download [https://perma.cc/JM44-XY2Q] [hereinafter DOJ Cyber Report].

[90] *Id*. at 49.

[91] *Id.*

"Indirect attribution" typically uses manifold techniques, including traditional intelligence gathering tools; law enforcement strategies; computer forensics and programs; and potential motives to circumstantially link an actor to a particular attack. Indeed, a determination of motive cannot be understated. And, at the very least, using process of elimination could reduce the number of suspects to consider; for instance, why would an allied State actor commit a ransomware[92] attack against the United States?[93] "Forensic attribution" uses malware evaluation, computer forensics, and code analysis to determine who authored the cyber-attack, while the specific category of technical attribution uses various computer science techniques to trace the signal to its origin.[94]

One of the biggest difficulties lies in the speed at which malicious cyber actors incorporate and use new technology. Every advancement in technical attribution computer science is learned and employed quickly by those malicious cyber actors. Similarly, a malicious cyber actor need not generate original code to deploy malware. Instead, these actors often reuse malware programs; thus even if the malware is traced to its original author, it may not elucidate who actually used it in a specific attack.[95] Likewise, the increased use of Tor, IP masking through spoofing and/or through proxy servers, as well as false attribution trails make the trail of breadcrumbs for forensic scientists long, slow, and difficult to follow.[96]

---

[92] "Ransomware" is a malicious program activated when an unsuspecting person clicks a hyperlink in an email, for instance. When the ransomware program runs, it spreads through the victim network and locks the network completely, threatening to erase the network's data or indefinitely deny access unless the victim pays a ransom, usually in cryptocurrency. Once the ransom is paid, the network is usually released.

[93] Interview with Leonard Bailey, *supra* note 88.

[94] Jolley, *supra* note 15, at 141. These technical attribution techniques include recursive tracebacks, stepping stones, honey pots, authorship attribution, attribution of files, manual attribution, *inter alia*.

[95] *Id.*

[96] For a discussion of how Tor networks work, see PANAYOTIS A. YANNAKOGEORGOS, STRATEGIES FOR RESOLVING THE CYBER ATTRIBUTION CHALLENGE 14–18 (May 2016), https://www.airuniversity.af.edu/Portals/10/

What if the cyber-threat arose from a CyDAE, not controlled by a human handler, orchestrating cyber-activities on its own prerogative? The rules of legal attribution are difficult enough for attributing acts to human actors and States without a sophisticated CyDAE intermediary further obscuring attributive links. The next section addresses this pressing issue.

V. THE THREAT OF A CYDAE

Imagine a decentralized intelligent entity, existing on thousands of computers, disembodied and intangible, residing only in cyberspace. Both its origin and its existence are unknown. Perhaps it is a web or network of "swarm intelligences," different, smaller AI organs that play off each other, contributing to the greater CyDAE,[97] exemplifying the adage "the whole is greater than the sum of its parts." What is known is that it was created by humans, but it is unclear whether human involvement continues to contribute to its development. Indeed, "[b]y definition, the initiator of a[] [CyDAE] would neither own the entity nor control it after launch. The initiator would, however, have the opportunity to set the algorithm's objectives prior to launch."[98] Still more worrisome, unlike the ArtDAO with a purpose to create art, or autonomous self-owning cars with a purpose to offer rides to humans, the CyDAE's apparent purpose may not always be benevolent and could easily undertake a purpose to orchestrate cyber-attacks against the United States and its allies.

The CyDAE with its veritable panoply of malicious capabilities[99] coupled with inhumanly fast operating speed, begins

---

AUPress/Papers/cpp_0001_yannakogeorgos_cyber_Attribution_challenge.PDF [https://perma.cc/U2G7-U6U7].

[97] One of the foremost thinkers in the decentralized AI space, Ben Goertzel, predicts the AGI arising from a network of smaller AI that share information and grow together. *See, e.g.*, SINGULARITYNET, WHITEPAPER 2.0 (Feb. 2019), https://public.singularitynet.io/whitepaper.pdf [https://perma.cc/6RZQ-AAVB] [hereinafter SingularityNET Whitepaper].

[98] LoPucki, *supra* note 55, at 900.

[99] CyDAE-like entities "are capable of inflicting massive damage on social and economic systems. They could shut down human computing, steal and release confidential information, and wreak havoc by seizing control of the internet of things." *Id*. at 902.

its cyber-operations against both State and private actors in the United States by surreptitiously monitoring private networks; parsing and changing vast datasets after copying them; or infecting thousands of Internet of Things ("IoT") devices[100] with "smart-botnet"[101] malware-like programs.[102] The following section illustrates the characteristics of a CyDAE—based on existing AI characteristics and qualities—that would support this CyDAE's malicious cyber-activity.[103]

A. Uniquely Effective Characteristics and Cyber-Capabilities of a CyDAE

In cybersecurity, speed is everything. The term "breakout time" is a metric for sophistication of a malicious cyber-activity; it "measures the speed with which adversaries accomplish lateral movements in the victim [network] environment after their initial [access]."[104] This is important "because it represents the time limit for defenders to respond and contain or remediate an intrusion before it spreads widely in their [network] environment and leads to a major breach."[105] Cybersecurity firm CrowdStrike measured the fastest State-attributed cyber-actors (Russia) to have an average breakout time of eighteen minutes, meaning it took those actors an

---

[100] "IoT" refers to the networks created by everyday devices with built-in computers, which range from refrigerators to doorbells to thermostats to children's stuffed animals.

[101] In simplistic terms, a "botnet" is a network of devices hijacked and controlled by a hacker through malware. The hacker can harness the computing power of such a network for nefarious purposes like denial of service attacks (flooding victim networks with signals causing the system to slow down or crash).

[102] Maybe to achieve decentralization, the CyDAE's propagates its code through IoT devices much in the same way a botnet operates.

[103] This analysis does not consider destructive attacks on critical infrastructure, although that kind of attack would prove to have kinetic effects, elevating a CyDAE's malicious cyber-activities into the realm of *jus ad bellum* and *jus in bello*. Still, a CyDAE coordinating digital attacks in warfare terms is an interesting proposition.

[104] CROWDSTRIKE, 2019 GLOBAL THREAT REPORT, ADVERSARY TRADECRAFT AND THE IMPORTANCE OF SPEED 14 (2019), https://www.crowdstrike.com/resources/reports/2019-crowdstrike-global-threat-report/ [https://perma.cc/6ZJH-T6LK] [hereinafter CrowdStrike Report].

[105] *Id.*

average of eighteen minutes from when they accessed the victim network to begin causing harm.[106] The CyDAE, which can process information at beyond-human speeds, could hypothetically achieve a breakout time of seconds or less. Such unfathomable speed already dominates securities markets. Many Wall Street firms practice "high-frequency trading," a technique in which they use algorithms so fast that they can "front-run" purchases of stock. In short, the algorithm detects the signal of another entity making a large buy order for stock before the buy order signal reaches the exchange, races ahead, and purchases all the shares before the first buy order goes through, then immediately sell the shares at a higher price to the entity that made the initial first buy order.[107] The same kind of "front-running" technique could be used by a CyDAE to scan networks, find defense mechanisms, and modify or adapt its behavior in accordance with the type of defense mechanism. Alternatively, the CyDAE could hypothetically "front-run" decoy signals indicative of normal activity around the network to lull defenders into a false sense of security.

According to CrowdStrike, "[a]fter attackers obtain their initial foothold, their first order of business is to get oriented within their newly accessed environment before determining next steps toward their objective."[108] This deliberative period, for major attacks, occurs because the human hacker(s) strategize and plan the next steps of the attack, based on the network landscape. It is not farfetched to think that a CyDAE, with speed undetectable by humans, could deliberate for fractions of seconds and unfurl its digital tendrils through the network at the speed of a lightning strike. Even if it were detected by network defenses, by the time that the signal indicating a breach reaches a human overseer (assuming the CyDAE cannot reroute the signal), the CyDAE could have moved, spread, attacked, changed its strategy, or deployed decoys many times over. The high-frequency trading algorithms mentioned above can execute

---

[106] *Id.* at 14–15.

[107] Elvis Picardo, *Understanding High-Frequency Trading Terminology*, INVESTOPEDIA (last updated May 30, 2019), https://www.investopedia.com/articles/active-trading/042414/youd-better-know-your-highfrequency-trading-terminology.asp [https://perma.cc/U7TD-BU7Z].

[108] CrowdStrike Report, *supra* note 104, at 21.

thousands of trades *per second*, a speed considered slow in the industry, which measures performance as number of trades in microseconds, or *millionths* of a second.[109] A logical inference then is that if an algorithm can execute one trade per microsecond, it could conceivably execute one million trades per second.[110] Using the metrics of Wall Street for the CyDAE hypothetical, a CyDAE could cause substantial havoc on a network in the same eighteen minute time frame (a time frame equal to one billion and eighty million microseconds) it took the fastest human cyber-operators in the world in 2018 to get started on their attack once they infiltrated the system. By the time the world's best human cyber-defense operator finds and identifies a cyber intrusion from a CyDAE, picks up the phone and calls leadership, the CyDAE could have already stolen or copied the company's information, transferred funds, and/or modified important data, before making itself undetectable again.

B. Malware Infiltration

The first step would be for the CyDAE to infiltrate its target system—whether the system be a government network or private network. The CyDAE's machine learning ANN would attempt to penetrate the system with a phishing attack to gain access to the network. For example, the CyDAE could "scrape" LinkedIn for employees with access to the network, mimic one of those employees' writing styles, and draft a decoy email to another employee the CyDAE determines is a "friend" of the first employee. This decoy email would act as a vehicle for the deployment of malware on the victim network.

Many malicious cyber-activities come from "malware," an umbrella term for malicious computer code with many

---

[109] *Getting Up to Speed on High-Frequency Trading*, FINRA (Nov. 25, 2015), http://www.finra.org/investors/getting-speed-high-frequency-trading [https://perma.cc/L42C-QWJ3].

[110] For comparison's sake, a fast-human typist can type 70–100 words per minute. Kimberlee Leonard, *What is a Good Typing Speed Per Minute?*" CHRON (Jan. 29, 2019), https://smallbusiness.chron.com/good-typing-speed-per-minute-71789.html [https://perma.cc/69MU-HHP2].

expressions.[111] Malign actors often use malware as a method to infiltrate a system initially—tricking a user into executing the code, a strategy often referred to as "phishing."[112] Once executed the malware can execute a ransomware[113] program; send "bots"[114] across the network to hijack computers to use in distributed denial-of-service attacks ("DDoS") against other, outside networks[115]; or install keyloggers that record every keystroke on a particular computer (including passwords).[116] Malware, however, has detectable signatures, and in some cases, sends traceable "command and control" signals to its operator requesting instructions.[117] In that sense, malware is "noisy"; cyber-defense infrastructure becomes suspicious when an unknown program runs in its environment.[118]

But, a smarter malware CyDAE could adapt itself to be less noisy, release CyDAE "smartbots" that mutate and either keep in constant communication with the CyDAE, act completely autonomously, or even sit dormant in a system until it detects a trigger to act. If a smartbot program includes some kind of evolving

---

[111] DOJ Cyber Report, *supra* note 89, at 25–28.

[112] *See* Jennifer Lynch, *Identity Theft in Cyberspace: Crime Control Methods and their Effectiveness in Combating Phishing Attacks*, 20 BERKELEY TECH. L.J. 259, 259–261 (2005) (defining and describing "phishing").

[113] A "Ransomware" attack "blocks a victim's access to data on its systems, typically by encrypting the data and demanding that the victim pay a ransom, often in the form of a difficult-to-trace virtual currency, to restore the data." DOJ Cyber Report, *supra* note 89, at 24.

[114] The term "bot" refers to a program that executes commands for a simple purpose, but sometimes that program is malware. In the cyber world, malign actors send malware to tens or hundreds of IoT devices, and if executed, the malware turns the IoT devices into bots that the actor can use to coordinate cyber-attacks.

[115] DDoS "involves the orchestrated transmission of communications engineered to overwhelm the victim's network's connection to the internet in order to disrupt that network's ability to send or receive communications." DOJ Cyber Report, *supra* note 89, at 23.

[116] *See* Paul Koob, *Not Enough Fingers in the Dam: A Call for Federal Regulation of Keyloggers*, 28 TEMP. J. SCI. TECH. & ENVTL. L. 125, 126–27 (2009) (defining and describing "keyloggers").

[117] Interview with Leonard Bailey, *supra* note 88.

[118] *Cyber Adversary Olympics: It's Russia for the Gold and North Korea (!) for the Silver*, CYBERLAW PODCAST (Feb. 25, 2019, minute 45:00), https://www.steptoe.com/feed-Cyberlaw.rss [https://perma.cc/H276-29LY].

machine learning algorithm, it would not need to send the easily detectable command and control signals to the controlling entity (whether human or CyDAE), but could instead know the circumstances or triggers in the network that signal when they should mount a DDoS attack or execute their malicious program.[119]

Theoretically CyDAE smartbots could stifle attempts to discover it by changing their code as a reaction to their environment analogous to the high-frequency algorithm anticipating other entities' stock orders described above. That is, a bot-detecting software sends a signal to a human controller alerting him/her that the computer is infected by a "bot," identified by a signature, Code A. In the time the signal was sent to the human controller, the CyDAE smartbot changes its signature from Code A to Code G. The bot-detecting software then sends a signal to the human controller identifying Code G, but the CyDAE smartbot then changes from Code G to Code W, perpetuating the cycle. The human controller starts receiving the signals from the bot-detection software, but cannot keep up with the changing smartbot code, and thus cannot find it. Indeed, a CyDAE smartbot could presumably operate without sending any command or control signals back to the CyDAE itself, thus insulating the CyDAE from easy detection. With the placement of its CyDAE smartbots, the CyDAE's DDoS infrastructure is set, and it can start harnessing the vast computing power of its smartbots.

C. Data-Based Attacks

For years, the dark web and sites like WikiLeaks existed as repositories for leaked, sensitive, classified, or stolen information and data. Malicious cyber-activities aimed at stealing data for monetary gain or exposing protected information has become an increasingly popular form of attack. For example, self-proclaimed "hacktivists" infiltrated Ashley Madison, the dating website aimed at people seeking extramarital affairs, and they released user data from the site, including "users' real names, banking data, credit card

[119] DOJ Cyber Report, *supra* note 89, at 23.

transactions, [and] secret sexual fantasies."[120] Indeed, the United States government arrested the Chinese national accused of the notorious hack of United States Government Office of Personnel Management ("OPM") which resulted in the release of millions of federal employees' personal information being released on the dark web.[121] This sort of harassment, which includes "broadcasting personal information about the victim on the Internet, exposing him or her to . . . harassment by others," is known as "doxing."[122]

Once the CyDAE gains access to the organization's network, it spreads, and rapidly scans for easily identifiable personal information, targeting high ranking individuals in the organization. With incredible speed, it sends the data back to itself and into a decentralized repository akin to a blockchain. The CyDAE may have its own site on the Internet to release the information publicly, or could send the information to another website like WikiLeaks.

Alternatively, the CyDAE could speed through the network, systematically deleting the organization's data, in a similar fashion to the Sony hack,[123] maneuvering to avoid detection by the cyber-defenses of the victim organizations. Or, the CyDAE could *modify* the data as it shoots though the network, in an attack called "data poisoning," which targets the learning process of AI systems. As discussed above, AI systems, particularly machine learning and ANN systems, require vast amounts of labeled and categorized data to learn. Data poisoning attacks occur when "malicious users inject false training data with the aim of corrupting the learned model."[124]

---

[120] *A Dating Site and Corporate Cyber–Security Lessons to Be Learned*, PANDA MEDIACENTER (Oct. 6, 2018), https://www.pandasecurity.com/mediacenter/security/lessons-ashley-madison-data-breach/ [https://perma.cc/GHL7-X8HC].

[121] Complaint, United States v. Pingan, No. 17MJ2970 (S.D. Cal. Aug. 21, 2017).

[122] DOJ Cyber Report, *supra* note 89, at 33.

[123] Raphael Satter, *North Korean Programmer Charged in Sony Hack, WannaCry Attack*, PBS NEWS HOUR (Sept. 6, 2018), https://www.pbs.org/newshour/nation/north-korean-programmer-charged-in-sony-hack-wannacry-attack [https://perma.cc/JQ2S-JX9H].

[124] JACOB STEINHARDT, PANG WEI KOH, & PERCY LIANG, 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, CERTIFIED DEFENSES FOR

In an age when States strive to achieve AI superiority, ruining an AI's ability to learn would be an effective tactic to disrupt the development of an AI system, especially when many ANNs and other machine learning algorithms are unexplainable black boxes, preventing developers from seeing the process from which the AI system derived its output or conclusion. A CyDAE could scan vast quantities of data, relabeling them (from "cat" to "dog," for example), or replacing legitimate data with fake or inaccurate data points.

Beyond seizure, manipulation, or theft of information, malicious cyber activities can result in physical consequences. The Stuxnet malware was built to infiltrate industrial systems software, which allow computers to control physical industrial processes like opening and closing valves.[125] Stuxnet caused centrifuges that processed material for the creation of nuclear energy at the Natanz Iranian nuclear facility to malfunction by spinning extraordinarily fast, while simultaneously sending signals to the monitoring computer systems that showed the centrifuges were working properly.[126] Alternatively, some malware can render entire computer systems completely useless. If a CyDAE were capable of infiltrating industrial control systems in a similar fashion to the Stuxnet virus, it could result in repercussions beyond data destruction. Opening valves in chemical plants could result in noxious chemical leaks or spills. Hydroelectric dams that control the flow of thousands of tons of water could malfunction and result in flooding. Manipulating the software that controls the timing of traffic lights could result in extreme obstructions in vehicular traffic, or even worse, traffic collisions. The cyberthreat to such critical infrastructure is real, and an entity like a CyDAE could cause substantial physical damage and

---

DATA POISONING ATTACKS (2017), https://papers.nips.cc/paper/6943-certified-defenses-for-data-poisoning-attacks.pdf [https://perma.cc/ZWL6-2UUT].

[125] Stuxnet is probably the closest malware to a CyDAE. It was capable of replicating itself, sending false signals to hide its presence from outside detection. Bruce Schneier, *The Story Behind the Stuxnet Virus*, FORBES (Oct. 7, 2010) https://www.forbes.com/2010/10/06/iran-nuclear-computer-technology-security-stuxnet-worm.html#6ebd5d5c51e8 [https://perma.cc/Q5X9-P8AW].

[126] *See id.*

casualties if it interrupted or interfered with the function of such industrial control systems.

VI. THE CYDAE'S ATTRIBUTION SHIELD

If it were to exist, a CyDAE's beyond-human capabilities would present difficulty to even the most sophisticated human cybersecurity experts. Whether or not a CyDAE could be detected and identified are two major questions, but even if they are detectable and identifiable, to what degree could the CyDAE's activity be attributed to a human actor in the current legal landscape? This section attempts to examine the relevant legal issues.

The general characteristics of a CyDAE present inherent problems to attribution. International law's two current tests were developed in the analog world, where physical beings and things could cross territorial boundaries or kill human beings with guns, knives, and bombs.

These legal tests have high burdens. The act-centric "effective control test" will permit legal attribution of an independent actor's acts to a State if clear and convincing evidence demonstrates that the actor was sufficiently under the State's control, because the State "directed and enforced the perpetration of the acts."[127] This test would always fail to permit attribution if the factual scenario involved a CyDAE. Even if a State released a CyDAE into cyberspace, the CyDAE is not a group, or a person, or any kind of legal entity. Neither international law nor any State in particular proscribes a framework for an autonomous legal AI entity. Moreover, traditional technical attribution techniques already experience difficulty tracing a cyber-event to the computer or operator. Even if a forensic whiz traced the cyber-activity to a CyDAE, the analysis would stop there, as the operator would likely have no way of piercing the CyDAE's unexplainable system to gather clues about its programming. Assuming *arguendo*, that a computer forensics expert could infiltrate the black box system of a CyDAE, the scientist could potentially gather clues about its origin and motives, but considering it acts autonomously, with no "command and control" architecture, there would be no

[127] *See* Nicaragua Case*, supra* note 81.

breadcrumbs to follow from the CyDAE to its human creator(s). Even if forensic technicians could trace a CyDAE's code to one node in its decentralized network, the destruction or examination of that node would reveal little, considering the nature of decentralization.

The actor-centric "complete dependence" test requires proving, with clear and convincing evidence, that the actor is an "organ" of the State and that an agency relationship exists between the two, and is even more difficult to prove than the "effective control test" above.[128] Showing "complete dependence" of a CyDAE on a State would be virtually impossible—demonstrating an agency relationship between a CyDAE acting autonomously and a State would require specific forensic evidence. The same issues arising under the effective control test also arise under the "complete dependence test." The unexplainability of a CyDAE would obfuscate any connection to the State, and the fact that the CyDAE needs no instruction or commands to pursue its programmed goals further distances the CyDAE from any States. Furthermore, the CyDAE cannot be considered an actor because it lacks legal personhood or entityhood.

The difficulties in attributing malicious cyber-activities to a State multiply when applied to CyDAE hypotheticals. Policymakers must begin to address such issues so regulatory structures can be built before CyDAEs emerge on the world stage. The next section contains a series of proposals intended to challenge and encourage policy and lawmakers to consider the importance of anticipating the existence of near-future, multiform CyDAEs.

VII. PROPOSALS

The idea of preemptively regulating AI is not new. Elon Musk, for instance, declared: "I'm increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish."[129] Many thinkers in the AI field endorse and embrace

---

[128] *Id*.

[129] Elon Musk, *We Are "Summoning a Demon" with Artificial Intelligence*, UPI (Oct. 27, 2014), http://www.upi.com/Business_News/2014/10/27/Elon-Musk-

the responsible development of AI systems.[130] This analysis suggests four general, but synergistic proposals with the goal of presenting a holistic approach to regulating future CyDAEs, and AI systems broadly.[131]

Proposal 1: AI Registry

All AI systems should be required to be registered with a legislatively or executively mandated agency or commission. The registration would be similar to the way that money services businesses must register with the Financial Crimes Enforcement Network in the United States. The registration would require basic information about the creator(s) and/or owner, whether it be an individual, corporation, or other legal arrangement. Such an idea is not novel: the first decentralized AI marketplace requires registration for AI to participate in its platform.[132] The registry would make basic information publicly available and accessible, as well as promote transparent use of CyDAEs and AI systems generally. Notably, a CyDAE released for the specific purpose of orchestrating cyber-attacks would likely not be registered by the entity that created it, so there must be some paradigm to deal with unregistered and malicious CyDAEs.

Proposal 2: Explainable AI Systems

AI systems should be primarily explainable. Some see benefits to unexplainable AI systems, or at least, they view unexplainable systems as the result of AI algorithms that process information at a level so advanced, that explaining its processes would be ineffective.[133] Perhaps every step of an AI system's process need not be explainable so to not impede "the cases that give machine

---

We-are-summoning-a-demon-with-artificial-intelligence/4191414407652 [https://perma.cc/WEY5-8EFJ].

[130] *See* SINGULARITYNET WHITEPAPER, *supra* note 97, at 6, 54.

[131] Importantly, this analysis focused on CyDAEs as a cybersecurity and national security threat. However, CyDAEs may arise in other forms, and so many of these proposals overlap with benevolent CyDAEs as well.

[132] SINGULARITYNET WHITEPAPER, *supra* note 97, at 17.

[133] *See, e.g.*, Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018).

learning its greatest value: true patterns that exceed human imagination."[134] At the very least, "[w]e need the developers to show their work."[135] In other words, policymakers should implement an explainability standard that requires, at a minimum, that the programmer or an expert reviewing the programmer's notes be able to explain the rules built into its AI system. Of course, that assumption requires that the programmer or creator be known, which underscores the importance of a registry.

Proposal 3: Legal Personhood Structure with Human Control

The most important feature of an AI regulatory regime is human control; thus, any AI system, and especially autonomous AI systems, should have human-controlled fail-safe mechanisms built in so no "loss of general control" occurs.[136] Regardless of the form, autonomous AI entities need a personhood arrangement so the law can handle what an AI entity *is*. Perhaps the personhood arrangement should be an autonomous entity, allowing an AI to govern itself. Should that be the case, as a part of that AI entity's incorporation or legal personhood creation, a human or corporate overseer entity should be tethered to that AI, with some affirmation and/or description of how that human-controlled legal entity maintains oversight, no matter how attenuated that oversight may be.[137]

Proposal 4: Universal Jurisdiction for CyDAEs

What happens under the framework above if a State discovers an unregistered, unclaimed, unbridled CyDAE operating in cyberspace, and attribution cannot be made? CyDAEs should be subject to something close to universal jurisdiction, considering the threats and difficulties—both practical difficulties and legal challenges—presented in this article. Cyberspace is everywhere and

---

[134] *Id*. at 1089.

[135] *Id.* at 1138.

[136] *See* Scherer, *supra* note 35, at 366 (discussing "control").

[137] A kill-switch or other mechanism controlled by a human that can deactivate the AI entity. Additionally, autonomous entity ownership of other AI entities should be prohibited (no AI "shells" should be allowed). The discussion of what mechanisms of control should be required is a different and expansive discussion.

nowhere, and it confounds traditional notions of sovereignty, jurisdiction, and territoriality. Decentralization further complicates the issue because it allows an entity or program to exist universally with only being partially located on a physical server. Moreover, data and computing can shift around the globe instantly, so what may be under one State's jurisdiction in one second may be in another State's jurisdiction the next second, and yet another State's jurisdiction the third second.

Jurisdiction over such evanescent and transient entities requires multilateral agreements on jurisdiction. The likelihood of such a multilateral agreement seems slim, considering current global multilateral understandings of sub-armed conflict and malicious cyber-activities are uncertain, and the global community cannot achieve consensus on even those fundamental rules.[138]

Universal jurisdiction is the doctrine that permits "any nation [to] prosecute universal offenses, even over the objection of the defendants' and victims' home states."[139] Historically, universal jurisdiction applied primarily to pirates because "traditional jurisdictional categories did not cover piracy."[140] This proposal seeks universal jurisdiction—not necessarily to prosecute—but to "summarily execute" a CyDAE because of its existence in cyberspace, which is "a global commons" that "lay[s] outside the territorial jurisdiction of any nation."[141] Indeed, a CyDAE cannot be captured and brought to court in a State jurisdiction. Nor, without attribution, can the creator be held responsible for the acts of the CyDAE. Therefore, this proposal suggests that an unregistered, unattributable CyDAE should be subject to summary destruction by any State. This universal jurisdiction would not permit one State to destroy the physical computer systems located in another State, rather, it would permit a State to use cyber-capabilities to disrupt,

---

[138] *See* TALLINN MANUAL, *supra* note 80, at 45–53.

[139] Eugene Kontorovich, *The Piracy Analogy: Modern Universal Jurisdiction's Hollow Foundation*, 45 HARV. INT'L L.J. 183, 183 (2004).

[140] *Id*. at 190.

[141] *Id.* While the analogy between cyberspace and the high seas fails in many ways, the evolution of the decentralization of cyberspace creates a problem of jurisdiction.

confound, destroy, undo, erase, corrupt, or otherwise destroy the program, code, or signals that make up the CyDAE.

VIII. CONCLUSION

The ways in which AI will shape and transform society are difficult to predict, considering the inscrutable nature of AI systems. AI systems may exist on a different level, outside the bounds of human-imposed strictures. As this new and exciting technology develops, the world must, at the very least, provide legal and policy guardrails for AI technology to prevent it from careening out of human comprehension and control. This analysis used the hypothetical CyDAE to exemplify potential worst-case scenarios in AI evolution. Although hypothetical, the CyDAE idea presented here is an amalgamation of existing technologies, AI models, and realistic theories. The four proposals proffered above aim to guide humanity into a symbiotic relationship with AI. Setting parameters for AI system development now will allow a future in which society is not threatened by CyDAEs but improved by their existence. Indeed, monitoring and regulating AI in the physical world seems intuitive—robots have physical effects and liability will attach to *someone* for the results of robotic acts. The esoteric and tangible effects resulting from action in the world of cyberspace however, are much more difficult to comprehend and fit into existing legal frameworks. In *Neuromancer*, William Gibson described cyberspace as a "consensual hallucination experienced daily by billions of legitimate operators, in every nation . . . ."[142] Much like a hallucination or dream, events in cyberspace are difficult to comprehend and harder still to predict. In Gibson's cyberspace, several AGI-level CyDAE's sought to achieve goals not discernable to humans; killing, stealing, and maiming in the process. It is the illegitimate operators in cyberspace, like the incomprehensible and complex CyDAEs dreamt up by Gibson—beyond human understanding, beyond the reach of human senses, and with opaque motivations—that will challenge humanity the most in the future to come.

[142] GIBSON, *supra* note 3, at 56.