Journal of Legislation

Volume 45 | Issue 2 Article 1

6-7-2019

The Perils and Promises of Artificial General Intelligence

Brian S. Haney

THE PERILS AND PROMISES OF ARTIFICIAL GENERAL INTELLIGENCE

Brian S. Haney[†]

INTRODUCTION

Most people think the fusion of man and technology might happen in the distant future; the truth is that human beings are already cyborgs. With a smartphone, a human being can quickly answer virtually any question, store limitless information in memory, and complete any calculation.[1] Modern technology companies collect data about humans from smartphones and feed it directly through advanced artificial intelligence ("AI") systems.[2] By design, AI systems maximize electrical impulses to consumers' limbic systems, the brain's reward center, to stimulate economic growth and development.[3] At the National Governors Association's 2017 Summer Meeting, Elon Musk stated, "[t]he biggest risk that we face as a civilization is artificial intelligence."[4]

Musk is not alone; in fact, there is a growing list of scholars and industry leaders that have directed attention to the existential threats that AI poses to man.[5] Yet, legal scholarship on the topic of artificial intelligence has either denied or relatively ignored the accelerating rate of AI advancement.[6] Instead, current legal scholarship devoted to AI regulation has encouraged regulators not to be distracted by claims of an "AI apocalypse" and to focus their efforts on "more immediate harms."[7] In sum, legal scholarship in the field of AI regulation is far behind and provides misguided advice to regulators and scholars.[8] Indeed, every task humans use intelligence to perform is a target for AI automation.[9] Further, it has often been the case that once an AI system reaches human level performance at a given task, shortly thereafter that same AI system exceeds the performance of the most skilled

[†] Brian S. Haney, J.D. Notre Dame Law School 2018, is the Chief Executive Officer of Martian Technologies. Special thanks go to LaDarien Harris, Mike Gallagher, Ryan Claudeanos, Delaney Foreman, Maria Munoz-Robles, Brian Wongchaowart, Bill Green, Ned Rooney, and the Journal of Legislation.

[1] *The Joe Rogan Experience #1169 – Elon Musk*, THE JOE ROGAN EXPERIENCE (Sept. 6, 2018), https://www.youtube.com/watch?v=ycPr5-27vSI.

[2] *Id*.

[3] *Id*.

[4] *Elon Musk at the National Governors Association 2017 Summer Meeting*, C-SPAN, 50:00 (July 15, 2017), https://www.c-span.org/video/?c4676772/elon-musk-national-governors-association-2017-summer-meeting.

[5] *See* MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION 12 (2018); *see also* MAX TEGMARK, LIFE 3.0 BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE (2017).

[6] *See* Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 432 (2017).

[7] *See id*. at 431.

[8] *See id*.

[9] BRUNDAGE ET AL., *supra* note 5.

humans in completing that task.[10] Many AI researchers expect that AI systems will eventually reach and then exceed human-level performance in all tasks.[11]

AI technology is sculpting a future where fake pictures and videos are inexpensive, widely available, and indistinguishable from the real thing, which is completely reshaping the way in which humans associate truth with evidence.[12] Even those who doubt whether Artificial General Intelligence ("AGI"), AI capable of accomplishing any goal,[13] will be created in the future, still agree that AI will have profound implications for all domains, including: healthcare, law, and national security.[14] The purpose of this Article is twofold. First, this Article defines and explains AI's cutting-edge technology with a specific focus on deep reinforcement learning, a breakthrough type of machine learning developed by Google in 2013.[15] Second, this Article identifies three hurdles for regulators to overcome in regulating AI.

This Article contributes to current legal and AI scholarship in three main ways. It is the first to focus on deep reinforcement learning, specifically on the existential threats posed by AI and it is the first to engage with the formal models that underpin AI. This Article proceeds in three parts. Part I explains basic terms and concepts in AI and explores several practical applications of AI in modern industry. Part II explains deep reinforcement learning, a relatively recent breakthrough in AI that many scholars believe provides a path to AGI. Part III explores legal scholarship on the topic of AI regulation and discusses three issues regulators must address to develop a strong regulatory framework for AI.

I. ARTIFICIAL INTELLIGENCE

Contemporary scholars have presented several different definitions of AI. For example, MIT Professor Max Tegmark concisely defines AI as "non-biological intelligence."[16] Google's Ray Kurzweil has described AI as "the art of creating machines that perform functions that require intelligence when performed by people."[17] Additionally, according to Stanford Professor Nils Nilsson, AI is "concerned with intelligent behavior in artifacts."[18] Generally, and for the purposes of this Article, AI refers to the study and development of intelligent machines that can replicate the thought processes of human cognitive functions like making predictions, speech processes, or playing games.

While AI includes different categories, two types of AI are most important in the context of AI regulation. The first is narrow AI, also known as weak AI.[19] Narrow AI has the ability to accomplish a limited set of goals[20] and is associated with attempts to develop AI to improve human intelligence, as opposed to duplicating

[10] *Id*. at 16.

[11] *Id*.

[12] GREG ALLEN & TANIEL CHAN, ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY 31 (2017).

[13] TEGMARK, *supra* note 5, at 31 (2017).

[14] ALLEN & CHAN, *supra* note 12.

[15] Methods and Apparatus for Reinforcement Learning, U.S. Patent Application No. 14/097,862 (filed Dec. 5, 2013) (available at https://patents.google.com/patent/US20150100530A1/en); *see also* TEGMARK, *supra* note 5, at 84.

[16] TEGMARK, *supra* note 5.

[17] RAY KURZWEIL, THE AGE OF INTELLIGENT MACHINES 14 (1992).

[18] NILS J. NILSSON, ARTIFICIAL INTELLIGENCE: A NEW SYNTHESIS 1 (1998).

[19] NILS J. NILSSON, THE QUEST FOR ARTIFICIAL INTELLIGENCE 388 (2010).

[20] *See* TEGMARK, *supra* note 5.

human intelligence.[21] The second type of AI is artificial general intelligence ("AGI"), also known as strong AI.[22] To demonstrate AGI, an AI agent must have the ability to accomplish any goal.[23] AGI is associated with the claim that a programmed computer could be a mind and could think at least as well as humans do.[24] Ultimately, AGI is the current goal for many AI researchers.[25] For example, OpenAI, a non-profit organization funding pioneering research in the field, states on its website that its mission is "[d]iscovering and enacting the path to safe artificial general intelligence."[26] Yet, it appears for the time being, that only narrow AI has been developed and successfully deployed.[27]

A. AI in Modern Professional Industries

The implementation of narrow AI is disrupting modern industries worldwide.[28] Even the legal industry is not exempt from this corrosive force.[29] Indeed, technology assisted review ("TAR") is revolutionizing the discovery process and AI is at the forefront of this innovation.[30] Litigators are now commonly called on by clients to establish e-discovery relevancy hypotheses and to implement predictive coding models (a type of TAR) for the discovery of electronic information.[31] In this process, litigators will first identify keywords to search and identify an initial set of documents to be reviewed.[32] Then, document review attorneys review, code, and score the initial set of documents based on the occurrence of certain keywords in relation to a document's relevance.[33] As this review takes place, e-discovery attorneys train and model supervised learning algorithms to classify documents based upon the document review attorneys' decisions in classifying documents in the initial set of documents.[34] In other words, the algorithm learns what documents are relevant by analyzing and replicating the decisions of real attorneys.[35] Additionally, predictive-coding models are capable of classifying millions of discoverable documents based on relevance.[36]

[21] *See* NILSSON, *supra* note 19, at 388–89.

[22] *See* NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 23 (2017).

[23] TEGMARK, *supra* note 5.

[24] NILSSON, *supra* note 19.

[25] *Id*.

[26] *About OpenAI*, OPENAI https://openai.com/about/ (last visited May 10, 2019).

[27] *See generally* Nick Bostrom, *Are You Living in A Computer Simulation?*, 53 PHILOSOPHY Q. 211, 243 (2003).

[28] ALLEN & CHAN, *supra* note 12; *see also* HEMANT TANEJA, UNSCALED: HOW AI AND NEW GENERATION OF UPSTARTS ARE CREATING THE ECONOMY OF THE FUTURE 1 (2018).

[29] RICHARD SUSSKIND, TOMORROW'S LAWYERS 11 (2d ed. 2017).

[30] Scott D. Cessar, Christopher R. Opalinski, & Brian E. Calla, *Controlling Electronic Discovery Costs: Cutting "Big Data" Down to Size*, ECKERT SEAMANS (Mar. 5, 2013), https://www.eckertseamans.com/publications/controlling-electronic-discovery-costs-cutting-big-data-down-to-size; *see also* Nicholas Barry, *Man Versus Machine Review: The Showdown Between Hordes of Discovery Lawyers and a Computer-Utilizing Predictive-Coding Technology*, 15 VAND. J. ENT. & TECH. L. 343, 344 (2013).

[31] KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS 240–42 (2017).

[32] Barry, *supra* note 30, at 351.

[33] GORDON V. CORMACK & MAURA R. GROSSMAN, EVALUATION OF MACHINE-LEARNING PROTOCOLS FOR TECHNOLOGY-ASSISTED REVIEW IN ELECTRONIC DISCOVERY 154 (2014), http://plg2.cs.uwaterloo.ca/~gvcormac/calstudy/study/sigir2014-cormackgrossman.pdf.

[34] Barry, *supra* note 30, at 354.

[35] *Id*.

[36] *See e.g.* ASHLEY, *supra* note 31, at 250.

A second example of an industry that is rapidly evolving due to AI is healthcare.[37] In another decade, the healthcare industry will look very different from today due to AI.[38] Currently, AI driven by big data is creating a noticeable shift in the practice of medicine from mass-market to personalized care.[39] Indeed, medical professionals practicing in modern hospitals now store patient data in electronic databases with Electronic Healthcare Records ("EHRs").[40] This allows machine-learning algorithms to analyze patient healthcare data and drastically improve patient care.[41] These data-driven resources not only allow a doctor to know virtually everything about a patient's medical history without ever meeting the patient, but also drastically reduce costs associated with healthcare by assisting in medical work.[42] For example, in 2016, researchers at Stanford developed AI that was able to diagnose lung cancer more accurately than human pathologists.[43] Another example is D-Wave's Adiabatic Quantum Computer, which is capable of running machine learning algorithms for cancer diagnostics.[44] In short, EHRs, big data, and AI are transforming the health-care landscape.[45]

A third example of AI disruption is occurring in the defense industry. AI is already an essential tool in cybersecurity.[46] Admiral Mike Rogers, Director of the National Security Administration has argued that AI and machine learning are foundational to the future of cyber security.[47] On March 2, 2017, a report was issued to the White House stating that Russian programmers launched an AI cyber-attack on the personal social-media accounts of over 10,000 employees at the Department of Defense.[48] Additionally, AI is used on the battlefield in modern warfare settings.[49] For example, the U.S. Phalanx missile-defense system for naval ships uses AI to detect, track, and attack threats from enemy missiles and aircraft.[50] However, terrorist misuse of commercial AI systems is a serious problem.[51] Terrorist organizations are already using AI systems in drones to deliver explosives and cause crashes.[52]

Narrow AI continues to change the way professional industries such as law, healthcare, and defense operate.[53] Several AI researchers have cited observable

[37] TEGMARK, *supra* note 5, at 102.

[38] TANEJA, *supra* note 28, at 73.

[39] *Id*.

[40] Kate Monica, *Apple EHR Patient Data Viewer Now in Use at 39 Health Systems*, EHRINTELLIGENCE (Apr. 2, 2018), https://ehrintelligence.com/news/apple-ehr-patient-data-viewer-now-in-use-at-39-health-systems.

[41] *See* XIAOQIAN JIANG ET AL., A PATIENT-DRIVEN ADAPTIVE PREDICATION TECHNIQUE TO IMPROVE PERSONALIZED RISK ESTIMATION FOR CLINICAL DECISION SUPPORT 137 (2012).

[42] *See* Alvin Rajkomar et al., *Scalable and Accurate Deep Learning with Electronic Health Records*, NATURE PARTNER J. (2018), https://www.nature.com/articles/s41746-018-0029-1.pdf.

[43] *See* Lloyd Minor, *Crunching the Image Data Using Artificial Intelligence to Look at Biopsies*, STANFORD MED. (2017), https://stanmed.stanford.edu/2017summer/artificial-intelligence-could-help-diagnose-cancer-predict-survival.html.

[44] *See* Brian S. Haney, *Quantum_Machine_Learning_Cancer_Diagnostics*, GITHUB (Feb. 24, 2019), https://github.com/Bhaney44/Leap/blob/master/Quantum_Machine_Learning_Cancer_Diagnostics.py.

[45] *Id*.

[46] *See generally* BRUNDAGE ET AL., *supra* note 5.

[47] ALLEN & CHAN, *supra* note 12, at 18.

[48] *See* Massimo Calbresi, *Inside Russia's Social Media War on America*, TIME (May 18, 2017), http://time.com/4783932/inside-russia-social-media-war-america/.

[49] *See United States Navy Fact File: MK 15 – Phalanx Close-In Weapons System (CIWS)*, U.S. DEP'T NAVY (last visited May 13, 2019), http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2.

[50] TEGMARK, *supra* note 5, at 111 (2017).

[51] *See generally* BRUNDAGE ET AL., *supra* note 5.

[52] *See id*.

[53] *See generally* TEGMARK, *supra* note 5.

patterns in historic information technology price and performance kinetics to support the argument that the rate of advancement of AI technologies will happen far more rapidly than expected.[54] Moreover, these researchers hypothesize AI technologies will continue to advance at an accelerating rate.[55]

B. The Law of Accelerating Returns

The Law of Accelerating Returns ("LOAR") states that fundamental measures of information technology will generally follow a predictable and exponential trajectory.[56] Indeed, information technologies build upon themselves in an exponential manner; this phenomenon has been named Moore's Law and is readily measurable in most processes where patterns of information evolve.[57] It describes the LOAR's application to the price and performance of computing[58] and was proposed by Gordon Moore, the founder of Intel, in 1965.[59] Moore's Law predicts that every eighteen months, the processing power of computers will double, while costs are cut in half.[60] It generally represents that the power of information technology doubles every one and a half years.[61] The past fifty-three years have proven Gordon Moore's prediction correct;[62] a smartphone today has more computing power than all of NASA had in 1969—when Apollo 11 landed on the Moon.[63] Applied to AI, Moore's Law has led many AI researchers to believe that we are currently at the cusp of developing super-intelligent AI.[64]

Irving J. Good first introduced the concept of superintelligence in 1965.[65] Good stated, "[l]et an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever."[66] According to Good, "[s]ince the design of machine is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an intelligence explosion, and the intelligence of man would be left far behind."[67] Indeed, Good predicted that the first ultraintelligent machine would be "the last invention that man need ever make."[68] Recent scholars have embraced Good's analysis and have defined superintelligence similarly. For example, Oxford Professor Nick Bostrom defines superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest."[69] Max Tegmark states that superintelligence is "[g]eneral intelligence far beyond human level."[70]

[54] *See* BOSTROM, *supra* note 22, at 85.

[55] *See* RAY KURZWEIL, HOW TO CREATE A MIND 250 (2012).

[56] *See id*.

[57] *See id*. at 256.

[58] *See* MARTINE ROTHBLATT, VIRTUALLY HUMAN 48 (2014).

[59] *See* KURZWEIL, *supra* note 55, at 251.

[60] *See* SUSSKIND, *supra* note 29, at 11.

[61] *See* ROTHBLATT, *supra* note 58, at 28.

[62] *See* KURZWEIL, *supra* note 55, at 251.

[63] *See* MICHIO KAKU, THE PHYSICS OF THE FUTURE 23 (2011).

[64] *See generally* BOSTROM, *supra* note 22; *see also* KURZWEIL, *supra* note 55; *see also* TEGMARK, *supra* note 5.

[65] *See generally* Irving J. Good, *Speculations Concerning the First Ultraintelligent Machine*, 6 ADVANCES IN COMPUTERS 31 (1966).

[66] *Id*. at 33.

[67] *Id*.

[68] *Id*.

[69] BOSTROM, *supra* note 22, at 22.

[70] *See* TEGMARK, *supra* note 5, at 39.

The application of the LOAR to AI is evidence that a transition from narrow AI to AGI and superintelligence may be much closer than commonly thought.[71] For now, the earliest estimate of AGI is 2029.[72] Indeed, Ray Kurzweil argues that the twenty-first century will yield what today may seem like 20,000 years of technological progress and innovation because of the LOAR.[73] Additionally, Bostrom and AI theorist Eliezer Yudkowsky have predicted a public perception of rapid kinetics in AI development due to anthropomorphism of AI.[74] Anthropomorphism of AI refers to the ascription of human levels of intelligence to non-human entities.[75] Humans may consider a village idiot and Albert Einstein extreme ends of the intelligence spectrum,[76] yet the difference between the two on a larger relative scale is actually *de minimis*.[77] Thus, the advancement of an AI system from the intelligence of the village idiot, to the intelligence of Einstein, to the intelligence of AGI, and finally superintelligence may be faster than expected.[78]

Interestingly, these predictions are supported by the massive amount of information humans began collecting at the dawn of the digital age.[79] Indeed, the amount of information humans collect is also accelerating.[80] Data, defined as a digital representation of information about the world,[81] is created at an astounding rate. Every two days, humans create more than five quintillion bytes of data, as much data as they did from the dawn of civilization up until 2003.[82]

Harvard professor and economist Michael Kremer argues, "the fundamental driver of human progress is not raw materials but technological solutions to problems."[83] In the context of AI, data is the driving force behind technological development instead of human programmers.[84] And the driving force of technological solutions is the realization that every piece of information can be represented as numbers.[85] The amount and type of data available for a particular problem largely determines the strength of AI systems that can be developed.[86] Thus, the LOAR will have a profound impact on the development of AI toward AGI and superintelligence.

Yet, some argue that AGI may never happen.[87] For example, the late Microsoft co-founder, Paul Allen asserts scientific progress is irregular and hypothesizes that by the end of the twenty-first century, humans will have yet to

[71] *See id*. at 157.

[72] *See* KURZWEIL, *supra* note 55, at 261.

[73] *See* Ray Kurzweil, *The Law of Accelerating Returns, in* KURZWEIL NETWORK (2001), http://www.kurzweilai.net/the-law-of-accelerating-returns.

[74] *See* BOSTROM, *supra* note 22, at 85.

[75] *See* Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, MACHINE INTELLIGENCE RES. INST., 21 (2008), https://intelligence.org/files/AIPosNegFactor.pdf.

[76] *See* BOSTROM, *supra* note 22, at 85.

[77] *See* Yudkowsky, *supra* note 75.

[78] *See* BOSTROM, *supra* note 22, at 86.

[79] *See* ETHEM ALPAYDIN, MACHINE LEARNING 11 (2016).

[80] *See* SUSSKIND, *supra* note 29 at 11.

[81] ALPAYDIN, *supra* note 79, at 3.

[82] SUSSKIND, *supra* note 29, at 11.

[83] Michael Kremer, "*Population Growth and Technological Change: One Million B.C. to 1990*," 108 Q. J. OF ECON. 3 (1993). (*quoting* SAIFEDEAN AMMOUS, THE BITCOIN STANDARD: THE DECENTRALIZED ALTERNATIVE TO CENTRAL BANKING (2018).

[84] ALPAYDIN, *supra* note 79, at 12.

[85] *Id*.

[86] SEBASTIAN RASCHKA & VAHID MIRJALILI, PYTHON MACHINE LEARNING, 2 (2d. ed. 2017).

[87] Paul G. Allen & Mark Greaves, *The Singularity Isn't Near*, MIT TECH. REV. (Oct. 12, 2011), https://www.technologyreview.com/s/425733/paul-allen-the-singularity-isnt-near/.

achieve AGI.[88] On the other hand, Max Tegmark suggests that the fundamental truth of the debate—whether humanity will ever build AGI—remains uncertain.[89] But Tegmark also explains that most AI experts project AGI will occur around 2047.[90] As one scholar argues, the questions about AI's impact will only become more urgent as we draw nearer to the exponential inflection point and its growth takes a sudden and dramatic vertical trajectory.[91] For now, the question is whether society is approaching that inflection point or if it is still in the slower gradual development phase.[92] Today, the clearest path that humanity has toward creating AGI is deep reinforcement learning.

II. AGI Development

Machine learning is a subfield of AI that focuses on the ability of machines to learn and replicate cognitive behaviors associated with the human mind.[93] Generally, machine learning involves data mining, pattern recognition, and natural-language processing.[94] These techniques have become increasingly popular in recent years due to the explosion in the amount of data humans have produced and collected since the dawn of the internet.[95] The most recent breakthrough in machine learning is deep reinforcement learning.[96] Deep reinforcement learning combines two traditional models of machine learning—supervised learning and reinforcement learning—to allow algorithms to learn independently from humans.[97]

Most scholarship in AI regulation focuses on either supervised or unsupervised methods of machine learning because until 2014, those were the only two types of machine learning in popular use.[98] Indeed, deep neural networks, a type of supervised learning algorithm, are the focus of most legal scholarship.[99] However, in 2013, Google developed a new type of learning called "deep reinforcement learning," which it subsequently patented.[100] Pioneered in the 1980s, reinforcement learning is a machine learning technique inspired by behaviorist psychology, where an intelligent agent's tendency to act in a certain way is influenced by a reward structure.[101] An intelligent agent is an entity that collects information about its environment from sensors and then processes that information to decide how to respond to its environment.[102] Deep reinforcement learning combines reinforcement

[88] *Id*.

[89] TEGMARK, *supra* note 5, at 54.

[90] *Id*. at 157.

[91] Michael Guihot et al., *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385, 400 (2017).

[92] *See id*.

[93] *See generally* ALPAYDIN, *supra* note 79.

[94] *See* Michael Simon et al., *Lola v. Skadden and the Automation of the Legal Profession*, 20 YALE J.L. & TECH 234, 253 (2018) (*quoting* Bernard Marr, *What Everyone Should Know About Cognitive Computing*, FORBES (Mar. 23, 2016, 3:28 AM), https://www.forbes.com/sites/bernardmarr/2016/03/23/what-everyone-should-know-about-cognitive-computing/#5630f9005088.

[95] *See id*. at 252.

[96] *See* TEGMARK, *supra* note 5, at 85.

[97] *See id*.

[98] *See id*. at 83.

[99] *See generally* Calo, *supra* note 6; *see also* John O. McGinnis, *Accelerating AI*, 104 NW. U. L. REV. 1253 (2010).

[100] '862 Application, *supra* note 15.

[101] RICHARD S. SUTTON & ANDREW G. BARTO, REINFORCEMENT LEARNING: AN INTRODUCTION 55 (2017); *see also* TEGMARK, *supra* note 5, at 85.

[102] TEGMARK, *supra* note 5, at 84.

learning with the use of deep neural networks.[103] Deep reinforcement learning refers to a reinforcement-learning algorithm using a deep neural network as a function approximator, which will be explained later in this Part.[104] First, this Part will explain deep neural networks. Second, this Part will explain reinforcement learning. Third, this Part will explain deep reinforcement learning.

A. Deep Neural Networks

The human brain is composed of processing units called "neurons."[105] Each neuron in the brain is connected to other neurons through structures called synapses.[106] A biological neuron consists of dendrites—receivers of various electrical impulses from other neurons—that are gathered in the cell body of a neuron.[107] Once the neuron's cell body has collected enough electrical energy to exceed a threshold amount, the neuron transmits an electrical charge to other neurons in the brain through synapses.[108] This transfer of information in the biological brain provides the foundation for the way in which modern neural networks operate.[109]

Indeed, artificial neurons are essentially logic gates modeled off of the biological neuron.[110] Both artificial and biological neurons receive input from various sources and map input information to a single output value.[111] An artificial neural network is a group of interconnected artificial neurons capable of influencing each other's behavior.[112] In an artificial neural network, the neurons are connected by weight coefficients modeling the strength of synapses in the biological brain.[113] Neural networks are trained using large data sets.[114] The training process allows the weight coefficients to adjust so that the neural network's output or prediction is accurate.[115] After a neural network is trained, new data is fed through the network to make predictions.[116]

In 1957, Frank Rosenblatt published an algorithm—the perceptron—that automatically learns the optimal weight coefficients for an artificial neural network.[117] The perceptron model is illustrated below:[118]

[103] Fei-Fei Li, Justin Johnson, & Serena Yeung, *Lecture 14: Deep Reinforcement Learning*, STANFORD U. SCH. OF ENG'G (2017), https://www.youtube.com/watch?v=lvoHnicueoE (last accessed May 13, 2019).

[104] *Id*.

[105] ALPAYDIN, *supra* note 79, at 86.

[106] *Id*.

[107] RASCHKA & MIRJALILI, *supra* note 86, at 18.

[108] *Id*.

[109] ALPAYDIN, *supra* note 79, at 86.

[110] KURZWEIL, *supra* note 55, at 38.

[111] JOHN D. KELLEHER & BRENDAN TIERNEY, DATA SCIENCE 131 (2018).

[112] TEGMARK, *supra* note 5, at 72.

[113] ALPAYDIN, *supra* note 79, at 88.

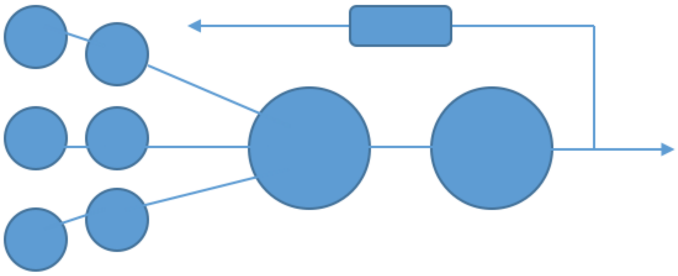
[114] *Id*. at 89.

[115] *Id*.

[116] KELLEHER, *supra* note 111, at 127.

[117] RASCHKA & MIRJALILI, *supra* note 86, at 18.

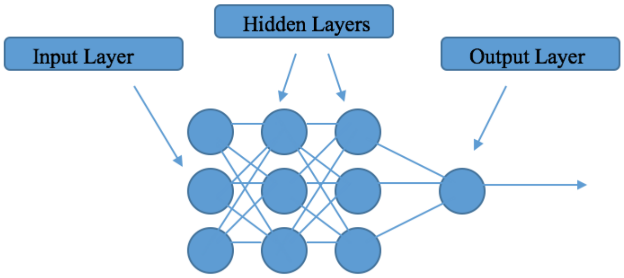
[118] *Id*.



In the perceptron, the three circles on the far left represent the input values $x_{j\ldots m}$ and the associated weight values $w_{j\ldots m}$ are the three circles to the right of the input values.[119] The input values and the weight values are aggregated, typically with a summation equation represented by the first big circle (from left to right).[120] The second large circle represents the threshold function, a predetermined value that, if exceeded, signals an output of 1.[121] If the threshold function is not exceeded, the model outputs a 0.[122] The output is represented by the arrow pointing right.[123] The box at the top of the model represents an error function.[124] In the event that the model's output is incorrect, then the error function is triggered.[125] If the error function is triggered, the weight values are updated pursuant to the perceptron learning rule.[126] The formal representation of the perceptron learning rule is defined as: $\Delta w_j = \eta\left(y^{(i)} - \hat{y}^{i}\right)x_j^{(i)}$, where η is the learning rate, $y^{(i)}$ is the true class label of the *i*th training sample, and $\hat{y}^{i}$ is the predicted class label.[127] The true class label is the output label, and the predicted class label is the perceptron's output.[128]

Every neural network has an input layer and an output layer.[129] However, in between the input and output layer, neural networks contain multiple hidden layers.[130] The number of hidden layers may vary and is dependent on the particular model.[131] It is important to note that while perceptron models are generally limited to linear classification tasks, this restriction does not apply to multi-layer networks.[132] Indeed, a multi-layer perceptron model is a universal approximator, which is an algorithm that can approximate any function with desired accuracy given enough neurons.[133] A deep neural network is a network that has multiple hidden layers.[134] This allows the neural network to account for several layers of abstraction.[135] The illustration below is a simple model of a deep neural network.[136]

[119] *Id*. at 19.
[120] *See* ALPAYDIN, *supra* note 79, at 89.
[121] *See Id*.
[122] *See* RASCHKA & MIRJALILI, *supra* note 86, at 18.
[123] *See* KURZWEIL, *supra* note 55, at 132.
[124] *See* RASCHKA & MIRJALILI, *supra* note 86, at 18.
[125] *See* ALPAYDIN, *supra* note 79, at 90.
[126] *See id*.
[127] *See* RASCHKA & MIRJALILI, *supra* note 86, at 21.
[128] *See id*. at 22.
[129] *See* KURZWEIL, *supra* note 55, at 132.
[130] *See* ALPAYDIN, *supra* note 79, at 100.
[131] *See* KELLEHER & TIERNEY, *supra* note 111, at 132.
[132] *See* ALPAYDIN, *supra* note 79, at 99.
[133] *See id*.
[134] *See* TEGMARK, *supra* note 5, at 76.
[135] ALPAYDIN, *supra* note 79, at 88.
[136] KELLEHER & TIERNEY, *supra* note 111, at 132 (model based on illustration at the following citation).



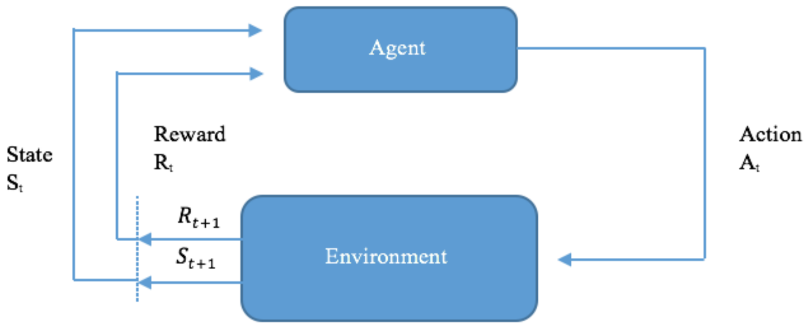
Each neuron represents a hidden unit in a layer and defines a complex feature of the model.[137] Hidden units correspond to hidden attributes defined in terms of what is observed, but not directly observed.[138] And the successive layers of hidden units correspond to increasing layers of feature abstraction.[139]

Indeed, each layer of hidden units acts as a feature extractor by providing analysis of slightly more complicated features.[140] Feature extraction is a method of dimensionality reduction—a method of decreasing input attributes—that allows raw input to be converted into output in a manner that allows data scientists to observe hidden features in data.[141] The later hidden units extract hidden features by combining the previous features in a slightly larger part of the input space.[142] The output layer observes the whole input to produce a final prediction.[143] In other words, deep neural networks learn more complicated functions of their initial input when each hidden layer combines the values of the preceding layer.[144] Additionally, deep neural networks have proven to be excellent for making predictions in several contexts.[145] However, these models require data to learn and at least a minimal amount of human intervention to supervise the learning process.[146] Reinforcement learning is a newer machine learning technique that requires neither.[147]

B. Deep Reinforcement Learning

Reinforcement learning is a type of machine learning technique inspired by behaviorist psychology.[148] Formally, reinforcement learning is described through an agent-environment interaction, with the Markov Decision Process ("MDP").[149] The model below describes the agent-environment interaction in an MDP.[150]

[137] ALPAYDIN, *supra* note 79, at 100.
[138] *Id.*
[139] *Id.*
[140] KELLEHER & TIERNEY, *supra* note 111, at 135.
[141] ALPAYDIN, *supra* note 79, at 102.
[142] *Id.*
[143] KURZWEIL, *supra* note 55, at 132.
[144] ALPAYDIN, *supra* note 79, at 104.
[145] *See generally* ASHLEY, *supra* note 31.
[146] ALPAYDIN, *supra* note 79, at 106.
[147] *Id.* at 127.
[148] SUTTON & BARTO, *supra* note 101, at 38.
[149] Alex Kendall et al., *Learning to Drive in a Day*, 1 (Sept. 11, 2018) (unpublished paper) (accessed through https://arxiv.org/abs/1807.00412).
[150] SUTTON & BARTO, *supra* note 101, at 38(model based on illustration at the preceding citation).



The environment is made up of states for each point in time in which the environment exists.[151] The agent's actions in each state determine the probabilistic evolution of the environment.[152]

Initially, the agent is presented with a state of the environment, which includes several possible actions.[153] Then, the agent takes an action in the state and advances to the next state of the environment, where a reward is returned.[154] The agent chooses which action to take when presented with a state based upon the agent's policy.[155] A policy is the way in which an agent makes decisions or chooses actions within a state.[156] For example, a person with a high amount of integrity has a policy that routinely guides their decision making to choose to do the right thing when faced with ethical dilemmas. Similarly, a greedy person has a policy that routinely guides their decision making to choose the action returning the highest dollar value. The goal of the policy is to allow the agent to advance through the environment so as to maximize a reward.[157]

A value-function defines the value of being in a state *s* and following a policy π until the final state of the environment, which is called the terminal state.[158] The terminal state concludes the episode, which is made up of all of the states in an environment.[159] The expected value of executing a policy π given state *s* is denoted as $V^{\pi}(s)$.[160] In the context of a MDP, the value function V^{π} is equal to the expected sum of the discounted rewards for executing policy π:[161]

$$V^{\pi}(s) \ = \mathrm{E}[R(s_0) + \gamma R(s_1) + \cdots | s_0 = s, \pi(s)]$$

The expected future rewards are discounted with a discount factor γ.[162] The discount factor is typically defined: $0 < \gamma < 1$.[163] This allows the value function to be defined

[151] Li, Johnson, & Yeung, *supra* note 103.
[152] MYKEL J. KOCHENDERFER, DECISION MAKING UNDER UNCERTAINTY 77 (2015).
[153] SUTTON & BARTO, *supra* note 101, at 39.
[154] KOCHENDERFER, *supra* note 152, at 77; *see also* SUTTON & BARTO, *supra* note 101, at 39.
[155] SUTTON & BARTO, *supra* note 101, at 39.
[156] KOCHENDERFER, *supra* note 152, at 77.
[157] SUTTON & BARTO, *supra* note 101, at 50.
[158] Li, Johnson, & Yeung, *supra* note 103.
[159] *Id*.
[160] KOCHENDERFER, *supra* note 152, at 80.
[161] Ahmad El Sallab et al., *Deep Reinforcement Learning Framework for Autonomous Driving*, (Apr. 8, 2017), (unpublished paper) (accessed at https://arxiv.org/pdf/1704.02532.pdf. *See Appendix A for Summary of Notation*).
[162] SUTTON & BARTO, *supra* note 101, at 92.
[163] *Id*.

in finite terms and allows the value of present rewards to be more valuable than future rewards.[164] The optimal policy $\pi^*(s)$ is defined as the policy that maximizes the expected value relative to other policies.[165] The objective of the MDP model is to find the optimal policy:[166]

$$\pi^*(s) = arg\max_{\pi} V^{\pi}(s)$$

The problem of finding the optimal policy for a given MDP is commonly solved with Q-learning.[167] Q-learning solves this problem by maximizing a Q-value function: Q(s, a).[168] A Q-value function describes the value of a state-action pair.[169] Indeed, the goal of a Q-learning algorithm is to discover the optimal Q-value function Q^* for any state-action pair.[170] The Bellman equation expresses the relationship between the value of a state and the values of its successor states.[171] The algorithm continues perpetually until the convergence of the Q-value function.[172] The convergence of the Q-value function represents Q^* and satisfies the Bellman Equation, defined as:[173]

$$Q^*(s,a) = \mathbb{E}_{s'\sim\varepsilon}\left[r + \gamma \max_{a'} Q^*(s',a')|s,a\right]$$

An agent's optimal policy π^* corresponds to taking the action in each state defined by Q^*.[174] However, one issue that arises is that the value of $Q(s,a)$ must be computed for every state-action pair, which may be computationally infeasible.[175] For example, computing the value of every state-action pair, where the raw input is pixels in an Atari game, would require tremendous computational power.[176] One solution is to use a function approximator to estimate the Q-value function:[177]

$$Q(s,a;\emptyset) \approx (s,a)$$

Here, $\emptyset$ represents the function parameters.[178] And if $\emptyset$ is determined by a Deep Neural Network, the algorithm is a deep reinforcement learning algorithm called a Deep Q-Network ("DQN").[179]

A DQN is a deep learning model that combines a Deep Neural Network ("DNN") with a Q-learning algorithm.[180] The DQN uses experience replay to

[164] *See* KOCHENDERFER, *supra* note 152, at 78.
[165] *See id*. at 79.
[166] Sallab et al., *supra* note 161, at 71–72.
[167] SUTTON & BARTO, *supra* note 101, at 108.
[168] *Id*. at 107.
[169] '862 Application, *supra* note 15, at 1.
[170] Li, Johnson, & Yeung *supra* note 103.
[171] SUTTON & BARTO, *supra* note 101, at 47.
[172] Sallab et al., *supra* note 161, at 72.
[173] '862 Application, *supra* note 15, at 5.
[174] MAXIM LAPAN, DEEP REINFORCEMENT LEARNING HANDS-ON, 102 (2018).
[175] Li, Johnson, & Yeung, *supra* note 103.
[176] LAPAN, *supra* note 174, at 125.
[177] '862 Application, *supra* note 15, at 5.
[178] *Id*.
[179] Li, Johnson, & Yeung, *supra* note 103.
[180] Manon Legrand, *Deep Reinforcement Learning for Autonomous Vehicle Control Among Human Drivers*, at 26 (academic year 2016–17) (unpublished C.S. thesis, Université Libre de Bruxelles) https://ai.vub.ac.be/sites/default/files/thesis_legrand.pdf.

maintain a buffer of old experiences of the algorithm to train a neural network.[181] An experience consists of an observed state-action pair, the immediate reward obtained, and the next state observed.[182] "An agent's experience at a time step *t* is denoted e_t and is a tuple (s_t, a_t, r_t, s_{t+1}) consisting of the current state s_t, the chosen action a_t, the reward r_t, and the next state s_{t+1}."[183] The experiences for all the time steps are stored in a replay memory, over many episodes, and are used to train the DNN.[184] The DNN's output corresponds to one valid action because the DNN serves as an approximator for the Q-value function.[185] Thus, after a feedforward pass of the network, the outputs are the estimated Q-values of the state-action pair.[186] This allows the algorithm to generalize from collected data of past experiences.[187] Indeed, according to MIT Professor Max Tegmark, "deep reinforcement learning is a completely general technique."[188]

III. ARTIFICIAL INTELLIGENCE REGULATION

Legal scholarship on the threat of AI is divided into two distinct camps.[189] One camp recognizes the potential threats posed by malicious and reckless use of AI, and the other argues an AI apocalypse is merely the talk of science fiction.[190] Neither of these camps truly grapple with the existential threats AI poses with the sense of immediacy required to prevent disaster.[191] First, this Part will discuss the arguments associated with the notion that AI poses no threat to humanity. Next, this Part will discuss arguments that have advanced scholarship in AI regulation to mirror the concerns of industry leaders. Lastly, this Part will address three ongoing and unanswered questions in AI regulation.

Scholars who argue an AI apocalypse is merely science fiction are wrong.[192] These scholars are represented by one in particular—John McGinnis—who notes, "the existential dread of machines that become uncontrollable by humans and the political anxiety about machines' destructive power on a revolutionized battlefield" are overblown.[193] Indeed, McGinnis attributes the problems associated with AGI to an error in thinking, where humans anthropomorphize AI and cause the mistaken fear that AGI will necessarily reflect human malevolence.[194] Thus, McGinnis suggests the possibility of friendly AI and encourages disposing of the assumption that AI

[181] HADO VAN HASSELT, ARTHUR GUEZ, & DAVID SILVER, ASS'N FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE, PROCEEDINGS OF THE THIRTIETH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE: DEEP REINFORCEMENT LEARNING WITH DOUBLE Q-LEARNING 2094, 2095 (2016).

[182] Volodymyr Mnih et al., *Human-Level Control Through Deep Reinforcement Learning*, 518 NATURE INT'L J. SCI. 529, 529 (2015).

[183] Legrand, *supra* note 180, at 72. A "tuple" is a data storage format similar to list or an array. *Id*. at 9.

[184] VAN HASSELT, GUEZ, & SILVER, *supra* note 181.

[185] Legrand, *supra* note 180, at 27.

[186] Mnih et al., *supra* note 182.

[187] KOCHENDERFER, *supra* note 152, at 124.

[188] TEGMARK, *supra* note 5, at 85.

[189] *See* Calo, *supra* note 6, at 432; *see also* Matthew U. Scherer, *Regulating Artificial Intelligence Systems:* 29 HARV. J.L. & TECH. 353 (2016).

[190] *See generally* McGinnis, *supra* note 99; *see also* Scherer, *supra* note 189, at 394.

[191] *See generally* BOSTROM, *supra* note 22, at 85.

[192] *See id.*; *see also* TEGMARK, *supra* note 5.

[193] *See* McGinnis, *supra* note 99, at 1254.

[194] *Id*.

must have willpower like a human.[195] He supposes that a lack of willpower should negate the fear surrounding evil AI.[196]

Interestingly, the anthropomorphic argument cuts both ways. In fact, Nick Bostrom and Eliezer Yudkowsky have convincingly made the anthropomorphic argument to explain why human beings will drastically underestimate the advancement of AI.[197] Bostrom and Yudkowsky argue that there will be a public perception of rapid kinetics in AI development due to human anthropomorphism of AI.[198] Again, human anthropomorphism of AI refers to the ascription of human levels of intelligence to non-human entities.[199] As illustrated by the comparison in Part I of Einstein and the village idiot, the difference between levels of intelligence on a larger relative scale is *de minimis*.[200] Thus, the advancement of AI to AGI and superintelligence, will be faster than expected because the difference in the two levels of intelligence on a broader scale is much narrower than humans realize.[201]

A second legal scholar, Ryan Calo, more bluntly argues that AI does not present an existential threat to humanity and that AGI is merely the "stuff of graphic novels."[202] Further, Calo contends that "devoting disproportionate attention and resources to the AI apocalypse has the potential to distract policymakers from addressing AI's more immediate harms. . . ."[203] He argues nothing in the field of machine learning suggests that humanity will soon be capable of modeling mammalian, let alone human intelligence.[204] This claim is patently misguided. Indeed, reinforcement learning and Markov Decision Processes quite literally model the human cognitive functions of decision making, rational agency, and intelligence.[205] Additionally, exponential increases in data production, computing power, and global GDP all lend support to the conclusion that AGI will arrive sooner than humans think.[206]

Therefore, the existential threat that AGI poses to mankind is an immediate harm. This threat is not analogous to a terminator-like robot taking over the world. Instead, this threat is the product of AGI developed from deep reinforcement learning agents.[207] Once AGI level agents are created, they will rapidly have the ability to improve their software architecture more efficiently than any human. These agents will be capable of accomplishing any goal correlated with a reward system in a virtual environment. Deep reinforcement learning systems are already capable of controlling missiles, rockets, cars, and aircraft.[208] And, the software for these

[195] *Id*. at 1263–64.

[196] *Id*.

[197] BOSTROM, *supra* note 22, at 85; *see* YUDKOWSKY, *supra* note 75, at 21.

[198] BOSTROM, *supra* note 22, at 85.

[199] YUDKOWSKY, *supra* note 75, at 21.

[200] *Id*.

[201] BOSTROM, *supra* note 22, at 86.

[202] Calo, *supra* note 6, at 432.

[203] *Id*. at 431.

[204] *Id*. at 432.

[205] *See also* TEGMARK, *supra* note 5, at 85; *see generally* BOSTROM, *supra* note 22, at 239.

[206] BOSTROM, *supra* note 22, at 1–4.

[207] C-SPAN, *supra* note 4.

[208] Mnih et al., *supra* note 182, at 529; *see also* Legrand, *supra* note 180, at 27; *see also* Kendall et al., *supra* note 149 ;*see also* U.S. Patent No. 8,678,321 to Bezos et al. (Mar. 5, 2014).

applications is open sourced.[209] So, today everyone with internet access also has potential access to the most sophisticated weapons control systems on the planet.[210]

And yet, legal scholarship completely ignores this unavoidable truth. But, some legal scholars have taken steps in the right direction without specifically addressing the issue of regulating AGI. For example, Matthew Scherer argues the starting point for regulating AI should be a statute establishing the general principles of AI regulation.[211] He proposes the Artificial Intelligence Development Act ("AIDA"), which would create an agency tasked with certifying the safety of AI systems.[212] The agency would be required to promulgate rules defining AI.[213] The main idea is that AIDA would delegate the substantive task of assessing the safety of AI systems to an independent agency staffed by specialists, thus insulating decisions about the safety of specific AI systems from the pressures exerted by electoral politics.[214]

Other scholarship discusses different regulatory frameworks that can be applied to analyze issues in AI as they arise, as well as a few concrete examples of problems in AI regulation.[215] The piece convincingly argues that AI, "no matter its potential, should be carefully handled."[216] Its authors advocate for a nuanced, responsive, and adaptive regulatory framework to foster innovation.[217] While limited progress has been made in the field of AI regulation, the rapid growth of research in intelligent-machine ethics and safety has not brought real progress.[218] As one piece notes, "[t]he great majority of published papers do little more than argue about which of the existing schools of ethics, built over centuries to answer the needs of human society, would be the right one to implement in our artificial progeny."[219] Further, even the more progressive scholarship in this field focuses quasi-exclusively on narrow AI rather than AGI.[220] Thus, none of the regulatory frameworks proposed by scholars have adequately addressed several important issues in the AGI development.

[209] *Welcome to Spinning Up in Deep RL!: User Documentation*, OPENAI (last visited Mar. 20, 2019) spinningup.openai.com.

[210] *TensorFlow 2.0 Alpha is Available*, TENSORFLOW, (2019), https://www.tensorflow.org/install; *see also* RICHARD WU ET AL., AAAI 2017 FALL SYMPOSIUM SERIES, A FRAMEWORK USING MACHINE VISION AND DEEP REINFORCEMENT LEARNING FOR SELF-LEARNING MOVING OBJECTS IN A VIRTUAL ENVIRONMENT (2017), https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16003/15319.

[211] Scherer, *supra* 189, at 394.

[212] *Id*. at 393.

[213] *Id*. at 394.

[214] *Id*. at 393.

[215] *See generally* Guihot et al., *supra* note 91.

[216] *Id*. at 454.

[217] *Id*.

[218] *See* Roman Yampolskiy & Joshua Fox, *Safety Engineering for Artificial General Intelligence*, 32 TOPOI 217 (2012).

[219] *Id*.

[220] *Id*.

A. Contemporary Issues in AGI Development

Experts suspect cyber-attackers will soon begin implementing strategies that use deep reinforcement learning agents to craft attacks that current technical defense systems are incapable of preventing.[221] Indeed, one scholar specifically details guidelines for the development of malicious AI software.[222] The scholarship was written to demonstrate that it is practically possible to develop machine learning algorithms that are capable of harming humans.[223] Additionally, humans today already have the power to destroy life on planet Earth with the use of nuclear weapons, and an AGI would certainly have the same capability.[224] Modern AI scholars analogize the process of building AGI, specifically deep reinforcement agents, to the building of nuclear weapons.[225] This Part proceeds by identifying three specific issues that any adequate regulatory framework for AI would need to accommodate.

The first issue is the competition problem. If regulators attempt to provide oversight to companies developing AGI, then this oversight will stifle innovation and will allow countries like China and Russia to develop AGI before the United States.[226] Indeed, there is a strong possibility that any entity that creates AGI will have a decisive advantage over the rest of the world.[227]

For example, DQN algorithms are commonly used to trade stocks, where an agent is able to take the actions of buying, selling, or holding a stock in each given state.[228] The agent's goal is to maximize the value of a portfolio.[229] The use of DQN algorithms for portfolio management has been successful.[230] If an entity could create AGI, then it could be used to create an agent capable of manipulating markets in a way that would allow a single actor to garner extraordinary amounts of wealth over a minimal period of time.[231] This would allow such an entity to evolve to become a unified central power of authority unbeknownst to the masses.[232]

The competition problem becomes even more daunting considering the major players in AI today are publicly traded companies. Companies like Google, Facebook, Apple, and Microsoft are some of the biggest players in AI development, and their technology is rapidly increasing in power and scalability.[233] The power disparity between these corporate actors, foreign governments, and the United States poses further problems for regulators.[234] If the federal government begins regulating AI, it must be wary that slowing the pace of progress domestically will surely put the United States at a disadvantage against foreign actors. The ultimate issue of the

[221] Brundage et al., *supra* note 5, at 34; *see also* Hyrum S. Anderson, et al., Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning, 2 (2018) (accessed at https://arxiv.org/abs/1801.08917).

[222] Federico Pistono & Roman Yampolskiy, *Unethical Research: How to Create Malevolent Artificial Intelligence* (2016), (unpublished article) (accessed at https://arxiv.org/abs/1605.02817).

[223] *Id*.

[224] Yampolskiy & Fox, *supra* note 218.

[225] Bostrom, *supra* note 22, at 104.

[226] Tegmark, *supra* note 5, at 9.

[227] Bostrom, *supra* note 22, at 103.

[228] Lapan, *supra* note 174, at 217.

[229] *Id*. at 220.

[230] *See generally* Zhipeng Liang et al., *Deep Reinforcement Learning in Portfolio Management*, (Aug. 29, 2018), (unpublished paper) (accessed at https://arxiv.org/abs/1808.09940).

[231] Tegmark, *supra* note 5, at 15–16.

[232] *Id*.

[233] Guihot et al., *supra* note 91, at 437.

[234] *Id*.

competition problem is that regulators are faced with a balancing of interests between security and freedom. If regulators place a heavier emphasis on security, they do so at the expense of the freedom that has allowed domestic industry leaders in technology to innovate. On the other hand, if regulators place a heavier emphasis on freedom, they do so at the expense of the security of the electorate. Therefore, regulators must design a framework that is sensitive to the competition between corporations, foreign governments, and national-security agencies.

The second issue regulators face is the "lone-wolf" concept in which a threat is viewed as an isolated incident as opposed to a broad societal issue. In many ways, regulating AI is analogous to the regulation of mathematics or computer science. Indeed, AI research requires only a personal computer.[235] Interestingly, scholars are torn as to the size of a potential project to develop AGI.[236] One scholar notes that the path to AGI could be achieved as part of a massive government project from the work of a small group or even the work of single individual.[237] The scale of the path to AGI in large part depends on the methods used to achieve AI.[238] For example, if the current methods of whole-brain emulation are employed it is likely that massive amounts of code will need to be developed by expert computer scientists and engineers to develop AGI.[239] It is important to note that while a project itself may be massive in scale, the individual group tasked with making the breakthrough from AI to AGI may be very small.[240] For example, the Manhattan Project employed roughly 130,000 people at its peak.[241] Yet the atomic bomb was created by a smaller group of scientists and engineers, led by J. Robert Oppenheimer and General Leslie Groves at the Los Alamos Scientific Laboratory.[242]

Another issue of the lone-wolf problem will manifest if the field of AI experiences a breakthrough by a single individual. In which case, it is possible that everything we currently know about AI could fall by the wayside. Science is no stranger to simple yet revolutionary breakthroughs that radically alter the way in which humankind understands the natural world.[243] Yet, one scholar argues it is likely that regulatory bodies could be aware of most people potentially capable of developing AGI.[244] Although it should be noted that an epiphany in AI, like that of Einstein's in physics expressed in the *Annus Mirabilis* papers, should not be ruled out of the realm of possibility. Therefore, it is possible that a single individual could be the first to create AGI and could shortly thereafter attain an unprecedented degree of power.[245] Regulators will need to design a framework that allows for the implementation of technology to identify and prevent lone-wolf AGI attacks and threats.

The third issue technology regulators face is the control problem. The control problem can be analyzed through a principal agent framework in two distinct ways.[246]

[235] BOSTROM, *supra* note 22, at 103.

[236] *Id*.

[237] *Id*. at 101.

[238] *Id*.

[239] KURZWEIL, *supra* note 55, at 124.

[240] BOSTROM, *supra* note 22, at 101.

[241] F.G. GOSLING, THE MANHATTAN PROJECT: MAKING THE ATOMIC BOMB 54 (1999).

[242] *Id*. at 35.

[243] *See generally* Albert Einstein, *On the Electrodynamics of Moving Bodies* (1905) *in* 2 THE COLLECTED PAPERS OF ALBERT EINSTEIN: THE SWISS YEARS: WRITINGS, 1900-1909 140 (Anna Beck trans., 1990), https://einsteinpapers.press.princeton.edu/vol2-trans/154; *see also* BRIAN GREENE, FABRIC OF THE COSMOS 128 (2005).

[244] BOSTROM, *supra* note 22, at 103.

[245] *Id*.

[246] *Id*. at 155.

The first framework exists where there is a project sponsor acting as a principal and a group of scientists and engineers acting as agents of the project sponsor.[247] In this framework, the control problem manifests if the scientists and engineers developing AGI use the knowledge and information they gain in the course of their work for malicious purposes.[248] For example, as a result of their respective companies' AI development, researchers at Apple, Google, and Facebook have gained immense power[249] and the capability of developing or altering advanced AI systems for their own personal gain or to the detriment of others.

In the second framework, the principal is the human creator and the agent is the AI system.[250] In this framework, the control problem manifests if an AGI system is developed and its actions are uncontrollable by its creator.[251] Several different methods of containing AGI have been presented. For example, Nick Bostrom has proposed boxing methods to subdivide and contain AGI's access to information.[252] Additionally, Max Tegmark has suggested the creation of a "Gatekeeper AI," a superintelligence with the goal of interfering as little as necessary to prevent the creation of another superintelligence, is possible.[253] Therefore, regulators will need to design a framework that controls the way in which AI researchers use their power and a framework which allows for the regulation of AGI systems, so they can be controlled by human actors.

In sum, three major problems faced by AI regulators are competition, the lone-wolf concept, and control. One practical way in which these problems may be addressed practically is with self-regulating AGI technology.[254] In essence, this will require the programming of AGI values in alignment with the values of the AGI's creator.[255] One major benefit of this solution is that it allows for public regulators to stay relatively in the dark regarding how AI technology works.[256] However, there are two major outstanding issues with this solution. First, if there is an AGI system capable of regulating all other AI systems, there will need to be a regulatory mechanism to contain the regulatory AGI so that it does not become a unified power with control over humans. Second, humanity must ensure that the regulatory AGI is not outmatched or overpowered by any other AI or AGI. Indeed, if there was an AGI capable of improving itself, the abilities of any human programmer would be swiftly left far behind, while Irving J. Good's infamous words, "…the last invention that man need ever make…" will echo in prophetic nature.[257]

CONCLUSION

There is a spectrum of possibility laid out in scholarship. On one end are those who argue that AI will forever change human life in the near future, and on the other are those who argue an AI apocalypse is merely science fiction. The truth is

[247] *Id*. at 155–56.
[248] *Id*.
[249] Guihot et al., *supra* note 91, at 455.
[250] BOSTROM, *supra* note 22, at 156.
[251] TEGMARK, *supra* note 5, at 187.
[252] BOSTROM, *supra* note 22, at 156.
[253] TEGMARK, *supra* note 5, at 176.
[254] Guihot et al., *supra* note 91, at 433–37.
[255] TEGMARK, *supra* note 5, at 261.
[256] BOSTROM, *supra* note 22, at 156.
[257] Good, *supra* note 65.

that neither camp fully understands AGI or the impacts it could have on our world.[258] At the time of his death in 1988, Nobel Prize-winning physicist Richard Feynman's blackboard contained the words "[w]hat I cannot create, I do not understand."[259] It follows that until mankind creates AGI, it is beyond the comprehension of mankind. This reality poses an ironic fate for humankind. Indeed, humanity must first understand AGI to control it, yet humans cannot understand AGI until it is created. Further, according to Max Tegmark, "we have no idea what will happen if humanity succeeds in building human-level AG."[260] Thus, we cannot take for granted that the outcome will be positive if AGI is created.[261] Indeed, the general consensus in the field of AI is that no set of rules is capable of controlling the totality of what AGI regulation requires.[262] But no matter how many patterns can be recognized and trends can be traced, the future of AI will not happen on its own.[263]

Most people think that the past has a deterministic relationship with the future, but the truth is that the future is fundamentally uncertain.[264] Since the early twentieth century, humanity has possessed conclusive evidence that the entirety of the space-time that humans perceive in their everyday experience only exists relative to an individual's subjective observation.[265] And, in quantum physics, as well as at the center of black holes, the laws of classical physics and the laws of space-time breakdown completely.[266] This is important because without space-time at a fundamental level of existence, the independence of massive particles evaporates and the forward flow of time humans perceive ceases to exist.[267] This is evidenced by the principles of superposition, non-locality, time-symmetry, and quantum entanglement.[268] Indeed, the future, as well as the past, exists in a fundamental state of quantum uncertainty. AI is the key to maximizing the probability of prosperity in the face of such uncertainty. Thus, we need to immediately work to create a safe and prosperous future.[269]

[258] *See* BOSTROM, *supra* note 22 (arguing that we already live in a simulation; this is a real possibility and a strong the argument exists that some entity, possibly in space-time does understand AGI).

[259] *See* Michael Way, "*What I Cannot Create, I Do Not Understand,*" J. OF CELL SCI. (2017), http://jcs.biologists.org/content/joces/130/18/2941.full.pdf.

[260] *See* TEGMARK, *supra* note 5, at 156.

[261] *See* PETER THEIL, ZERO TO ONE 195 (2014).

[262] *See* Yampolskiy & Fox, *supra* note 218.

[263] *See* THEIL, *supra* note 261.

[264] *See generally* Einstein, *supra* note 243.

[265] *See generally id*.

[266] *See* STEPHEN HAWKING, A BRIEF HISTORY OF TIME 111 (1996); *see also* ROBERT J. SPITZER, NEW PROOFS FOR THE EXISTENCE OF GOD 123 (2010).

[267] *See* GREENE, *supra* note 243, at 192–93.

[268] *See id*. at 199–208.

[269] *See* THEIL, *supra* note 261.

Appendix A

Summary of Notation	
Notation	**Meaning**[270]
$Q^*(s, a)$	Value of taking action *a* under the optimal policy
γ	Discount factor
$\mathbb{E}[x]$	Expectation of random variable
$arg\max_a f(a)$	A value of a, at which $f(a)$ takes its maximal value.
r	Reward
s_t	State at time t
π	Policy
π^*	Optimal policy
$V^\pi(s)$	Expected value of executing policy from a given state.

[270] *See generally* SUTTON & BARTO, *supra* note 101.