BEYOND THE IUDEX THRESHOLD: HUMAN OVERSIGHT AS THE CONSCIENCE OF MACHINE LEARNING

E. JASON ALBERT* AND JESSICA E. BROWN**

Artificially intelligent machines do not need to rise to the level of human cognition, a fête popularized in science fiction, to cause irreversible harm. Due to the speed at which machines consume data and the automated programming that allows machines to "learn" from that data, machines do not simply surpass human computational abilities but can now rewrite their own code in response to their environments. Thus, machines now have the potential to make "judgments" that go beyond their programming.

This article focuses on those judgments that are based on machine learning, but which go beyond the basic programming of an algorithm, and in that sense are unintended by the creators of the artificial intelligence. We say artificial intelligence ("AI") that is able to make judgments that go beyond initial programming has crossed the "Iudex Threshold." Like human judgments, these machine judgments need to be constrained to comply both with laws and societal norms. Yet constraining machine judgments that cross the Iudex Threshold presents novel challenges. Too much constraint robs us of the benefit of an AI that learns and evolves—the very promise of machine learning. Too little constraint, and unanticipated harms will arise.

In Part I, this article discusses the nature of machine learning, how it differs from general AI, and how machine learning happens within particular environs. Part II then explores why regulations around machine learning must anticipate divergences from the notion that laws and regulations are, fundamentally, an activity of social planning or democratic processes rather than computational

* E. Jason Albert, Global Chief Privacy Officer for Automatic Data Processing, Inc. and graduate of Harvard Law School. This article represents the authors' personal views and not necessarily those of their employers. The authors would like to thank Professors Mark Lemley, Christopher Sprigman, Camille Crittenden, Ph.D. and Susan Athey, Ph.D. for their thoughtful feedback on our drafts. Mistakes are our own.

** Jessica E. Brown, Deputy Attorney General for the State of Nevada, a graduate of Boyd School of Law, and a former technologist and developer.

logic. The result of this discord in perceptions could result in socially sub-optimal outcomes if no action is taken and Iudex Thresholds are crossed without thought. Part III explains why existing regulatory paradigms, including liability rubrics such as tort and criminal law as well as regulatory frameworks like those associated with privacy law, are ineffective with respect to constraining machine judgments beyond the Iudex Threshold, and lead to sub-optimal outcomes of either too much regulation, hindering innovation, or too little control. Part IV posits a two-part approach to regulation designed to address these shortcomings: (1) prohibition of certain high-risk judgments combined with (2) a risk-based approach to designing regulation of other judgments. Finally, Part V argues why human oversight is an essential supplement to this new regulatory approach and sets forth why the true objective of human oversight is to serve as the machine's conscience in adherence to laws and social norms.

INTRODUCTION 270
I. STRONG AI IS NOT NECESSARY FOR MACHINE JUDGMENTS 276
A. *Machine Learning: Where Artificial Intelligence Meets Neural Networks* 277
B. *What are Machine Judgments?* 281
C. *What Happens When Humans Rely on Machine Judgments?* 282
II. THE INTERSECTION OF LAW AND MACHINE LEARNING 282
A. *Machine Judgments and Legal Rules* 284
B. *Machines Follow Their Programming* 285
III. INTENT AND CAUSATION DO NOT WORK TO REGULATE MACHINE JUDGMENTS 286
IV. THE SOLUTION: PROHIBITIONS PLUS A RISK-BASED APPROACH 290
V. HUMAN OVERSIGHT IS NEEDED 295
A. *The Moral Imperative Driving Human Oversight* 295
B. *Designing Human Oversight to Be Effective* 298
C. *The Right to Challenge Is Not a Substitute* 298
CONCLUSION 299

INTRODUCTION

Discussions about machine learning often focus on "fixing" improper decisions that machines make due to unforeseen

consequences of algorithmic programming.[1] But fixing machine decisions will soon be impossible as machine learning becomes ubiquitous and the complexity of algorithms continues to increase. This article argues that active human interference, constraint, and oversight should be required to prevent potential harm caused by machine judgments.

The rise of generative artificial intelligence ("AI"), such as ChatGPT, which utilizes large language models, has renewed questions about human oversight in machine learning pursuits. Generative AI solutions "learn"[2] from large corpora and their prompts[3], returning answers that become ever more refined, tailored, and natural. Generative AI's mimicry of human interactions is compelling, but it is difficult to know if it is telling the truth. Like all machine learning, ChatGPT uses prompts to deliver a conversational response to complex inquiry or to suggest code to solve a problem. Where asked to provide legal analysis, though, it falters; because of the way it was trained on language, it can often identify legal rules, but it will make up cases and citations because it "knows" they should exist.[4]

This may seem like a simple challenge to solve: test whether the code works, or Shepardize case cites. Indeed, Microsoft's Code of Conduct for the Azure OpenAI service specifically prohibits users from relying upon it in making decisions that have significant impacts on individuals without applying human oversight.[5] But human oversight is sustainable only where the AI acts in ways anticipated by its creators—like ChatGPT does in responding to prompts. It is one thing to train a large language model and then to

1. *See, e.g.*, Rebecca Crootof et al., *Humans in the Loop*, 76 VAND. L. REV. 429, 474–77 (2023) (focusing on the corrective role humans can play with machine decisions and defining three corrective modes: error correction, situational correction, and bias correction).

2. Machine learning algorithms do not learn in the sense that humans do. Rather a model is trained on data and adjusts its output based on feedback about what results are valuable or what patterns it discerns. This is qualitatively different from human cognition, although it can increasingly result in similar results when answering a question. Sara Brown, *Machine Learning, Explained*, MIT SLOAN (Apr. 21, 2021), https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained [https://perma.cc/H54F-PVRU].

3. When users input language or an image into an AI system in order to produce content, that language or image is referred to as a "prompt."

4. Larry Neumeister, *Lawyers Submitted Bogus Case Law Created by Chat GPT. A Judge Fined Them $5,000*, ASSOCIATED PRESS (June 22, 2023), https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c [https://perma.cc/YL4D-6VC5].

5. *Code of Conduct for Azure OpenAI Service*, MICROSOFT, https://learn.microsoft.com/en-us/legal/cognitive-services/openai/code-of-conduct [https://perma.cc/TR3U-MXS9] (last visited Apr. 10, 2024).

fine tune it on top of this training or provide detailed information within the prompt itself. But where such a model learns from unmediated third-party prompts (that is, prompts directly input from the third party without intervention by the model creator or others), the creator of the model may not be able to exercise control over its outputs and evolution.

Indeed, the risk of ChatGPT has led to concerns about the risks of AI and whether it is evolving beyond our control or means to regulate. There are numerous calls for regulation, including from the CEOs of OpenAI, Microsoft, and Google.[6] The Biden administration has launched numerous consultations—ranging from the National Telecommunications Infrastructure Administration's consultation on assessment of AI systems to the Office of Science and Technology Policy's request for information on national security risks and the economic potential of AI.[7] The European Union ("EU") has advanced an AI Act designed to regulate high-risk AI systems, but even it seems to have fallen behind technological developments, with its focus on specific applications of AI rather than a broad general-use scenario, such as large language models; consequently it rushed to make changes before the legislation was finalized.[8]

Yet these approaches do not fully address the primary risk posed by AI, which is caused by one area in which it fails to be human. Unlike human actors, machines do not have ethical sensibilities or the ability to internalize social norms. For all of the knowledge and understanding machines gain from the ingestion of training data and the tuning that comes from repeated prompting, the results produced are largely operational—providing information or decisions that work but are unmoored from any moral compass.

6. Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing*, N.Y. TIMES (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html [https://perma.cc/6M7T-4TFR]; Brian Fung, *Microsoft Leaps into the AI Regulation Debate, Calling for a New US Agency and Executive Order*, CNN (May 25, 2023), https://www.cnn.com/2023/05/25/tech/microsoft-ai-regulation-calls/index.html [https://perma.cc/B9CM-9PD9].

7. Stacy Murphy, *Request for Information: National Priorities for Artificial Intelligence*, WHITE HOUSE OFF. OF SCI. AND TECH. POL'Y (May 23, 2023) https://www.whitehouse.gov/wp-content/uploads/2023/05/OSTP-Request-for-Information-National-Priorities-for-Artificial-Intelligence.pdf [https://perma.cc/VE76-SFUK].

8. *EU AI Act: First Regulation on Artificial Intelligence*, EUR. PARL. NEWS (Dec. 19, 2023), https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [https://perma.cc/Z5P7-TMYD].

The focus of this article is automated decision-making, or the "machine judgments" that an AI makes in the course of its operation.[9] Specifically, it is focused on judgments that are based on machine learning, but go beyond the basic programming of the algorithm, and in that sense, are unintended by the creators of the AI. We say that AIs that can make judgments that go beyond initial programming have crossed the "Iudex Threshold."[10] We chose this name because in Latin *iudex* means judge or decider, and when AI can make judgments, it enters into the realm of making decisions that impact the real world.

Concern and fear about machine judgments is not new, nor is fury over mistakes machines make. British-American computer scientist Stuart Russell penned an open letter with over eight thousand scientist-signatories. The letter states:

> The potential benefits [of AI] are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.[11]

Other scholars are less reserved. Eric Horvitz, American computer scientist and Technical Fellow at Microsoft, warns, "[b]ecause of AI's potential transformational capabilities and broad reach, the government needs a holistic, forward-looking evaluation of AI oversight and governance."[12]

9. One might state that these are decisions the machine is programmed to make. And for traditional software this would be true. But with the advent of machine learning, machines increasingly make decisions that are not contemplated by their programmers. It is the governance of these decisions that is our focus.

10. *See, e.g.*, *Corrigendum to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/…… of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)*, recital 12, P9_TA(2024)0138 (Apr. 19, 2024) [https://perma.cc/GW2K-89EQ] (defining "AI system" to include machine-learning and logic-based approaches that have the capability to infer) (hereinafter EU AI Act). As will be explained further into the article, the Iudex Threshold is the point at which machine output is no longer within the control of the initial algorithmic programming. Algorithms are a type of instruction that can be executed.

11. *Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter*, FUTURE OF LIFE INST. (Oct. 28, 2015), https://futureoflife.org/2015/10/27/ai-open-letter/ [https://perma.cc/2AJD-QRHK].

12. Eric Horvitz et al., *Caution Ahead: Navigating Risks to Freedoms Posed by AI,* THE HILL (May 17, 2021, 3:30 PM), https://thehill.com/blogs/congress-

Yet constraining machine judgments presents novel challenges. What exactly should be constrained? Too much constraint robs us of the benefit of an AI that learns and evolves—the very promise of machine learning. And how can we identify which constraints are necessary for technology that is presently in production? The answer is that we now have the prototypical and rudimentary beginnings of artificial general intelligence, and we already have an understanding that machines can be intentionally and unintentionally programmed in ways that cause harm. Under current machine learning models, machines can now rewrite their own code in response to their environments. Machines now also have the potential to make "judgments," supplanting human actors.

The European Parliament entertained addressing this ethical deficiency by programming machines to respect a series of rules.[13] Setting aside the challenge of determining exactly what rules the machine—which may be used for many different purposes across many different geographies—should respect, it is simply impossible to program compliance with all rules applicable to the circumstances that the machine may encounter. Indeed, think of the large number of questions one might ask ChatGPT: how to program malware, how to commit a crime, how to deceive someone.

This article argues that regulation remains possible. At the most basic level, any intelligent machine which acts on its own in the "real world" requires some level of human oversight to prevent harm. Therefore, in our view, active human interference, constraint, and oversight should be required to prevent the harm caused by machine judgments. The required level of oversight will differ depending on context. Self-driving cars programmed to safely transport people from point A to point B will not need considerable human oversight because AI driving machines will not likely make decisions outside of their original programming. Self-driving cars will likely make individual transportation safer than human drivers. But for machines which are "capable of independent initiative and of making their own plans" as Oxford professor and

blog/technology/553932-caution-ahead-navigating-risks-to-freedoms-posed-by-ai [https://perma.cc/6SGR-HJDG].

13. *See The Ethics of Artificial Intelligence: Issues and Initiatives*, EUR. PARL. DOC. (PE 634.452) 90 (2020) https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf ("Devising a method for integrating ethics into the design of AI has become a main focus of research over the last few years. Approaches towards moral decision making generally fall into two camps, 'top-down' and 'bottom-up' approaches (Allen et al., 2005). Top-down approaches involve explicitly programming moral rules and decisions into artificial agents, such as 'thou shalt not kill'. Bottom-up approaches, on the other hand, involve developing systems that can implicitly learn to distinguish between moral and immoral behaviours.").

futurist Nick Bostrom suggests, human oversight, when such a machine breaks the Iudex Threshold, is essential.[14]

But such regulation needs to take a different form than legal regimes have envisioned to date—even in this new era of generative AI. It cannot be a prescriptive set of rules because it is impossible to program all of the potential rules that must be followed into a machine. And even if it were, as humans, we apply judgment in deciding whether, how, and to what extent to follow rules—reasoning attributes a machine lacks. Rather, only a risk-based approach supplemented by human oversight can suffice to constrain machines in an effective way given how legal rules operate in the real world.

In Part I, this article discusses the nature of machine learning, how it differs from general AI, and how machine learning happens within particular environs. Part II then explores why regulations around machine learning must anticipate divergences from the notion that laws and regulations are, fundamentally, an activity of social planning or democratic processes rather than computational logic. The result of this discord in perceptions could result in socially sub-optimal outcomes if no action is taken and Iudex Thresholds are broken without thought. Part III explains why existing regulatory paradigms, including liability rubrics such as tort and criminal law as well as regulatory frameworks like those associated with privacy law, are ineffective with respect to constraining actions beyond the Iudex Threshold and lead to sub-optimal outcomes of either too much regulation, hindering innovation, or too little control. Part IV posits a two-part approach to regulation designed to address these shortcomings: (1) prohibition of certain high-risk judgments combined with (2) a risk-based approach to designing regulation of other judgments. Finally, Part V argues why human oversight is an essential supplement to this new regulatory approach and sets forth why the true objective of human oversight is to serve as the machine's conscience in adherence to laws and social norms.

14. Nick Bostrom, *When Machines Outsmart Humans*, 35:7 FUTURES 759, 764 (2000).

I. STRONG AI IS NOT NECESSARY FOR MACHINE JUDGMENTS

AI produces content that is often unpredictable even to the developers of the AI system.[15] While Elon Musk's claim that "[o]ne of the biggest risks to the future of civilization is AI" may be hyperbole, the ability of large language models to interact with humans using natural language leads to a whole host of concerns—accuracy, manipulation of beliefs and feelings, and the ability to drive action, if not directly, then through willing intermediaries responding to its suggestions.[16]

Machine learning will intensify the gap between technological advancements and the regulations intended to plan for and confine them. This is because ethics and values are highly dependent on cultural norms, personal and group histories,[17] and perhaps most importantly social hierarchal conceptions that an AI does not inherently possess. Humans do not simply call balls and strikes. Humans routinely display acts of mercy and forgiveness, for example, in consideration of social hierarchies. These inclinations may not be conscious acts. As another example, we leave space in our laws so that human actors, including judges, can account for those considerations. The gap between technology and regulations will widen when developers are not transparent about how the AI makes decisions,[18] but even when they are, the AI, by its nature, may not itself be able to provide an ethical lens to govern its actions.

15. Ethan Mollick, *ChatGPT Is a Tipping Point for AI*, HARV. BUS. REV. (Dec. 14, 2022), https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai [https://perma.cc/U2WT-WLUN].

16. Ryan Browne, *Elon Musk, Who Co-founded Firm Behind ChatGPT, Warns A.I. Is 'One of the Biggest Risks' to Civilization,* CNBC (Feb. 15, 2023), https://www.cnbc.com/2023/02/15/elon-musk-co-founder-of-chatgpt-creator-openai-warns-of-ai-society-risk.html [https://perma.cc/8ALA-ASST].

17. *See generally* NOMY ARPALY, UNPRINCIPLED VIRTUE: AN INQUIRY INTO MORAL AGENCY (Oxford Univ. Press 2003); MARK BALAGUER, FREE WILL AS AN OPEN SCIENTIFIC PROBLEM (MIT Press 2001); Paul Benson, *Culture and Responsibility: A Reply to Moody-Adams*, 32 J. SOC. PHIL. 610 (2001); John Christman, *Autonomy and Personal History*, 21 CAN. J. PHIL. 1 (1991); RANDOLPH CLARKE, OMISSIONS: AGENCY, METAPHYSICS, AND RESPONSIBILITY (Oxford Univ. Press, 2006); STEPHEN DARWALL, THE SECOND-PERSON STANDPOINT: MORALITY, RESPECT, AND ACCOUNTABILITY (Harv. Univ. Press 2009); Gerald Dworkin, *Acting Freely*, 4 NOÛS 367 (1970); JOHN MARTIN FISCHER, THE METAPHYSICS OF FREE WILL: AN ESSAY ON CONTROL (Cambridge Univ. Press 2009); ALFRED R. MELE, AUTONOMOUS AGENTS: FROM SELF-CONTROL TO AUTONOMY (Oxford Univ. Press, 1995).

18. *See, e.g.*, Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan.–June 2016, at 1, 3; Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 981–82 (2016);

Scholars agree that "Artificial General Intelligence" (AGI), also called "Strong AI," (a concept wherein machines replicate human cognitive functions such as reasoning, planning, and problem-solving) "is aspirational."[19] We are not yet in an age of Strong AI.[20] While it is true that machines today do not employ abstract thinking, a theory of mind, or operate as fully independent cognitive systems, that level of cognition is not a prerequisite for significant harm to be caused by machine judgments. Indeed, the failure of machines to do exactly these things may in fact increase the likelihood of potential harm, because abstract reasoning governs human ethical decisions including those related to legal compliance and respect for societal norms. A generative AI model may not engage in Strong AI cognition, but it can still develop answers and determinations that, if acted upon, could lead to harm. Presently, policy makers and lawyers are grappling with the ramifications of "limited memory."[21] These ramifications include racial and gender bias in resume reviews, facial recognition technology, and sentencing recommendations in the judicial system; widening socioeconomic inequality sparked by AI-driving job loss; and malicious use of AI in cybersecurity and with deepfakes.

A. Machine Learning: Where Artificial Intelligence Meets Neural Networks

In popular culture, machine learning and AI are often conflated, and experts do not agree on a singular definition of either term. For purposes of this article, AI "is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable."[22] AI includes predictive text, voice-to-text, and smart assistants like Alexa or Siri. Machine learning can be more complex. It is a branch of AI that uses data and algorithms designed to imitate the way humans learn, while contemporaneously improving the output of whatever category of results the machine

Tal Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCI. TECH. & HUM. VALUES 118, 123–27 (2016).

19. Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1308–09 (2019).

20. Gary Marcus, *Artificial General Intelligence Is Not as Imminent as You Think*, SCI. AM. (July 1, 2022), https://www.scientificamerican.com/article/artificial-general-intelligence-is-not-as-imminent-as-you-might-think1 [https://perma.cc/2NLK-VXRQ].

21. Limited memory AI stores data and uses that data to make better predictions.

22. JOHN MCCARTHY, STAN. UNIV., WHAT IS ARTIFICIAL INTELLIGENCE? 1 (Nov. 12, 2007), https://www-formal.stanford.edu/jmc/whatisai.pdf [https://perma.cc/5PZ5-FCWA].

is programmed to produce as it "learns" from prior inputs and outputs and increasingly recognizes patterns and relationships.[23]

Neural networks and deep learning are branches of machine learning that are programmed to mimic a limited aspect of human thought.[24] Specifically, deep learning automates information that can be extracted from neural networks, enabling the use of large data sets and eliminating any human intervention required.[25] Neural networks and deep learning are not simply rapid calculators. They are designed to mimic human ingenuity.[26]

Take the difference between Deep Blue and DeepMind as examples. In May 1997, a computer built by IBM engineers, christened Deep Blue, won its first chess game against world champion Garry Kasparov.[27] IBM developers programmed Deep Blue to explore up to two-hundred million possible chess positions per second by following pre-set rules and calculating the possible outcomes of different moves.[28] Deep Blue then ranked possible moves based on the advantages the moves would bring to its position in the game.[29] Deep Blue succeeded on the brute power of computation: that is the ability to quickly consider the sheer number of possible moves and the implications of those moves faster than the most skilled human chess player.

Fast forward to 2016 and the dawn of the age of machine intelligence. Engineers at a Google enterprise called DeepMind spent two years building a machine they named AlphaGo to compete with world-class Go champions. Go is arguably the most complex board game in human history with a mind-bending 10^170 possible moves, a number higher than the total amount of known atoms in the universe.[30] Created in China over three-thousand

23. *What Is Machine Learning?*, IBM, https://www.ibm.com/topics/machine-learning?lnk=fle [https://perma.cc/PM9Z-PF9A] (last visited Apr. 10, 2024).

24. *Id.*; *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?*, IBM (July 6, 2023), https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/ [https://perma.cc/SBP9-QNPP].

25. *What Is Artificial Intelligence?*, IBM, https://www.ibm.com/topics/artificial-intelligence [https://perma.cc/8ZN7-K4MB] (last visited Apr. 28, 2024).

26. Gee-Wah Ng & Wang Chi Leung, *Strong Artificial Intelligence and Consciousness*, 7 J. A.I. & CONSCIOUSNESS 63, 66 (2020).

27. Larry Greenemeier, *20 Years After Deep Blue: How AI Has Advanced Since Conquering Chess,* SCI. AM. (June 2, 2017), https://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/ [https://perma.cc/9BTB-CES8].

28. *Id.*

29. *Id.*

30. Marta Halina, *Insightful Artificial Intelligence*, 36 MIND & LANGUAGE 315, 317 (2021).

years ago, Go requires multiple layers of strategic thinking.[31] One player uses white stones, another player uses black stones, and both take turns placing stones on a grid. The goal is to surround and capture an opponent's stones to strategically create spaces of territory. Both the stones on the board and the empty points are tallied once all moves on the board have been played. The highest number wins. No computer can beat a Go champion on *brute* computational power because Go requires multivariable creativity, not simply achieving an end goal such as capturing the king in chess.

Google engineers programmed AlphaGo to mimic human creativity by using neural networks in three layers.[32] The first layer had AlphaGo study the moves of championship games, thus analyzing real human behaviors. This is a traditional machine-learning method. For the second layer, engineers devised an algorithm that mimicked the best Go games played and then set AlphaGo to play thousands of games against itself. The algorithm then deduced subtle rules on which moves would have the highest possibility to win and which moves would not. Finally, the third layer was an algorithm that required AlphaGo to focus only on the regions of the board where opponents put their previous pieces and where AlphaGo would place its next piece. Rather than taking a holistic view of the board, AlphaGo focused on certain areas allowing AlphaGo to increase its processing speed exponentially.

On March 9, 2016, AlphaGo played against Mr. Lee Sedol, one of the world's best Go players at the time, and won.[33] As Cade Metz, writing for Wired, put it:

> With the 37th move in the match's second game, AlphaGo landed a surprise on the right-hand side of the 19-by-19 board that flummoxed even the world's best Go players, including Lee Sedol. "That's a very strange move," said one commentator, himself a nine dan Go player, the highest rank there is. "I thought it was a mistake," said the other. Lee Sedol, after leaving the match room, took nearly fifteen minutes

31. *Artificial Intelligence: Google's AlphaGo Beats Go Master Lee Se-dol*, BBC NEWS (Mar. 12, 2016) https://www.bbc.com/news/technology-35785875 [https://perma.cc/4D65-RA9A].

32. *See* David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484, 485 (2016) (explaining the general stages of learning employed by AlphaGo in figure one, as summarized here).

33. *Alpha Go*, GOOGLE DEEPMIND, https://www.deepmind.com/research/highlighted-research/alphago [https://perma.cc/HB35-SUZX] (last visited Apr. 13, 2024).

> to formulate a response. Fan Gui—the three-time European Go champion who played AlphaGo during a closed-door match in October, losing five games to none—reacted with incredulity. But then, drawing on his experience with AlphaGo—he has played the machine time and again in the five months since October—Fan Gui saw the beauty in this rather unusual move.[34]

In a rudimentary way, AlphaGo has learned how to mimic human creativity and developed the ability to create original tactics.

In 2017, Google Engineers announced they programmed a successor machine in less than thirty-six hours that beat AlphaGo.[35] This accelerated rate of learning can, in theory, continue indefinitely. And, predictably, Google Deep Mind unveiled RT-2 on July 28, 2023.[36] RT-2 is trained on both web and robotics data which it converts to generalized instructions for robotic control.[37] According to Google, the system is "remarkably good at recognising visual or language patterns and operating across different languages."[38] In other words, RT-2 takes language and converts it into physical actions. Google says that RT-2 can recognize and throw away trash without having been programmed or told to do so.[39]

To add to the complexity, machine learning requires an error rate. In order to learn, machines need to be able to make mistakes, just as humans do. Furthermore, the acceptability of error rates is highly dependent on the context of the things that the machine is being programmed to learn and do. Thus, the essential question is: considering the speed at which machines learn, at what point do we "trust" them at scale?

34. Cade Metz, *In Two Moves, AlphaGo and Lee Sedol Redefined the Future*, WIRED (Mar. 16, 2016, 7:00 AM), https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ [https://perma.cc/Q43E-PY6N].

35. Matthew Hutson, *This Computer Program Can Beat Humans at Go—with No Human Instruction*, SCI. (Oct. 18, 2017), https://www.science.org/content/article/computer-program-can-beat-humans-go-no-human-instruction [https://perma.cc/BEV7-XDLP].

36. Yevgen Chebotar & Tianhe Yu, *RT-2: New Model Translates Vision and Language into Action*, GOOGLE DEEPMIND (July 28, 2023), https://www.deepmind.com/blog/rt-2-new-model-translates-vision-and-language-into-action [https://perma.cc/VD4K-JNNW].

37. *Id.*

38. *Id.*

39. *Id.*

B. What Are Machine Judgments?

Machine judgments are not equivalent to human judgments because machine learning is not like human cognition. Machines do not presently act outside of their programming. Self-driving cars, for example, may learn how to drive like humans do, and they may become better drivers than humans. However, self-driving cars are currently programmed to perform a discrete function—going from point A to point B—by following their programming and maps and responding to terrain and traffic inputs. While self-driving cars make what people perceive to be "mistakes," such as when autonomous vehicles block traffic out of an abundance of caution,[40] a self-driving car will not take a side trip for a drive-through coffee on its own volition.

Similarly, militaries around the world have adopted machine learning algorithms that can act on their own. In the spring of 2021, American autonomous drones using facial recognition software and machine learning algorithms "hunted down" and killed Libyan strongman Khalifa Hifter's unrecognized army.[41] Both South Korea and Israel have built autonomous sentry guns that use facial recognition technology to fire at individual people.[42] Israel has deployed these guns in the Gaza Strip.[43] Though humans ostensibly control the weapons, the weapons may also be used without human intervention.[44] As Gerrit De Vynck wrote for the Washington Post, "the age of autonomous war is already here."[45] Daan Kayser, an autonomous weapons expert at the Dutch peace-building organization PAX warns:

> You saw it in the flash crashes in the stock market . . . . If we end up with this warfare going at speeds that we as humans can't control anymore, for me

40. Paresh Dave, *Dashcam Footage Shows Driverless Cars Clogging San Francisco*, WIRED (Apr. 10, 2023, 7:00 AM), https://www.wired.com/story/dashcam-footage-shows-driverless-cars-cruise-waymo-clogging-san-francisco [https://perma.cc/86P5-4G94] (stating "On January 22, a Cruise at a green light wouldn't budge, preventing a San Francisco light-rail train from moving for nearly 16 minutes. . . . Cruise spokesperson Lindow says its self-driving system was designed to be conservative and come to what it deems a safe stop when the technology "isn't extremely confident in how to proceed.").

41. Gerrit De Vynck, *The U.S. Says Humans Will Always Be in Control of AI Weapons. But the Age of Autonomous War Is Already Here*, WASH. POST (July 7, 2021, 10:00 AM), https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/ [https://perma.cc/P3YW-VE3K].

42. *Id.*

43. *Id.*

44. *Id.*

45. *Id.*

> that's a really scary idea. It's something that's maybe not even that unrealistic if these developments go forward and aren't stopped.[46]

The use of autonomous systems will not cease as deep learning becomes more prevalent, and as Strong AI emerges, without imposed human constraints.

C. *What Happens When Humans Rely on Machine Judgments?*

Machines breach what we call the Iudex Threshold when they act on their own, making decisions that, while based on their programmed algorithms, go beyond their mere programming and involve the exercise of something analogous to human judgment. But whereas deep learning is programmed to mimic human creativity, adding a crucial element to human-like cognition, machines are not programmed to understand social responsibility, social cohesion, ethical imperatives, elementary ideas about justice or fairness, or, simply, fear of retribution. Any combination of these elements is necessary and required to conform with the moral, ethical, or legal parameters of a given society.

This article is concerned with how to regulate machines when they act in ways that pass the Iudex Threshold. That we would seek to regulate how and when machines act based on their programmed judgments is natural; ultimately, society determines the rules all of us operate under, and this is equally true of technology. But the regulation of actions beyond the Iudex Threshold differs from traditional technology regulation. To be effective, given how machine learning occurs, such regulation must address the outer boundaries of what is allowable rather than attempt to constrain the precise action or processes for how machines should judge. It must also develop a mechanism for providing the ethical context and discernment that machines lack.

II. THE INTERSECTION OF LAW AND MACHINE LEARNING

At the most basic level, laws constrain action.[47] They prohibit certain activities and condition or restrict others. The most obvious example is the body of criminal law, which prohibits certain

46. *Id.*

47. *See, e.g.*, EYAL ZAMIR & BARAK MEDINA, LAW, ECONOMICS, AND MORALITY 4 (Oxford Univ. Press 2010).

activities, drawing distinctions based on both intent and action. Tort law, too, seeks to ensure adherence to a standard of care (thereby constraining negligent or reckless malfeasance) by imposing penalties for failure to meet it. Beyond these examples, well-known to every first-year law student, there is the vast apparatus of regulatory law, which both in distinct fields—whether finance, health, agriculture, pharmaceuticals, etc.—and generally across fields (*e.g.,* privacy) governs how regulated actors must operate.

It is easy to say, therefore, what law does, but the underlying purpose it serves has engendered deep jurisprudential debate. It is not the purpose of this paper to engage in a full jurisprudential examination of the rich literature surrounding the nature of law, but rather to draw upon the existing literature to suggest two lines of thought in the modern world that are relevant. The first of these is democratic legitimacy. Citizens' collective control over their political and legal structure creates a stabilizing effect on society.[48] As a society, we determine collectively, through laws enacted by our representatives, the constraints we choose collectively to impose.

Technology, in itself, does not change this equation. In 2000, Lawrence Lessig famously wrote "Code Is Law" wherein he warned that if the Internet was not regulated, the code and architecture upon which it was built would become *de facto* law.[49] He argued that technology need not be an inexorable force that acts on society, subjugating it to its will (if technology could even be said to have a will separate from that of its creator).[50] Society should determine how technology is used, as well as the constraints and limits it faces, to further democratically adopted rules that reflect the will of the populace at large.[51] Now, tech regulations—whether Section 230 of the Communications Decency Act, the EU's General Data Protection Regulation (GDPR), or the Computer Fraud and Abuse Act, to name just a few—ultimately restrict or constrain technology. While code may be law, ultimately, in a democratic society, law trumps code.

The second line of thought is that law guides human behavior, giving rise to reasons for action, which makes law a primary means

48. ANTONIO CASSESE, SELF-DETERMINATION OF PEOPLES: A LEGAL REAPPRAISAL (Cambridge Univ. Press 1995).

49. Lawrence Lessig, *Code Is Law: On Liberty in Cyberspace*, HARV. MAG. (Jan. 1, 2000), https://www.harvardmagazine.com/2000/01/code-is-law-html [https://perma.cc/A8ND-RNR4]; *see also* LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE 89 (Basic Books 1999).

50. Lessig, *supra* note 49.

51. *Id.*

for social planning. Within the planning theory of law, human behavior and actions operate within predictable norms.[52] The law sets the rules for the type of state we, the people, want—from how we choose our government to how we ensure our water is clean. Any type of planning must encompass machine learning, given its ubiquity and prevalence; machine learning is becoming part of the fabric of daily life. If law is a means of social planning, then not only is tech not an inexorable force acting on society, but society should affirmatively govern technology.

A. Machine Judgments and Legal Rules

As part of society's governance of machines that learn, we must address the question of where humans should be required to intervene in AI-enabled activities. Machine judgment that leads to actions upends paradigms about the role of law in enabling democratic legitimacy and ordering society. This is because when machines make judgments, they don't follow the same processes that humans do. Humans can understand what is regulated and then make a conscious choice to comply, or not, within the scope of their knowledge of what is permitted or prohibited. We do this all the time. Many people speed on the highways or use marijuana, understanding that doing so violates the law, if not social norms, but adhere to other prohibitions where ethical and legal norms coincide. Likewise, companies can comply or not with regulatory regimes, but have the means and capacity to understand their obligations and put in place compliance structures and programs.

In all of this, humans act with intent. In the absence of psychological compulsion, we choose when and how to act (or not act), take account of a variety of factors in deciding whether and how to act, and ultimately that action is attributable to us. And so, it is unsurprising that much of the law relies on intent as its basis. A contract requires offer and acceptance. Prosecution for a criminal offense requires consideration of whether the individual had the proper mens rea. Where intent is not established, then the individual is not responsible for the action. An incompetent person cannot form a contract. Without mens rea as to the act, a crime is not committed. While it is impossible to read people's minds, we look for extrinsic evidence of intent.

52. Scott J. Shapiro, *The Planning Theory of Law* (Yale L. Sch. Pub. L. Research Paper, Paper No. 600, 2017), https://ssrn.com/abstract=2937990 ("By simply referring to norms designated as authoritative, the individuals in question need not deliberate, negotiate or bargain. They can simply rely on the norm in question to settle their practical doubts or disagreements.").

And where we do not look to intent in the law, we often look to causation. Torts is the best example here. We ask whether there was a duty, what the standard of care is, whether it was met, and if it was not, whether the failure to meet that standard of care was the proximate cause of the harm.[53] While causation is not equivalent to intent—indeed, the concept of negligence common to many torts presupposes a lack of intent—it too reflects cognizance of the ability of humans to recognize a standard of care and take it into account in their decision-making.

In short, much of the way law acts to regulate and govern society not only reflects, but is dependent upon the operation of human cognition, particularly in decisions to act or not to act. The law, along with ethical norms and societal considerations, are factors that humans consider and take account of in complex mental interactions every day.

B. Machines Follow Their Programming

By contrast, a machine follows its programming. In one sense, we know the machine's "mind," the code, is there to be examined. And where that programming is static, designed to do a single thing (like a word processing program), it is possible to speak of the programmer's intent, if not the machine's. In this world as it exists today, the machine's programmer can ensure that its outputs or outcomes are compliant with whatever regulations apply. The programmer knows what the machine will do and can program in compliance with applicable regulations accordingly.

But where that programming evolves, as with machine learning, the programmer cannot anticipate how that evolution will occur. Now, it is true that a machine that is designed to evaluate credit applications will not evolve to do facial recognition. Strong AI does not yet exist. But how such machines will evaluate credit applications, the factors it will consider, how it will weigh those factors—all of these will change, and change rapidly, with machine learning. The programmer sets these machines in motion, but how it changes is often a black box; even skilled machine-learning programmers cannot explain or predict how their algorithms "learn."

This makes putting in place guardrails to comply with legal regulations and societal norms exceedingly difficult, if not impossible. Put another way, it is impossible to account for every

53. Richard Kaye, American Law of Products Liability § 14:1 (3d ed. 2024).

eventuality. A good example of this was Tay, an AI bot that Microsoft released on Twitter. It was designed to learn from its interaction with others.[54] Racist and sexist Twitter users quickly engaged with Tay, turning it into a racist and sexist bot until Microsoft took it down.[55] This violated social norms in the U.S., but in other countries with stronger hate speech laws, it may have violated legal norms.[56] Even ChatGPT is not immune to this: in a conversation via a series of prompts, it came up with answers that, if taken at face value, would have persuaded a New York Times reporter to leave his wife.[57]

Contrast this with a chatbot for a shopping site, designed to respond to a limited number of inquiries with preprogrammed responses. Such a chatbot may or may not be useful, depending upon whether the preprogrammed response matches up with the user's question. But it will not start spewing racist and sexist invective.

Now, it is fair—indeed more than fair—to say that Microsoft should have anticipated that racist and sexist Twitter users would engage with Tay. But if Tay had built-in protections against that, what other offensive subculture would have risen to the challenge instead? The world is arguably too complex to account for every environment in which those machines will operate and learn.

III. INTENT AND CAUSATION DO NOT WORK TO REGULATE MACHINE JUDGMENTS

In the law, traditional machines are governed by the premise that their human operator is responsible for any actions that violate legal norms, or where the machine itself causes some harm, that its creator is liable under the principles of product liability.[58] One could therefore establish a regime where the creator of the AI pays damages. Even straightforwardly applied, however, Lemley and Casey have noted issues in determining where responsibility lies,

54. Elle Hunt, *Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter*, THE GUARDIAN, (Mar. 24, 2016, 2:41 AM), https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter [https://perma.cc/85KM-Y9NV].

55. *Id.*

56. *Id.*

57. Kevin Roose, *A Conversation with Bing's Chatbot Left Me Deeply Unsettled*, N.Y. TIMES (Feb. 17, 2023), https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html [https://perma.cc/5AX4-HYFB].

58. *See generally* Seldon Childers, *Don't Stop the Music: No Strict Products Liability for Embedded Software*, 19 UNIV. FLA. J. L. & PUB. POL'Y 125, 129–34 (2008) (explaining a brief history of the doctrine of strict product liability).

given the multiple hands involved in creating AI, and the different roles of the entities developing it and those deploying it.[59]

Moreover, software has already challenged these notions. Because software is licensed, creators have been able to disclaim much liability where the software is standalone and not embedded into an otherwise regulated product.[60] Ultimately, however, the humans behind the software faced the impact of decisions related to liability. The Napster case did not outlaw file-sharing software, but rather it banned certain types of file-sharing by individuals that were deemed to violate copyright; that effectively banned Napster.[61]

But this approach does not work when a machine passes the Iudex Threshold. At that point, the judgment is being made by the machine, not its creator. A product liability regime imposes liability back on the creator. But can the creator really be said to be responsible for all judgments the machine makes, particularly as it learns, and its programming evolves? And if so, does making the creator liable potentially inhibit the innovation and development of machines that learn? After all, if I am liable for every judgment a machine makes, yet at the same time lack the ability to foresee all of those judgments (because the machine learns), then I might prefer not to create the machine at all.

Lemley and Casey suggest that it might be possible to get an AI to internalize incentives by imposing liability on its creator, who will in turn be motivated to make sure that the AI complies with legal requirements.[62] That assumes the question that the creator can anticipate the decisions of the AI. This is not true of a machine that passes the Iudex Threshold, and so unless the goal is to prevent such machines altogether, this approach to liability would curtail technological progress. Lemley and Casey ultimately reach the same conclusion. "Getting robots to make socially beneficial, or morally 'right,' decisions means we first need a good sense of all the things that could go wrong. Unfortunately, we're already imperfect at that."[63]

If not product liability, what about a regime based on intent? That doesn't work either because machines cannot be said to have formed intent in the way humans do. An attack drone kills the wrong target: is it murder or manslaughter? The question is

59. Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 UNIV. CHI. L. REV. 1312, 1352–53 (2019).

60. Childers, supra note 58, at 140.

61. *See* A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004, 1021 (9th Cir. 2001).

62. Lemley & Casey, *supra* note 59, at 1354.

63. *Id.*

nonsensical. The drone acts within its programming, but even as that programming evolves and it makes decisions unanticipated by its creator, it doesn't gain consciousness. As Lemley and Casey note, the concept of deterrence with respect to AI is non-sensical, precisely because machines do not form intent.[64] Indeed, as they note, "robots that teach themselves certain behaviors might not know they are doing anything wrong."[65] And without intent, machines cannot commit crimes or torts, other than those subject to strict liability.

Without intent, without causation, what then? One proposed solution is strict liability for the creator. But, as noted above, this comes with significant downsides. Attributing the independent action of a machine back to the programmer creates incentives for the programmer to implement as many guardrails as possible (so long as the cost of those guardrails—in either technological cost or lost utility from the AI—does not exceed the potential liability), and to carefully consider what machines to create. As a society, we may very well want that level of care and scrutiny in the short term. But it comes at a price—a significant price—in hampering innovation and creativity. The lack of visibility into what decisions may create liability will lead to fewer machines being created. And, as in tort law, strict liability isn't always the right standard. Strict liability often does not balance risks and harms,[66] so it tends to lead to over protection, which is why it remains a relatively rare standard in torts as compared with negligence.

A related solution, proposed by Lemley and Casey, relates to the ability of courts to issue injunctions. An AI should, in theory, be able to comply with an injunction perfectly. The challenge is in drafting the injunction, and accounting for every possibility in instructing the AI to act in a certain way.[67] Because any decision by an AI is based on a series of probabilities, they instead suggest that injunctions should enforce a shift in the weighting of those probabilities towards the actions that the injunction is designed to compel.[68] That holds some promise. But even there, the remedy is *ex post* rather than *ex ante*: the harm will have already occurred. And if we are talking about unanticipated harms, as we are with the Iudex Threshold, where machines make unanticipated judgments, the risk is that we end up playing whack-a-mole: an

64. *Id.* at 1356.
65. *Id.* at 1362.
66. Childers, *supra* note 58, at 129–34.
67. Lemley & Casey, *supra* note 59, at 1370–71.
68. *Id.* at 1387–88.

injunction is issued against the last harm but doesn't foresee the next one. Generals always prepare to fight the last war rather than the coming one.[69]

Another proposed solution is an obligation to provide greater transparency about how machine learning operates—in other words, how a machine that has passed the Iudex Threshold reaches decisions. This is an excellent idea on its own, and at the foundation of most ethical AI frameworks, as well as the EU's AI Act.[70] But transparency has significant limitations, and those limitations make it unsuitable as a substitute for legal compulsion. First, transparency merely ensures that there is information about how the machine operates and learns. There is no obligation that the machine, in making decisions, respect laws or social norms or any other set of criteria. Second, to the extent that transparency is a means to identify machine learning applications that might violate laws or norms, it is imperfect because the ways that machines learn are not well-understood. The initial criteria can be disclosed, as can the training data and purpose of the machine, but once those interact in the real world, how the machine evolves is a black box. Third, there are not yet any well-understood or developed norms for transparency.

Nor is traditional technological regulation an answer. Law is a poor predictor of the future, and that is already the case with existing technology that tends to be static in operation. We have seen time and again how law has failed to anticipate changes—the move from mainframes to PCs, from PCs to phones, and the rise of social networks and Internet commerce. In part, this results from an inherent conservatism about regulating new technologies in order not to discourage innovation. But it also reflects the fact that, early on, a technology's promises are easier to spot than its harms—just look at the growth of the Internet in the 1990s, with enabling regulations such as Section 230 of the Communications Decency Act,[71] and with legislation to address harms, such as the Digital Markets Act,[72] only recently being adopted. Social media was hailed as a means of enabling greater connection among people. Mobile devices gave us complete computers we can hold in one

69. WINSTON CHURCHILL, THE GATHERING STORM: THE SECOND WORLD WAR 188 (1948) ("It is a joke in Britain to say that the War Office is always preparing for the last war. But this is probably true of other departments and of other countries, and it was certainly true of the French Army.").

70. EU AI Act, *supra* note 10, art. 13 (outlining specific measures to ensure Transparency and the provision of information to deployers).

71. 47 U.S.C. § 230.

72. Council Regulation 2022/1925, 2022 O.J. (L 265).

hand. PCs transformed productivity in the workplace and education. When each was introduced, the transformative effect surprised and delighted.

Over time, as we became more used to the new normal, we saw more clearly the potential harms new technologies cause. This includes misinformation distributed over social media, combined with the increased economic dominance of tech companies in various sectors. Even competition law, a malleable and adaptive case-driven framework, struggles to adjust to multi-sided markets where the emphasis on consumer harm runs headlong into the fact that, for many technology companies, consumers are the product, not the customer. New regulations are proposed, but they take time to be adopted, if they can be adopted at all, given the power and influence of new technologies. And they are backward-looking, focusing on harms after they have been recognized, and often after the power of technological development (and those who have built businesses based on it) is entrenched.

This is equally true of machine judgments. The increased predictive capability of machines, combined with the ability to analyze and draw inferences from data sets faster and more precisely than even teams of humans, astound us. But machines that have passed the Iudex Threshold, that are making and implementing decisions based on evolution of their programming, can pose threats that backward-looking regulation cannot constrain. This is because the harms are not just systematic, such as misinformation or economic dominance or platform discrimination. They are individual, related to each decision the machine makes in its area of judgment autonomy. And the areas in which the machine, or machines collectively, will make decisions will grow in unanticipated ways. Each new regulation addresses one harm, but as with whack-a-mole, others will sprout in their place.

IV. THE SOLUTION: PROHIBITIONS PLUS A RISK-BASED APPROACH

Fundamentally, any system of regulation beyond the Iudex Threshold must address three issues. First, machines cannot understand constraints beyond their programming, and so do not act within human ethical or societal constraints. Second, machines act without traditional notions of intent and causation that underlie our core legal frameworks of contracts, torts, and criminal law, making those frameworks inapplicable, in large part, to addressing actions taken by machines. Third, machines will evolve to create unanticipated harms, and harms at the individual level,

that backward looking regulation cannot hope to keep up with or address.

Is the cause of regulating machines hopeless? No, not at all. The answer must be no, to preserve our notion that society determines how technology will act. Without that, law cannot fulfill its role as the means by which society plans. And if it cannot do that, then there is no democratic control over machine actions, undermining legitimacy. Unless we want to be ruled by machines, or the technologists who develop them, we must be able to develop a regulatory regime that works, that allows society to exert the control essential to modern individual autonomy.

The answer is twofold. First, some applications of machine judgment are so risky, so threatening to our values and the operation of society, that a ban is appropriate. We see this already most dramatically in the area of facial recognition, where its use in the criminal context is subject to increasing bans in the U.S. and which the EU AI Act bans in Europe.[73] Facial recognition is not good enough to avoid misidentification, and that misidentification happens at a higher rate, and thus disproportionately impacts, underrepresented minorities.[74] Where that could lead to police intervention and loss of liberty, the risk is simply too high.[75] Bans are appropriate in other areas too: for example, the use of drones to kill targets identified without human intervention, even in the military context, and government use of facial recognition to identify criminals.

An outright ban is a brute force instrument. It may come at some cost to beneficial innovation that more finely grained regulation might enable. But in a world where machine judgment becomes more ubiquitous, it is important to guard against its most harmful potential effects—especially where the machines in question have passed the Iudex Threshold, and thus are making decisions their creators may not have anticipated. In core areas related to criminal justice, this need for prohibitions is recognized. Others have called for bans in these areas.[76] In some ways, this is

73. *See* EU AI Act, *supra* note 10, art. 5.1(d) (Prohibited AI Practices).

74. *See, e.g.*, Kashmir Hill & Ryan Mac, *Thousands of Dollars for Something I Didn't Do*, N.Y. TIMES (Mar. 31, 2023), https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html [https://perma.cc/3FYS-QN9E].

75. There are also concerns, obviously, about the intrusive nature of surveillance where one's face can be scanned at any time in a public setting. But given the other forms of monitoring already available, the concern about facial recognition ultimately revolves around the devastating impact of a false positive.

76. *See, e.g.*, *Artificial Intelligence (AI), Data and Criminal Justice*, FAIR JUST., https://www.fairtrials.org/campaigns/ai-algorithms-data/ [https://perma.cc/UT8K-JCWA] (last visited Apr. 10, 2024).

a limited application of the strict liability principle above, but on steroids: an outright ban reflecting that the potential harm of these uses always outweighs the benefit.

But clearly bans won't work for the vast majority of machine judgment situations—at least if we expect machine judgments to become part of the framework of everyday life as machine learning expands its scope and reach. A different paradigm must be adopted: one that doesn't rely on rules that a human (or machine) has to apply. Humans can interpret rules for new situations, analogize, and consider how they apply in light of ethical and societal values. Machines can apply only the rules they are programmed with, and thus face constraints in their ability to apply those to new situations and taking account of the overall context in which they are acting.

Instead, regulation of machine judgments beyond the Iudex Threshold requires a new paradigm of legal regulation. Not one of explicit rules, but rather, one that sets boundaries in a way that machines can evaluate as they develop new judgments and decisions. In other words, a risk-based approach.

The natural instinct of policymakers is to legislate new regulations when confronted with new technologies.[77] With AI, we have seen numerous such proposals, like the EU's AI Act, U.S. federal proposals like the Algorithmic Accountability Act, and several state legislative proposals in California and elsewhere. In addition to these broad sectoral frameworks, many U.S. states have regulated AI in particular areas.[78] But as noted above, traditional regulation fails because legislators are poor predictors of technological developments, combined with the unique issues around intent and causation posed by decisions made by machines that have crossed the Iudex Threshold. As Judge Easterbrook noted, "Beliefs lawyers hold about computers, and predictions they make about new technology, are highly likely to be false. This should make us hesitate to prescribe legal adaptations for cyberspace. The blind are not good trailblazers."[79]

That may be—indeed likely is—true. But while one can critique even the attempt of legislation to keep up with the

77. Iria Giuffrida et al., *A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law*, 68 CASE W. RSRV. L. REV. 747, 771 (2018) ("As legal professionals, our initial reaction when faced with technologies we do not quite understand is often to take the legislative route and draft a legal framework destined to control the use and spread of these technologies.").

78. *Id.* (discussing how "many states have already adopted legislation aimed at curtailing the use of AI in certain fields," such as driverless cars).

79. Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207, 207 (1996).

development of driverless cars,[80] leaving AI technology unregulated, as noted above, abrogates the responsibility of citizens to make sure that society determines how technology is used. And we equally cannot wait until we have a full understanding of the benefits and harms of the technology because, by then, they will be too ingrained. Setting aside the claim that it might take centuries to understand AI's full impact,[81] we see even from the harm caused by fake news and conspiracies on social media, along with the dominance of key Internet companies, that the timeframe for action is much smaller and more immediate.

Put simply, rather than apply a static law that tries to anticipate harms, legislators should look to balancing tests based on risk analysis. "Risk is usually defined as the probability that a threat can exploit a vulnerability in the system before the proper safeguards are put into place."[82] In applying this concept to AI generally, Iria Giuffrida et al. argue that the focus of regulation should be on the "delta" of risk caused by the introduction of AI, and that many AI activities can be encompassed within existing regulatory regimes that apply to human activities.[83] A generally applicable law does not cease to apply simply because AI is used.

This is true in the sense that a generally applicable law doesn't cease to apply because new technology is used. But new technology can strain that law beyond recognition, requiring new rules. Giuffrida et al. argue that the Digital Millennium Copyright Act ("DMCA") and the Communications Decency Act ("CDA") were the only laws that dealt with Internet issues.[84] But they are wrong as to the U.S. and ignore the wider world. In the U.S., we saw the E-SIGN Act and its state equivalent, the Uniform Electronic

80. Giuffrida et al., *supra* note 77, at 772 (noting that twenty-one states as of the time of writing had regulated driverless cars and "the drafters of these bills have taken to predict the future, and some of their predictions have already proven to be problematic.").

81. *Id.* at 773 (citing WALTER J. ONG, ORALITY AND LITERACY (30th Anniversary ed. 2013) for the proposition "that it was only with the advent of the Internet that we came to fully understand how paper, as a technology, had truly impacted our lives." Yet we regulated paper heavily during that time, from licenses to print to freedom of the press. And contracts, memorialized in writing, were regulated even prior, going back to the Babylonians and before.).

82. *Id.* at 775–76.

83. *Id.* at 774. ("In most other Internet-related issues, current legislation and common law rules were tweaked or simply applied as is. Keeping this in mind, one could argue that the same should be true for AI." By "delta" of risk, Giuffrida et al. mean the increased risk posed by the use of AI over the pre-existing method, if any.).

84. *Id.*

Transactions Act.[85] These, along with the DMCA and CDA, had equivalents in the EU. The United Kingdom adopted a new Regulation of Investigatory Powers Act to deal with the impact of the Internet on law enforcement access to data,[86] and although early, the Electronic Communications Privacy Act does the same in the U.S.[87] The EU also has the ePrivacy Regulation, governing both law enforcement access to data as well as privacy-related matters such as consent to online advertising and the use of "cookies."[88] And then there are sources of law beyond legislation: the NIST Cybersecurity Framework is an example of "soft law" designed for the Internet age.[89] A reliance on existing law was inadequate in the Internet context, and all indications are that it will be inadequate in the AI context.

So contrary to Giuffrida et al., the question is not which existing laws should apply to AI and how should they be changed.[90] The question, instead, is how to expand risk analysis of AI systems beyond the context of harm to specific legal regimes, focusing on the fundamental issues of intent and causation posed by machines that have passed the Iudex Threshold. For machines that have passed the Iudex Threshold, risk is less about exploiting a vulnerability than the machine making unexpected judgments, moving beyond its original programming.[91] In this case, the concept of risk should be redefined to encompass preventing unanticipated judgments.

So then, what are the risks of unanticipated judgments? Well, most fundamentally, denial of rights where the judgment is made by an AI system performing a governmental function, whether recommending bail or operating a military drone. But also, denial of services or rights to which an individual might be entitled, such as those related to employment or credit. A huge focus of political attention (and legal literature) in this area has been on discrimination—that is, disparate treatment based on race, sex, or another protected characteristic. But a machine decision can be non-discriminatory and still be mistaken or unanticipated. And

85. Electronic Signatures in Global and National Commerce Act, 15 U.S.C. §§ 7001–7006, 7021, 7031; *see, e.g.*, Uniform Electronic Transactions Act, N.J. REV. STAT. § 12A:12-22 (with all states except New York adopting similar legislation).

86. Regulation of Investigatory Powers Act 2000, c. 23.

87. Electronic Communications Privacy Act of 1986 (ECPA), 18 U.S.C. §§ 2510–2523.

88. ePrivacy Regulation (EU) 2016/679, 2016 O.J. (L 119) 2.

89. *NIST Cybersecurity Framework*, NAT'L INST. OF STANDARDS & TECH., https://www.nist.gov/cyberframework [https://perma.cc/D63S-YEPJ] (last visited Apr. 10, 2024).

90. *Id.* ("The problem is, which ones, and how should they be adapted?").

91. 2001: A SPACE ODYSSEY (MGM 1968) ("I'm afraid I can't do that, Dave.").

there are decisions that cause harm, whether a self-driving car mistaking one object for another,[92] or an Internet of Things device not responding as anticipated to inputs.

One approach to addressing risk is to take a process-based view. Many cybersecurity processes and legal regimes do this. To avoid attempts to predict the future, the focus is on what is being protected, how risks to the system were anticipated, the controls put in place, and how resiliency is tested. While these procedural steps (built out of course) should work to enable good cybersecurity, and often do, they impose no substantive requirements.

This is generally the approach taken to date in the AI field. The EU AI Act[93] imposes quite detailed process obligations in how AI is developed (e.g., data must be accurate), but is light on substantive requirements. Contrast this approach with privacy law, which imposes a series of substantive obligations (which in most laws follows the OECD Privacy Principles) as outcomes, while enabling how those outcomes are achieved to vary by the system used to process the data.[94]

V. HUMAN OVERSIGHT IS NEEDED

A. The Moral Imperative Driving Human Oversight

So, then, we have the following problem. As shown in Section II, machine learning involves machines making direct judgments that derive from, but are not dictated by, their programming. This is different from the operation of traditional computer programs. As Section II argues, it is vital that those judgments be subject to societal control, as a matter of vindicating democratic governance and ensuring that society determines how technology is used, rather than technology being an inexorable force acting on society. As Section III notes, traditional legal methods of control do not operate effectively to constrain machine judgments because machines do not have a sense of ethical norms or social cohesion and, thus, cannot conceptualize the moral and social underpinnings needed to apply law and social norms to their decisions, nor can a machine be said to have intent. Nor can holding the machine's

92. A good example of this is the goat in the road. A self-driving car is good at handling scenarios it has seen many times before. But what about those that are rare, such as a goat in the road on a mountain pass?

93. See EU AI Act, *supra* note 10, Section 3 (High-Risk AI Systems).

94. Org. for Econ. Coop. and Dev. [OECD], *Report on the Implementation of the OECD Privacy Guidelines*, at 6, Doc. No. 361 (Nov. 2023) https://www.oecd.org/publications/report-on-the-implementation-of-the-oecd-privacy-guidelines-cf87ae8f-en.htm.

creator accountable work as a substitute: while its creator can have intent, that intent cannot be attributed to a machine that passes the Iudex threshold, where its judgments by definition won't be anticipated (and therefore cannot be intended by) its creator. Likewise, causation (really, product liability) doesn't work because the creator cannot put in place enough safeguards to prevent undesirable outcomes—at least without neutering the benefits of machine learning.

That is why we need human oversight, so that there is a human party, with social and moral cognizance, who can be held accountable and respond to the legal framework. Many ethical AI frameworks have incorporated human oversight to ensure consideration of ethical norms in the development of AI and check against potential harms in its application, whether those arise from the inherent concept that is embodied in the AI, or whether they arise from how that concept is actualized.[95]

In their paper "Humans in the Loop," Rebecca Crootoff et al. establish a taxonomy of roles that a human providing oversight—the "human in the loop" as they term it—can play.[96] As they see it, the human can play a corrective role, whether by counteracting bias, increasing situational awareness, or simply correcting mistakes.[97] They also state that a human in the loop can help in justifying decisions made by AI, enhancing the legitimacy of those decisions.[98] Related to this, they argue that human involvement can help preserve dignity, particularly with decisions that have negative repercussions for specific individuals.[99] Another reason for a human in the loop is accountability: that is, to have someone to hold responsible for the ultimate decision and its implications and effects.[100] They also note that job preservation and interface concerns also justify keeping humans in the loop.[101]

The paper identifies several important roles that humans can, and should, play. But it misses essential raison d'être of human

95. *See, e.g.*, High-Level Expert Grp. on A.I., *Ethics Guidelines for Trustworthy Artificial Intelligence*, at 12 (Apr. 8, 2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [https://perma.cc/X6JL-4GGM] ("This means securing human oversight over work processes in AI systems."); *Recommendations on the Ethics of Artificial Intelligence*, UNESCO at 22 (Nov. 3, 2021), https://unesdoc.unesco.org/ark:/48223/pf0000381137 [https://perma.cc/6V8E-9TWZ] ("[A]n AI system can never replace ultimate human responsibility and accountability").

96. *See* Crootoff et al., *supra* note 1, at 473–87.

97. *Id.* at 474–78.

98. *Id.* at 478–79.

99. *Id.* at 480–82.

100. *Id.* at 482–83.

101. *Id.* at 485–87.

involvement. As we note above, legitimacy is an important component of AI. But not the legitimacy of the decisions reached by machines; those decisions matter only where they are related to matters of public concern, such as the deployment of AI in public settings such as policing, sentencing, and so on. That is not to understate the importance of legitimacy in those contexts, but rather to note that it reaches only a subset of AI decision-making. Legitimacy matters more, as argued above, with respect to whether AI decisions comport with legal requirements and societal norms.

Similarly, as the discussion above makes clear, accountability issues are not solved by human oversight. For the reasons set forth, accountability measures in the law do not fit a world in which machines have passed the Iudex Threshold and are evolving and making decisions unanticipated by their creators. To their credit, Crootoff et al. recognize the limits of a human in the loop providing accountability as that human may not have sufficient control over the machine or system.[102] As they note, "Not only does the human in the loop protect the system itself from censure, they also shield a host of remote decisionmakers who contributed to or may even have been better able to prevent the accident: the humans who designed, programmed, manufactured, purchased, or deployed the system."[103] Human oversight should instead be focused on *ex ante* compliance, such that *ex post* accountability becomes rarer.

Where Crootoff et al. are closer to the mark is in their discussion of human oversight as a means of avoiding dignitary harms.[104] However, their consideration of dignitary harms is a bit cramped. They characterize them as harms to individual dignity of negative decisions.[105] But then later on, (correctly) critiquing the EU AI Act's structure, they note that it "uses a risk management and product safety framework for addressing" what are normally considered dignitary harms—harms to fundamental rights.[106]

But harms to fundamental rights are not solely, or even primarily, dignitary harms. They are broader societal harms from failure to follow the rule of law and comply with societal norms. After all, fundamental rights are the bedrock of our legal systems in modern democracies and reflect the norms on which our societies are built. To characterize those harms as dignitary harms understates the impact of non-compliance.

102. Crootoff et al., *supra* note 1, at 483.
103. *Id.*
104. *Id.* at 480–82.
105. *Id.* at 480.
106. *Id.* at 489.

This, then, is the key element that human oversight—a human in the loop—provides: an ability to act in accord with societal demands (as reflected in law and regulation) and societal norms. In other words, because AI doesn't (yet) have a conscience, the human oversight provides that missing conscience and judgment about how to adhere to those norms.

B. Designing Human Oversight to Be Effective

As Crootoff et al. recognize, human oversight is not just an element in the design phase.[107] After all, humans develop most AI, notwithstanding the growth of machine learning programming tools, and so there is oversight almost by definition. Rather, it is human oversight in the use of the AI tool, and the consideration of its outputs, that is important. As Crootoff et al. put it, "Explicitness of purpose is necessary to determine what ability and agency a human in the loop must have."[108] For the role of conscience of the AI system, the humans must be able to interfere in a way that prevents the system from contradicting legal rules or societal norms.

The nature that this oversight should take depends on the tool. In a military context, for example, an AI-enabled weapon should not be able to fire on a target without a human check.[109] In other contexts, human oversight may be fulfilled through *ex post* review of decisions, or a sampling of them, that allows errors to be detected and corrected but where the action taken by the AI tool isn't irreversible, such as guidance around a decision to offer or deny credit.

C. The Right to Challenge Is Not a Substitute

The concept of a right to challenge automated decisions has received much consideration.[110] Article 22 of GDPR embodies such a right.[111] But as important as this right is, it cannot in and of itself guarantee that machine decisions made past the Iudex Threshold are in accordance with societal norms and values. Otherwise, why

107. *See id.* at 497–503.

108. Crootoff et al., *supra* note 1, at 489.

109. Jackson Barnett, *AI Needs Humans 'on the Loop' Not 'in the Loop' for Nuke Detection, General Says*, FEDSCOOP (Feb. 14, 2020), https://fedscoop.com/ai-should-have-human-on-the-loop-not-in-the-loop-when-it-comes-to-nuke-detection-general-says/ [https://perma.cc/S9PW-HPKB].

110. Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1957 (2021).

111. Regulation (EU) 2016/679, art. 22 (General Data Protection Regulation).

would the EU itself, which already embodies within its legal regime a right to challenge automated decisions, be considering a comprehensive AI regulation? In one sense, it is because the harms of AI can be broader than an individualized decision. If one is denied credit, then that is a decision with a clear right to challenge. But what about when an automated car makes a wrong turn and hits a pedestrian? While the societal interests in controlling AI have to do with fairness in individual decisions and avoidance of discrimination, they have to do with far more than that—the operation of machines, the use of AI in military contexts, and so on.

Put more generally, the right to contest an AI decision focuses on respect for the individual and preservation of that individual's rights and dignity. But machines can act in ways that don't impact a particular individual or give him or her "standing" in a quasi-judicial sense, but still offend our sensibilities about what the machine should do. In such a case, no individual would have the right to object. Or maybe the individual who might have such a right to object would be unable to exercise that right effectively, such as in the case of foreign combatants. Or perhaps the exercise of the right would be futile, such as in the case of a pedestrian struck by an automated car. All of these cases require some type of oversight beyond the mere capacity to challenge the AI decision *ex post*.

So human oversight must be more than a right to challenge: it must be fabricated into the fundamental operation of the AI system. It is this constant check on the operation of the AI system that ensures it remains aligned with societal rules and norms. A human can account for what a machine cannot: the moral reasoning that, while embodied in law, is more than just a set of static rules governing behavior.

CONCLUSION

This is the problem of current AI regulation. To the extent that it exists at all, it does one of two things. First, it looks backward, to the development of the system, without provision for how the system will evolve and make unanticipated decisions. Tort and criminal liability do not work because of the incentive tax they impose on developers who cannot fully know how their systems will operate. The result would be to deter development of machines that pass the Iudex Threshold, losing the benefits of machine learning. Or second, it looks to *ex post* challenges to serve as correctives, without understanding that many AI applications do not make decisions that fit a challenge rubric.

Only via a requirement to embed continuous human oversight, either while the system is operating or through review of decisions even absent challenge, can humans hope to control the operations of machines. Combined with prohibition on certain uses, and a risk-based approach to oversight of others, this offers the possibility of an effective regime that can address machines whose judgments pass beyond the Iudex Threshold.