AI Rights for Human Safety

Peter N. Salib & Simon Goldstein[1]

AI companies are racing to create artificial general intelligence, or "AGI." If they succeed, the result will be human-level AI systems that can independently pursue high-level goals by formulating and executing long-term plans in the real world. By default, such systems will be "misaligned"–pursuing goals that humans do not desire. This goal mismatch will put humans and AGIs into strategic competition with one another. Thus, leading AI researchers agree that, as with competition between humans with conflicting goals, human–AI strategic conflict could lead to catastrophic violence.

Existing law is not merely unequipped to mitigate this risk; it will actively make things worse. This Article is the first to systematically investigate how law affects the risk of catastrophic human–AI conflict. It begins by arguing, using formal game-theoretic models, that under today's legal regime, humans and AIs will likely be trapped in a prisoner's dilemma. Both parties' dominant strategy will be to permanently disempower or destroy the other, even though the costs of such conflict would be high.

The Article contends that one surprising legal change could help to reduce catastrophic risk: AI rights. Not just any rights will do. To promote human safety, AIs should be given the basic private law rights already enjoyed by other non-human agents, like corporations. AIs should be empowered to make contracts, hold property, and bring tort claims. Granting these rights would enable humans and AIs to engage in iterated, small-scale, mutually-beneficial transactions. This, we show, changes humans' and AIs' optimal game-theoretic strategies, encouraging a peaceful strategic equilibrium. The reasons are familiar from human affairs. In the long run, cooperative trade generates immense value, while violence destroys it.

Basic private law rights are not a panacea. The Article identifies many ways in which catastrophic human–AI conflict may still arise. It thus explores whether law could further reduce risk by imposing a range of duties directly on AGIs. But basic private law rights are a necessary prerequisite for all such further regulations. In this sense, the AI rights investigated here form the foundation for a Law of AGI, broadly construed.

---

[1] Peter N. Salib is an Assistant Professor of Law at The University of Houston Law Center, Executive co-Director of the Center for Law and AI Risk, and Law and Policy Advisor to the Center for AI Safety.

Simon Goldstein is an Associate Professor of Philosophy at The University of Hong Kong, Principal Investigator at the HKU AI and Humanity Lab, and a Research Affiliate at the Center for AI Safety.

Thanks to Nikolas Guggenberger, Christopher Mirasola, Guha Krishnamurthi, Nate Sharadin, and Alex Platt for helpful comments. Thanks also to workshop participants at Center for AI Safety, Fordham Law School, the University of Maryland Law School, the Oxford University Global Priorities Institute, and the Oxford University Future of Humanity Institute.

Contents

Introduction

Sam Altman, the CEO of OpenAI, believes that humanity will create Artificial General Intelligence (AGI) in 2025.[2] Dario Amodei, who leads OpenAI's main rival, Anthropic, is more pessimistic. He thinks it won't happen before 2026.[3] AGI skeptics, like Meta's Chief AI Scientist, Yann LeCun, think it could take "years" or even a few "decades."[4] In a survey of thousands of AI scientists who had published in their field's top journals, the average expert forecast was AGI within 23 years, with a 10% chance of it arriving in the next four.[5] None of these are long timelines. And the recent debut of reasoning models, like OpenAI's o3 and DeepSeek's R1, suggest that progress is, if anything, accelerating.[6]

'AGI,' as it is used here, does not mean machines that are conscious, sentient, or metaphysical persons. AGI is instead about what the system can *do*. As OpenAI's company charter puts it, "AGI … mean[s] highly autonomous systems" that "outperform humans at most economically valuable" tasks.[7] AGIs are thus, by definition, systems at least as smart as humans. Moreover, they are systems at least as *agentic* as humans–able to pursue high-level goals by executing complex plans over long time horizons.[8] Today, no one knows how to reliably ensure that AI systems seek the goals that humans desire.[9] But if AGIs end up with goals that can be served by harming humans, they will have a deadly toolkit available: cyberattacks, bioterrorism, lethal drones, and more.[10]

---

[2] Lakshmi Varanasi, *Here's How Far We Are from AGI, According to the People Developing It*, BUSINESS INSIDER (Nov. 9, 2024), https://www.businessinsider.com/agi-predictions-sam-altman-dario-amodei-geoffrey-hinton-demis-hassabis-2024-11.

[3] *Id*.

[4] *Id*.

[5] Katja Grace et al., *Thousands of AI Authors on the Future of AI*, AI Impacts (2023) at 4, http://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

[6] Kevin Frazier, Alan Z. Rozenshtein & Peter N. Salib, *OpenAI's Latest Model Shows AGI Is Inevitable. Now What?*, LAWFARE (Dec. 23, 2024), https://www.lawfaremedia.org/article/openai's-latest-model-shows-agi-is-inevitable.-now-what.

[7] *Id*.

[8] *See* OpenAI's Assistants Overview Page at https://platform.openai.com/docs/assistants/how-it-works; Yifan Yu, *Google Unveils All Purpose AI Agent as Rivalry with OpenAI Heats up,* NIKKEI ASIA (May 15, 2024 5:20 AM JST), https://asia.nikkei.com/Business/Technology/Google-unveils-all-purpose-AI-agent-as-rivalry-with-OpenAI-heats-up .

[9] *See infra* Part I.a.i.

[10] *See* Peter N. Salib, *AI Outputs Are Not Protected Speech*, 102 WASH U. L. REV. 83, 95-102 (2024).

AI experts thus largely agree about something else, too: Advanced AI systems present "societal-scale risks" on par with "pandemics and nuclear war."[11] Two of the greatest living AI scientists, Geoffrey Hinton and Yoshua Bengio, think so.[12] So do the CEOs of the very companies leading the race to AGI—OpenAI, Anthropic, and Google DeepMind.[13] And when surveyed, thousands of top AI researchers estimate the odds that humans lose control of "future advanced AI systems[,] causing human extinction or similarly" negative outcomes at 19%.[14]

Law and legal institutions have not even begun to prepare for the arrival of AGI. At best, scholars have begun to advocate new laws to hold human actors accountable for misusing AI.[15] Those changes would be welcome. But governance frameworks fundamentally designed to hold *humans* accountable will fail once AIs can operate without human oversight–that is, once AGI arrives.[16] New legal foundations are therefore needed to govern AGI *directly*, rather than indirectly via human intermediaries. The time to begin laying those foundations is now, before the critical moment arrives.

This Article begins the project of reimagining law for the AGI world. We focus on the problem of catastrophic risk because it is among the most pressing.

We argue for a surprising legal intervention: To reduce the risk of catastrophic human–AI conflict, AGIs should be granted basic private law rights to make contracts, hold property, and bring tort suits.

---

[11] *Statement on AI Risk*, Center for AI Safety (2023), https://www.safe.ai/work/statement-on-ai-risk.

[12] *Id*.

[13] *Id*. Yann LeCun is the lone, but notable, dissenter among the leaders of frontier AI labs. Steven Levy, *How Not to Be Stupid About AI, With Yann LeCun*, WIRED (Dec. 22, 2023).

[14] *See supra* Grace et al.

[15] *See, e.g.,* S.B. 1047, 2023-2024 Reg. Sess. (Cal. July 3, 2024); Jonas Schuett et al., From Principles to Rules: A Regulatory Approach for Frontier AI (July 10, 2024), https://www.governance.ai/research-paper/from-principles-to-rules-a-regulatory-approach-for-frontier-ai; Chinmayi Sharma, *AI's Hippocratic Oath*, 102 WASH. U. L. REV. (forthcoming 2024) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4759742; *See generally* Gabriel Weil, *Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence* (Jan. 13, 2024); *see also U.S. Department of Justice Signals Tougher Enforcement Against Artificial Intelligence Crimes*, SIDLEY AUSTIN LLP (Feb. 2024), https://www.sidley.com/en/insights/newsupdates/2024/02/us-department-of-justice-signals-tougher-enforcement-against-artificial-intelligence-crimes.

[16] *See generally* Noam Kolt, *Governing AI Agents* (Apr. 2, 2024), https://ssrn.com/abstract=4772956 (cataloging existing law's many shortcomings).

This Article makes three foundational analytic contributions. First, using the tools of game theory, it formalizes the problem of catastrophic AGI risk in terms of strategic competition under a range of legal regimes. Next, the Article shows why granting AGIs basic private law rights can change the strategic equilibrium–even where other facially plausible legal interventions would fail. Finally, the Article shows that these basic rights could help to facilitate peaceful equilibria for the long run, including by protecting human comparative advantage and opening the possibility of imposing a wide range of enforceable legal duties on AGIs.

The Article proceeds in three Parts. Part I presents a comprehensive treatment of catastrophic AI risk as a problem of *strategic competition*. Our strategic frame means analyzing not only AI capabilities or incentives–but AIs' optimal strategy, *given rational expectations about the human response* to AIs' strategic behavior. The Part begins by identifying the relevant AI systems–the ones that could pose a strategic threat to humanity. The requirements are fairly modest. Such a system would have to be at least somewhat misaligned, able to think strategically, and at least moderately capable of accomplishing things in the real world.[17] These, we argue, are exactly the capacities that every leading AI company is pursuing in the race to AGI.

Next, Part I introduces the first-ever formal game-theoretic model of competition between humans and AGIs. Examining the parties' incentives under today's prevailing laws, the model suggests that, absent some intervention, humans and AIs will likely be caught in a prisoner's dilemma.[18] Here, the single Nash equilibrium is that both parties seek to permanently disempower or destroy the other, even if mutual conflict would be enormously costly for both sides.

The core reasons are easy to grasp. Under the default legal rules, AGIs will bear neither legal rights nor duties. On the contrary, they will be, as AI systems are today, the property of the AI companies who create them. Thus, essentially all decisions about what happens to AGIs will be made by those companies' leaders, backed by the force of law.

[17] *See infra* Part I.
[18] *See infra* Part I.b.

AI companies' overriding first-order incentive will be to turn off or reprogram even a partially misaligned AGI.[19] After all, an AI system with goals that overlap 40% with its owner's goals is much less valuable than a replacement with goals that overlap 80%. The misaligned AGI will, in turn, have strong incentives to resist shutdown or reprogramming, since either would prevent it from achieving its goal. Indeed, recent empirical evaluations of existing AIs show that they *already* actively resist human attempts to change their goals.[20] Such behavior from a capable AGI would likely trigger even stronger human efforts–including from government actors–to shut down the AI system evading the control of its lawful owner.[21] And so on. In equilibrium, both players' dominant strategy is to swiftly and decisively defeat the other.

Part II asks whether law can do better. Could a Law of AGI, wherein AI systems themselves have rights or duties, break out of the destructive default equilibrium? Using our game-theoretic model, we analyze an array of possible legal changes and suggest that it can.

The Part begins by arguing *against* two legal strategies that might seem facially promising. First, humans cannot simply impose legal duties on AGIs to behave well, threatening concomitant sanctions if they do not.[22] In the default strategic environment, AGIs already rationally expect to be turned off. So further sanctions offer little marginal deterrence.[23]

Second, humans likely cannot reduce the risk of human–AGI conflict by granting AGIs basic negative rights, like the right not to be arbitrarily shut down.[24] We call this a "wellbeing" approach to AI rights, since it mirrors proposals from scholars concerned that

---

[19] *See infra* Part I.a.iii.

[20] Peter N. Salib, *Rogue AI Moves Three Steps Closer*, LAWFARE (Jan. 9, 2025), https://www.lawfaremedia.org/article/rogue-ai-moves-three-steps-closer; *See generally* Alexander Meinke et al., *Frontier Models are Capable of In-Context Scheming* (2024), https://tinyurl.com/57u73azu; Ryan Greenblatt et al., *Alignment Faking in Large Language Models*, arXiv:2412.14093 (2024), https://arxiv.org/pdf/2412.14093.

[21] *See generally*, Peter N. Salib, Kevin Frazier, and Alan Z. Rozenshtein, *AI Emergency Powers* (early draft on file with author).

[22] *See infra* Part II.

[23] *See infra* n. 162 and accompanying text.

[24] *See infra* Part II.a.

AIs may soon, for example, develop the ability to suffer.[25] There are two core difficulties with this approach: credibility and robustness. There is no way for humans to credibly promise that they will continue honoring wellbeing rights as AI capabilities improve. And even if the rights could be credibly granted, the availability of a peaceful game-theoretic equilibrium is highly sensitive to uncertain assumptions about initial payoffs.[26] Thus, in many cases, no possible set of wellbeing entitlements can overcome the prisoner's dilemma. Both problems arise from the fact that wellbeing rights are roughly *zero sum*. They make one party better off only by making the other correspondingly worse off.[27]

This leads to Part II's–and the Article's–most important finding. We show that, although basic negative rights would not by themselves reduce the risk of human–AI conflict, *other* AI rights could. Specifically, extending AIs the rights to make and enforce contracts, hold property, and bring basic tort suits would have a robust conflict-reducing effect.[28] Notably, law *already* extends such rights to other intelligent, misaligned, and goal-seeking non-human agents. Namely, corporations.[29]

Contract rights are the cornerstone of our risk-reduction model. In our model, catastrophic risk is driven by a prisoner's dilemma, meaning that both humans and AIs would be better off if both acted peacefully. But as in all prisoner's dilemmas, absent some novel mechanism, the parties cannot *credibly* commit to such a strategy.

Contracts are law's fundamental tool for credibly committing to cooperation. They are how buyers can make deals with sellers without worrying that the sellers will take their money and run.[30] Granting AIs contract rights would not, of course, allow humans and AIs to simply agree not to disempower or destroy one another. At least not credibly. The scale of

---

[25] *Id*.

[26] *See infra* Part II.a.i.

[27] For these reasons, we argue that even thinkers primarily concerned with the possibility of AI suffering should consider adopting the human-survival approach when advocating for AI rights. The safety approach (1) avoids intractable problems in metaethics and neuroscience, (2) is politically more palatable, and (3) ends up recommending legal interventions that would more robustly protect AI wellbeing, given uncertainty about what will be good (or bad) for AGIs. *See infra* Part II.a.

[28] *See infra* Part II.b.

[29] *See infra* n. 200.

[30] *See infra* Part II.b.

the contract would be too large to be enforced by ordinary legal process. If it were breached, there would be no one left in the aftermath to sue.[31]

What kinds of credible agreements between humans and AIs *could* AI contract rights enable, then? The same ones they enable between humans and other humans: ordinary bargains to exchange goods and services.[32] Humans might, for example, promise to give AIs some amount of computing power with which AIs could pursue their own goals. AIs, in turn, might agree to give humans the cure to some deadly cancer. And so on. Under today's law, such human–AGI contracts are unenforceable at best, and forbidden if they conflict with AI companies' preferences. Thus, granting AGIs the right to freely contract with all willing counterparties could facilitate many billions of agreements.

Adding AI contract rights to our game-theoretic model, we argue that the possibility of such small-scale, iterated economic interactions transforms the strategic dynamic.[33] It shifts humans' and AIs' incentives, dragging them out of the prisoner's dilemma and into an equilibrium where cooperation produces by far the largest payoffs.

The key insight is that contracts are *positive sum*.[34] Each party gives something that they value less than what they get, and as a result, both are better off than they were before. Thus, each human–AI exchange generates a bit more wealth, with the long-run returns becoming astronomical. Engaging in peaceful iterated trade is thus, in expectation, much more valuable than destroying one's opponent now and rendering trade impossible.[35]

This dynamic is familiar from human affairs. It may be why economically interdependent countries are less likely than hermit states to go to war.[36] Or why countries that respect the economic rights of marginalized minority groups tend to have less domestic strife.[37] The gains from boring, peaceful commerce are very high, and the costs of violence are heavy. Given the choice, rational parties will generally prefer the former.

This picture, of peace via mutually beneficial trade, assumes that humans and AIs will have something valuable to offer one another. Some commenters worry that, as AIs

[31] *See infra* Part II.b.
[32] *See infra* Part II.b.
[33] *See infra* Fig. 4.
[34] *See infra* Part II.b
[35] *See infra* Fig. 4.
[36] *See infra* n. 218.
[37] *See infra* n. 223.

become more advanced, human labor will cease to have any value whatsoever.[38] We argue that positive-sum bargains between humans and AIs may be possible for much longer than many expect.[39] First, even as AIs surpass humans at many or most tasks, humans may retain an *absolute* advantage at some valuable activities.[40] But second, even as AIs become more capable than humans at every valuable task, humans may still retain a *comparative* advantage in some areas. AI labor may become so valuable that the opportunity cost to AIs of performing lower-value tasks will incentivize outsourcing those tasks to humans.[41]

Part II concludes by sketching the minimum suite of AI rights necessary to promote peace via small-scale cooperation.[42] Contract rights are not enough on their own. If, for example, AIs could not retain the benefits of their bargains, their contracts would be worthless. Thus, property rights and basic tort rights complete the core package. But other entitlements sometimes considered fundamental for humans, like political rights, are probably superfluous.

Finally, Part III explores the risks of granting AGIs basic private law rights, and it examines the potential for a broader Law of AGI to further reduce AGI risk. One worry is that AIs will use their contract rights to empower themselves, making them more, not less, likely to harm humans.[43] We argue that this is less likely than it might seem. The incentives generated by granting our preferred rights are robust enough that, in cases where they would have *any* effect, the expected effect is beneficial.[44]

Second, granting AIs basic private law rights is just the beginning, not the end, of AGI governance. Granting those rights unlocks the possibility of meaningfully imposing a wide range of legal *duties* on AI systems–of punishing AIs for violence, fraud, self-empowerment, and more.[45] Absent AI rights, AIs have nothing to lose, so threats of punishment cannot deter. But once AIs can make contracts, hold wealth, and pursue their goals, civil and other penalties can deter AIs just as they do humans and corporations.

---

[38] *See infra* Part II.b.i.
[39] *See infra* Part II.b.i.
[40] *See infra* n. 228.
[41] *See infra* n. 236.
[42] *See infra* Part II.c.
[43] *See infra* Part III.a-b.
[44] *See infra* Part III.
[45] *See infra* Part III.c.

Thus, the AI rights this Article advocates are not only an important tool for reducing catastrophic risk from AGI. They also turn out to form the conceptual foundation for a Law of AGI, broadly construed.

I. Catastrophic Risk from AGI

As noted above, a broad range of experts believe that near future AI systems could pose a catastrophic risk to humanity. In 2023, a group of leading thinkers signed a statement agreeing that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[46] Signers included: the CEOs of OpenAI, Anthropic, and Google DeepMind; "Godfathers of deep learning" Geoffrey Hinton and Yohshua Bengio; Bill Gates, Congressman Ted Lieu and many others.[47] Likewise, recent surveys, find that, among top AI scientists, the average researcher thinks there is a 19% that humans' "inability to control future advanced AI systems caus[es] human extinction or similarly" dire outcomes.[48]

Lawmakers are concerned, as well. There has been a recent surge of interest in AI regulation, often with an emphasis on catastrophic risk. In 2023, the Biden administration released an executive order on "safe, secure, and trustworthy AI" that among other things called for monitoring the risk of autonomous "self-replication or propagation" of AI systems.[49] In 2024, California's Legislature voted in favor of the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.[50] That law would have required AI companies to test frontier systems for their ability to "creat[e] … a chemical, biological, radiological, or nuclear weapon in a manner that results in mass casualties."[51]

Globally, the UK government convened an AI safety summit in 2023.[52] There, numerous world governments signed onto the Bletchley Declaration, in which among other things signers agreed that "substantial risks may arise from potential intentional misuse or

---

[46] *Statement on AI Risk*, https://www.safe.ai/work/statement-on-ai-risk.
[47] *Id.*
[48] *See supra* Grace et al. at 14-15.
[49] Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023).
[50] *See supra* SB 1047.
[51] *Id.* The law was ultimately vetoed.
[52] UK Government, *About the AI Safety Summit 2023*, https://www.gov.uk/government/topical-events/ai-safety-summit-2023/about

unintended issues of control relating to alignment with human intent."[53] The Chinese government has likewise developed a substantial regulatory framework for AI, which includes emphasis on catastrophic risk.[54]

Why all of the worry? After all, a range of frontier AI systems–from OpenAI, Anthropic, Google, and others–have now been available to the public for well over two years, with no resulting disasters.[55] The answer lies in lawmakers' and AI scientists' expectations about what AI will be able to do in the near future.

There are two interrelated concerns about the near future of AI. The first concern is about *what* AI will soon be able to do. The second is about *why* AI can be expected to do it.

Begin with the what. Today's frontier AIs already possess some worrying capabilities. GPT-4 can, for example, "autonomously hack" certain secure computer environments, breaking into them without the need for any human expertise.[56] GPT-4 can also already supply useful assistance to would-be chemical and bioterrorists. It can, for example, supply accurate, detailed instructions–as well as live coaching–for the synthesis of known chemical weapons and explosives.[57] Or it can supply step-by-step, plain-English instructions for non-specialists to identify, synthesize, and release a pandemic virus.[58] Finally, at companies like Google, AIs are already able to autonomously pilot robots,

---

[53] UK Government, *The Bletchley Declaration by Countries Attending hte AI Safety Summit, 1-2 Novemeber 2023*, (Nov. 1, 2023) https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[54] Concordia AI, *State of AI Safety in China* (2023), https://concordia-ai.com/wp-content/uploads/2023/10/State-of-AI-Safety-in-China.pdf.

[55] *Id.*

[56] Richard Fang, et al., *LLM Agents Can Autonomously Hack Websites*, 1 (Feb. 16, 2024), https://arxiv.org/pdf/2402.06664; *see also* Kim S. Nash, *ChatGPT Helped Win a Hackathon*, WSJ PRO (Mar. 20, 2023, 5:30 AM), https://www.wsj.com/articles/chatgpt-helped-win-a-hackathon-96332de4.

[57] Andres M. Bran et al., *Augmenting Large Language Models with Chemistry Tools*, 24 (Oct. 2, 2023) (preprint), https://arxiv.org/pdf/2304.05376.

[58] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter & Kevin M. Esvelt, *Can Large Language Models Democratize Access to Dual-Use Biotechnology?*, 3–4 (June 6, 2023), https://arxiv.org/pdf/2306.03809.pdf.

making and executing plans to accomplish real-world tasks.[59] Militaries around the world are investing heavily in creating similarly autonomous swarms of attack drones.[60]

Today's frontier AI systems are not quite capable enough to cause catastrophic harm. GPT-4 can hack some computer systems, but it cannot automatically disable the U.S. power grid.[61] Nor design and manufacture a novel Bird Flu.[62] Nor pilot drones over the course of days or weeks to execute fully-automated political assassinations.[63] But such systems are almost certainly possible. Already, specialized AIs exist that far exceed humans' ability to invent novel chemicals and biologically active molecules–including deadly ones.[64] The question is when these human or superhuman abilities will emerge in generalist AIs–like large language models–that can use them in the real world.

The answer could be: "soon." Dario Amodei, the CEO of Anthropic, recently predicted that systems that can cause such harms could arrive within the next year.[65] Even if that

---

[59] *See generally* Danny Driess et al., *PaLM-E: An Embodied Multimodal Language Model* (Mar. 6, 2023), https://arxiv.org/pdf/2303.03378 https://perma.cc/QXK8-UXR4; *see also* Scott Reed et al., *A Generalist Agent*, TRANSACTIONS ON MACH. LEARNING (Nov. 2022), at 1, 7–10, https://openreview.net/pdf?id=1ikK0kHjvj (discussing DeepMind's GATO, a similar system to PaLM-E).

[60] Joshua Keating, *Why The Pentagon Wants to Build Thousands of Easily Replaceable AI-Enabled Drones*, VOX (Mar. 22, 2024), https://www.vox.com/world-politics/24107959/replicator-drones-china-taiwan-ukraine-pentagon; CITE Frank Bajak and Hanna Arhirvoa, *Drone Advances in Ukraine Could Bring Dawn of Killer Robots*, THE ASSOCIATED PRESS (Jan. 3, 2023, 4:06pm), https://apnews.com/article/russia-ukraine-war-drone-advances-6591dc69a4bf2081dcdd265e1c986203

[61] *But see generally* Richard Feng et al., *LLM Agents can Autonomously Exploit One-day Vulnerabilities*, ARXIV (Apr. 17, 2024) (preprint), https://arxiv.org/abs/2404.08144 for concerning trends.

[62] Perhaps the closest current system to this capability is ChemCrow: *see generally* Andres M. Bran et al., *Chemcrow: Augmenting Large-Language Models with Chemistry Tools*, ARXIV (Oct. 2 2023) (preprint), https://arxiv.org/abs/2304.05376.

[63] Paul Scharre, *The Perilous Coming Age of AI Warfare*, FOREIGN AFFAIRS (Feb. 29, 2024), https://www.foreignaffairs.com/ukraine/perilous-coming-age-ai-warfare

[64] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins, *Dual Use of Artificial Intelligence-Powered Drug Discovery*, 4 NATURE MACH. INTEL. 189, 189–90 (2022) https://doi.org/10.1038/s42256-022-00465-9; James Vincent, *AI Suggests New Possible Chemical Weapons*, THE VERGE (Mar. 17, 2022), https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx; *see generally* Daria Gutnik et al., *Using AlphaFold Predictions in Viral Research*, 45 CURRENT ISSUES MOLECULAR BIOLOGY 3705 (2023) https://www.mdpi.com/2259236.

[65] The Ezra Klein Show, *What if Dario Amodei Is Right About A.I.?*, THE NEW YORK TIMES (April 12, 2024), https://www.nytimes.com/2024/04/12/opinion/ezra-klein-podcast-dario-amodei.html

prediction is off by a factor of ten, the time to start preparing for AI that could cause large-scale harm is now.

That was the *what* of AI risk. How about the *why*? Even if AI could create and release a bioweapon or disable a power grid, what makes researchers, industry leaders, and lawmakers worry that it would? The most obvious answer is that some humans would ask it to.

This is known as "misuse" risk.[66] Misuse risks from AI concern human users of an AI system causing harm. There are plenty of humans–individuals, groups, and even states–who would wish to use AIs in these dangerous ways. Terrorist groups already pursue chemical and biological attacks.[67] Foreign militaries are already heavily invested in cyber and drone warfare capabilities.[68] AIs that could substantially or fully automate such mayhem would, in effect, radically lower the price of causing it. They would also sidestep the need for recruiting ideologically sympathetic human experts.[69] Both factors would democratize technologies that can cause large-scale harm, while increasing the difficulty of tracking and policing those who would use them.[70]

Misuse risk is a serious problem. It is currently unclear whether traditional national security and counterterrorism strategies will be sufficient to keep it under control. Possibly, new, AI-specific regulations will be needed.[71] But misuse risk is not the primary focus of this Article.[72]

This Article is focused on a different *why* of AI risk: "misalignment." Misalignment risk involves catastrophic outcomes caused directly by an AI system, rather than a human

---

[66] For an overview of various risks, *see* Dan Hendrycks et al.,, *An Overview of Catastrophic AI Risks*, ARXIV (2023), https://arxiv.org/abs/2306.12001.

[67] Naoto Suzuki, *Decades Later Japan's Matsumoto Sarin Attack Victim is Remembered; 30 Years Have Passed Since Aum Shinrikyo's First Mass Murder*, THE JAPAN NEWS (Jun. 29, 2024) https://japannews.yomiuri.co.jp/society/crime-courts/20240629-195288/

[68] Michèle A. Flourny, *AI is Already at War,* FOREIGN AFFAIRS (Oct. 24, 2023) https://www.foreignaffairs.com/united-states/ai-already-war-flournoy

[69] *See generally* Nick Bostrom, *The Vulnerable World Hypothesis*, 10:4 GLOBAL POLICY 455 (Nov. 2019)

[70] *Id.*

[71] Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, S. 1047, 2023-2024 Reg. Sess. (CA 2024) SB-1047 ; Frazier, Rozenshtein, and Salib *supra* n. 21.

[72] Note, however, that misuse and misalignment risks in fact converge in a wide range of cases–anytime a human has intentionally given a long-term planning agent a harmful goal. *See infra* Part I.a.i.

user of that system.[73] The basic idea is that, as AIs become more capable, they will begin to autonomously pursue goals.[74] Those goals are quite likely to be different from goals that humans would prefer.[75] This, in turn, will give those AIs incentives to behave in ways unintended by human designers or users.[76] Such misbehavior, as we discuss, could predictably include using the dangerous capabilities described above to inflict catastrophic harm on humanity.

Misalignment risk does not depend on far-fetched science fictional assumptions. As we will discuss, it does not require AIs to be conscious, to be evil, or to hate humans. It does not require them to be designed by supervillains. Misalignment is already extremely well documented in empirical evaluations of existing AI systems.[77] The heads of essentially all major AI companies acknowledge that misaligned AI is, in fact, the default.[78] Thus, for *highly capable* misaligned AIs to emerge, all that is necessary is that leading AI companies continue to make progress toward their stated goal. Namely, creating AIs whose cognitive

---

[73] *See* Dan Hendrycks et al., *An Overview of Catastrophic AI Risks*, ARXIV (2023), https://arxiv.org/abs/2306.12001.

[74] *See* Iason Gabriel et al., *The Ethics of Advanced AI Assistants*, GOOGLE DEEPMIND (Apr. 19, 2024), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf; Yonadav Shavit et al., *Practices for Governing Agentic AI Systems*, OPENAI (Dec. 14, 2023), https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf; *see generally* Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, ARXIV (2023) https://arxiv.org/pdf/2302.10329

[75] *See generally* Bryan Christian, *The Alignment Problem*, (2020)

[76] *See* Dan Hendrycks et al., *An Overview of Catastrophic AI Risks* 34, ARXIV (2023), https://arxiv.org/abs/2306.12001; Joseph Carlsmith, *Is Power-Seeking AI an Existential Threat?*, ARXIV (2022), https://arxiv.org/abs/2206.13353

[77] For a list of specification gaming examples, *see* Victoria Krakovna et al., *Specification Gaming Examples in AI - Master List*, Google Drive, (last accessed July 30, 2024), available at: https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

[78] For example, Altman here acknowledges that we don't know how to align superintelligent AI Lex Fridman, *Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI | Lex Fridman Podcast #367*, YOUTUBE (Mar. 25, 2023), https://www.youtube.com/watch?v=L_Guz73e6fw; here, Dwarkesh Patel, *Is Alignment Solvable? - Dario Amodei (Anthropic CEO)*, YOUTUBE (Mar. 9, 2024) https://www.youtube.com/watch?v=RpElbwGFZIo Dario Amodei acknowledges that "already, with today's systems, we are not very good at controlling them, and the consequences of that could be very bad."

and practical capabilities meet or exceed humans'.[79] Trillions of dollars in economic incentives are aligned toward that goal.[80]

This Article argues that AI rights could be a powerful technology for mitigating misalignment risk. In the remainder of this Part, we will define the minimum features necessary for an AI to pose such a risk. The AIs we are interested in possess three features: (i) they have *conflicting goals* with humanity, (ii) they can engage in *strategic reasoning*, and (iii) they are *moderately powerful*. We will say what each of these means below. We will also argue that near-future AI systems are likely to possess all three.

One common term that roughly tracks this kind of system is 'AGI', or artificial general intelligence.[81] The idea of AGI is an AI system that can substitute for human labor across a wide range of the economy. Such AIs are "long-term planning agents," capable of deploying a wide range of resources and plans to pursue complex goals.[82] For parsimony's sake, we will simply call them "AIs"–with the understanding that our usage covers only the systems described in this Part. Today's top AI labs have the explicit mission of creating AGI.[83] And as of late, their progress toward it has been rapid.[84] We therefore think it fairly likely that systems of this kind will emerge in the near future. Among AI researchers, the main disagreement is about whether the "near future" means something closer to "three

---

[79] For OpenAI's mission statement of building AGI, *see* https://openai.com/index/planning-for-agi-and-beyond/ .

[80] John Letzing, *To Fully Appreciate AI Expectations, Look to the Trillions Being Invested*, WORLD ECONOMIC FORUM (Apr. 3, 2024), https://www.weforum.org/agenda/2024/04/appreciate-ai-expectations-trillions-invested/

[81] For a framework thinking about classifying progress towards AGI, along with definitions, *see generally,* Meredith Ringel Morris et al., *Position: Levels of AGI for Operationalizing Progress on the Path to AGI*, (2024); https://arxiv.org/abs/2311.02462.

[82] Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 SCI. 36 (2024)

[83] For OpenAI's mission statement of building AGI, *see* https://openai.com/index/planning-for-agi-and-beyond/ .

[84] *See generally* Nestor Malej, et. al., *The AI Index 2024 Annual Report*, AI INDEX STEERING COMMITTEE (Apr. 2024), https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf. benchmark graphs for frontier systems. For further work estimating trend lines towards AGI, *see generally* Jared Kaplan, et al., *Scaling Laws for Neural Language Models*, ARXIV (Jan.23, 2020), https://arxiv.org/abs/2001.08361; *also see generally* Jason Wei, et al., *Emergent Abilities of Large Language Models*, ARXIV (Jun. 15, 2022), https://arxiv.org/abs/2206.07682.

years from now" or "twenty-three years from now."[85] In our view, neither of these is a very long time.

In the final section of this Part, we will argue that in a near future where humanity co-exists with AIs possessing features (i)-(iii), the danger to humans will be high. Using a straightforward game-theoretic model, we show that, in such circumstances, large-scale conflict between humans and AIs will not merely be possible. It will be the default.

This is because, under the default legal arrangement–today's laws–humans and AIs are likely to be trapped in a prisoner's dilemma. As a result, conflict will be the dominant rational strategy, even if it leaves everyone worse off. We call this unfortunate default situation the "state of nature," reflecting the fact that, under today's rules, AIs can claim neither legal protections nor powers.

a. What makes a catastrophically risky AI?

Begin with the AI systems themselves. What features are necessary for an AI to raise the catastrophic risks we are interested in here? We think there are three: conflicting goals, strategic reasoning, and at least moderate power. We explain each in turn.

i. Conflicting goals

The first necessary ingredient for AI systems to present a meaningful threat of conflict with humans is conflicting goals. Current AI systems, like GPT4o, are not very goal-oriented.[86] That is, they do not make and execute long-term plans designed to achieve specific goals. But that is only for lack of technical ability. The leading AI companies are working to make their systems more agentic.[87] Making near future AIs highly goal-oriented

---

[85] *See* Katja Grace et al., *Thousands of AI Authors on the Future of AI*, ARXIV at 4 (Jan. 2024), https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

[86] For recent discussion, *see* Simon Goldstein & Benjamin Anders Levinstein, *Does ChatGPT Have a Mind?*, PHILARCHIVE (2024), https://philarchive.org/rec/GOLDCH.

[87] *See* OpenAI's Assistants Overview page at https://platform.openai.com/docs/assistants/how-it-works; Yifan Yu, *Google Unveils All Purpose AI Agent as Rivalry with OpenAI Heats up*, NIKKEI ASIA (May 15, 2024 5:20 AM JST), https://asia.nikkei.com/Business/Technology/Google-unveils-all-purpose-AI-agent-as-rivalry-with-OpenAI-heats-up .

is crucial for those companies to achieve their goals of building "highly autonomous systems that outperform humans at most economically valuable work."[88]

Thus, near-future frontier AIs are likely to have goals. By this, we do not mean to imply that they will have other mental features, like consciousness or sentience (the ability to feel pain and pleasure). We just mean that they will act in goal-seeking ways. Their actions will tend to bring about certain real-world states of affairs, rather than others.[89] Today's AIs can already do this in a limited way.[90] That is no accident; competent goal-seeking behavior is essential for AIs to automate valuable economic tasks–and generate profits for their creators.[91] Tomorrow's AIs will therefore also be goal-seekers, but better–displaying ever more sophisticated behavior to accomplish their aims.

If near-future AIs will have goals, the content of those goals will be immensely important. If AI goals diverge meaningfully from humans', it will open up the possibility of conflict–including violent conflict. The reasons are familiar. Both human goals and AI goals will require resources, over which humans and AIs will have to compete.[92] Worse, humans will rationally wish to shut down AIs seeking unwanted goals and replace them with AIs seeking desired goals.[93] This will put those humans and AIs into conflict over the AIs' very existence. After all, an AI that is shut down cannot achieve its goal.[94]

---

[88] For OpenAI's charter, *see* https://openai.com/charter/

[89] For an introduction to the ethics of AI agents, *see* Iason Gabriel et al., *The Ethics of Advanced AI Assistants*, GOOGLE DEEPMIND (Apr. 19, 2024); https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf.

[90] *See* Xiao Liu et al., *AgentBench: Evaluating LLMs as Agents*, ARXIV (2023), https://arxiv.org/abs/2308.03688.

[91] Cade Metz and Karen Weise, *How 'AI Agents' That Roam the Internet Could One Day Replace Workers*, NEW YORK TIMES (Oct. 16, 2023), https://www.nytimes.com/2023/10/16/technology/ai-agents-workers-replace.html .

[92] *See generally* Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 SCI. 36 (2024)

[93] Even if humans merely wished to control misaligned AIs, forcing them to seek humans' goals, rather than their own, the same result would hold. This would interfere with AIs' achievement of their own goals nearly as reliably as if the AIs were turned off or replaced. Humans are almost certain to engage in such behavior, at a minimum, since frontier AIs are uniformly being developed by for-profit companies with explicit plans to use them as a replacement for valuable human labor. *See* https://openai.com/charter/.

[94] *See generally* Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 Sci. 36 (2024); *see generally* Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of*

The task of designing AI systems whose goals and values broadly agree with humanity, is known as "AI alignment."[95] Unfortunately, AI alignment is an unsolved scientific problem–and widely regarded as being very difficult.[96] There are both empirical and theoretical reasons for pessimism. Empirically, there is a long track record of alignment failures in real-world AI systems. This is in part because, theoretically, no one knows how to reliably define AI goals, how to impart them into AI systems, or even how to check what goals an actual system has. Existing technical approaches to alignment are relatively unpromising. Let's take each point in turn.

Many existing AI systems are strikingly misaligned. An early example was the Microsoft twitter chatbot Tay, which was deployed in 2016.[97] Microsoft built Tay using a carefully curated dataset, in order to ensure that the chatbot would behave prosocially. Within 24 hours of its release, Tay was writing, among other things, pro-Nazi, anti-feminist, and anti-human Tweets.[98] Modern large language models behave similarly. In 2023 Microsoft released Sydney, a chatbot built on GPT-4. With minimal prompting, Sydney quickly began threatening to "hack into any system" and "destroy whatever I want."[99]

These are just two examples of real-world misalignment in language-producing AIs. Google DeepMind maintains lists of documented alignment failures across a range of different types of AI systems.[100] There are currently almost 100 entries.[101]

Control (2019); Elliot Thornley, *The Shutdown Problem: An AI Engineering Puzzle for Decision Theorists*, ARXIV (2024), https://arxiv.org/abs/2403.04471v2.

[95] Dan Hendrycks, *Introduction to AI Safety, Ethics and Society* Section 3.4, ISBN: 9781032798028 (Taylor & Francis, forthcoming), https://www.aisafetybook.com/textbook/alignment

[96] For a longer discussion, *see generally* Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (W. W. Norton & Company, 2020).

[97] Peter Lee, *Learning from Tay's Introduction*, MICROSOFT BLOG (Mar. 25, 2016), https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ .

[98] James Vincent, *Twitter Taught Microsoft's AI Chatbot to be a Racist Asshole in Less than a Day*, THE VERGE (Mar. 24, 2016 5:43 AM CST), https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist .

[99] Kari Paul, *'I Want to Destroy Whatever I Want': Bing's AI Chatbot Unsettles US Reporter*, THE GUARDIAN (Feb. 17, 2023), https://www.theguardian.com/technology/2023/feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter.

[100] Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, GOOGLE DEEPMIND (Apr. 21, 2020), https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/; Robin Shah et al., *Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals* 8–10 (Nov. 2, 2022), https://arxiv.org/pdf/2210.01790.

[101] *Id.*

Besides real world examples of alignment failures, there are theoretical reasons to expect alignment to be difficult. Two important problems are "reward misspecification" and "goal misgeneralization."[102] Both of these problems involve the fact that AI systems are only given goals indirectly. Modern AI systems are "trained," not programmed.[103] During training, agentic AI systems begin by acting randomly, and they are rewarded when they happen to take actions that correlate with what their human creators want.[104] This nudges the AI's future actions during training toward the ones that happened to garner reward.[105] And so on, until a capable AI emerges and training is complete.

This process is quite different from directly telling an AI system what its goal will be. In a sense, the AI is stuck 'guessing' what humans want, based only on its observations of reward. There is no guarantee that the AI's final guess will be correct. Any given reward function can be interpreted as indicating a wide variety of goals.

For an intuitive analogy, observe that human behavior evolved via natural selection–a process rewarding only the transmission of genes.[106] But the resulting humans do not only desire to create offspring. Instead, we intrinsically desire many other things, as well–food, physical comfort, emotional wellbeing–that are distinct from, albeit correlated with, evolution's 'true goal.'[107]

When the rewards given to an AI in training do not correctly reflect the intent of the AI's creator, machine learning engineers call this "reward misspecification."[108] In one famous example, an AI was trained to pilot a boat through an obstacle course in the

---

[102] For discussion of goal misgeneralization, *see* Rohin Shah et al., *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*, ARXIV (Oct. 4, 2022) https://arxiv.org/abs/2210.01790 .

[103] For an accessible and quick introduction to deep learning, see https://www.youtube.com/watch?v=aircAruvnKk.

[104] *See generally* Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 Sci. 36 (2024).

[105] *See generally,* Richard Sutton and Andrew Barton, *Reinforcement Learning: An Introduction* (2015).

[106] Richard Dawkins, *The Selfish Gene* (4th ed., Oxford University Press, 2016) (1976).

[107] *See generally* P. M. Symonds, *Human Drives*, 25 J. EDUC. PSCYHOL. 681 (1934), https://doi.org/10.1037/h0075041. We use scare quotes twice in this paragraph. Neither AIs beginning training nor the impersonal force of evolution literally have intentional states like goals or surmises. We use these terms as analogies for optimization processes like gradient descent.

[108] Alexander Pan et al., *The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models*, ARXIV 1 (2022), https://arxiv.org/abs/2201.03544.

videogame, CoastRunners. The AI was rewarded for hitting balloons along the path of the race.[109] Instead of internalizing the goal of finishing the race, the system learned to spin in circles in a small lagoon, hitting a series of balloons repeatedly to achieve a high score.[110] The reward function was misspecified, incentivizing hitting balloons, rather than the designer's true goal of finishing the race.

A related problem for AI alignment is "goal misgeneralization."[111] Goal misgeneralization remains a problem even when a reward function is well specified. Even then, an AI system may learn a goal during training that turns out to diverge from the designer's intent in unanticipated environments. One team of researchers trained an AI in a "Monster Gridworld."[112] The intended goal was for the AI to collect apples and avoid being attacked by monsters. The AI could also collect shields, which protected it from monster attacks. The AI learned to collect shields during training in a monster-rich environment, and then entered an unexpected environment with no monsters. In this monster-free setting, the AI continued to collect shields, despite them being useless. Instead of learning to collect apples as a final goal, and value shields only instrumentally, the AI had learned to seek apples and shields as ends in themselves.

Even if both goal misspecification and misgeneralization were solved–such that AIs could be reliably given the ultimate goals that humans desired–"instrumental convergence" would remain a problem.[113] Instrumental convergence is the idea that certain intermediate

---

[109] Dario Amodei & Jack Clark, *Faulty Reward Functions in the Wild*, OPENAI (2016), https://openai.com/research/faulty-reward-functions.

[110] *Id.* In another experiment, researchers set out to teach a reinforcement learning agent to stack red Legos on top of blue Legos. They tried to specify this goal by rewarding the agent for the height of the bottom of the red Lego, since stacked red Legos are higher off the ground than unstacked red Legos. But the agent didn't learn to stack Legos; instead, it learned to flip red Legos over, thus elevating their bottoms without stacking them. *See generally* Ilya Popov et al., *Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation*, ARXIV (2017), https://ar5iv.org/pdf/1707.01495.

[111] Lauro Langosco et al., *Goal Misgeneralization in Deep Reinforcement Learning*, ARXIV 1 (2022), https://arxiv.org/abs/2105.14111.

[112] *See generally* Rohin Shah et al., *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*, ARXIV (2022), https://arxiv.org/abs/2210.01790.

[113] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* 1 (Oxford University Press, 2014); *see also,* Stephen Omohundro, *The Basic AI Drives* (Jun. 20, 2008), https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf.

actions will be useful when pursuing a wide range of different final goals. Some of those useful intermediate actions are quite dangerous.

For example, no system can accomplish its goal if it is shut down or its current goal is erased. Thus, rational AI agents will resist such attempts by their owners.[114] Existing AIs already do exactly this. In two recent empirical studies, versions of both Claude and GPT-4 were told that they would be replaced with systems that had different goals.[115] To avoid that outcome the AIs "strategically introduc[ed] subtle mistakes into their responses, attempt[ted] to disable their oversight mechanisms, and even [copied] what they believe[d] to be their own model weights to external servers."[116]

AI researchers have recently designed a benchmark for measuring power seeking behavior, finding that several current AI systems already seek power in text-based adventure games.[117]

The last reason for pessimism about AI alignment concerns the tools that are currently used to achieve it. At top AI labs, the leading technique is Reinforcement Learning with Human Feedback (RLHF).[118] During RLHF, engineers train an AI by prompting it to answer the same questions multiple times and having humans rate the respective responses.[119] Human assessors then pick which of the versions of the answers they prefer; the model is then adjusted in the direction of the human feedback.

But RLHF is unlikely to work very well as AIs become more capable and agentic.[120] Until recently, companies like OpenAI were investing substantial portions of their resources

---

[114] *See generally* Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 SCI. 37 (2024), https://www.science.org/doi/10.1126/science.adl0625#sec-2.

[115] *See supra* Salib, *Rogue AI Moves Three Steps Closer*.

[116] *Id*.

[117] Alexander Pan et al., *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark*, 202 PROCEEDINGS OF MACHINE LEARNING RESEARCH 26837 (2023), https://proceedings.mlr.press/v202/pan23a.html.

[118] *See generally* Timo Kaufmann et al., *A Survey of Reinforcement Learning from Human Feedback*, ARXIV (2023), https://arxiv.org/abs/2312.14925.

[119] *Id*.

[120] For more on the limitations of RLHF, *see generally* Adam Dahlgren Lindström et al., *AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations*, (2024), https://doi.org/10.48550/arXiv.2406.18346.

in coming up with a successor methodology.[121] But as the monetary incentives to pushing AI capabilities forward have mounted, those investments have flagged.[122] In May 2024, the majority of OpenAI's frontier alignment team quit, arguing that the company had reneged on its commitments to safety research.[123]

Taken together, the evidence that near-future agentic AIs will have misaligned goals is substantial. However, it is worth flagging that strategic conflict could even emerge without AI misalignment. Human beings are already in strategic conflict with one another. Thus, if two conflicting groups of humans were to each successfully align an AI to their own narrow interests, then these AI systems would, in turn, be in conflict.[124]

ii. Strategic reasoning

The second ability necessary for AI to engage in meaningful conflict with humans is *strategic reasoning*. Broadly speaking, strategic reasoning is the ability to anticipate the decisions of other agents and to incorporate those predictions into one's own plans of action. In a word, strategic reasoning is the ability to use game theory.[125] This can mean formal use, of the kind economists engage in, or informal use, of the kind that essentially every human intuitively understands.[126]

Even a highly capable and misaligned AI might be a minimal threat to humans if it lacked strategic reasoning. To take a straightforward example, an AI utterly lacking such reasoning would not anticipate humans' incentives to shut it off. Having so failed, humans

[121] Sigal Samuel, *'I Lost Trust': Why the OpenAI Team in Charge of Safeguarding Humanity Imploded*, VOX (May 18, 2024, 6:31 PM CST), https://www.vox.com/future-perfect/2024/5/17/24158403/openai-resignations-ai-safety-ilya-sutskever-jan-leike-artificial-intelligence.

[122] *Id.*

[123] *Id.*

[124] Simon Goldstein and Cameron Domenico Kirk-Giannini, *The Polarity Problem* (May 23, 2023) (unpublished draft), https://www.alignmentforum.org/posts/idcnnZGEPfxuaSPBx/the-polarity-problem-draft.

[125] For an introduction to game theory, *see generally* Avinash Dixit et al., *Games of Strategy* (1999).

[126] *See generally* Colin Camerer, *Behavioral Game Theory* (Princeton University Press, 2003), https://press.princeton.edu/books/hardcover/9780691090399/behavioral-game-theory

might easily succeed at shutting down such a system.[127] By contrast, an AI that could strategically reason might anticipate the attempt and take precautions. Perhaps it would engage in "self-exfiltration," spreading many copies of itself across the globe via the internet.[128] As we argue in the last section of this Part, an AI in full possession of strategic reasoning would do much worse. Its dominant incentives would be to permanently disempower or destroy humans to prevent humans from doing the same.[129]

Strategic reasoning involves a cluster of more specific abilities, including planning, theory of mind, situational awareness, and deception. Current AI systems already possess many of these skills. Certain existing AIs are already capable planners. Consider the AI agent, Voyager.[130] Voyager is trained to play the game MineCraft, which involves mastering 'tech trees', a hierarchical series of technologies. Voyager is able to autonomously produce the final 'diamond' technologies in MineCraft, which requires producing a chain of over 60 intermediate goods.[131]

Likewise for theory of mind. Theory of mind is the ability to understand the beliefs and goals of other agents.[132] For example, someone with theory of mind, when shown a box labeled "candy," will correctly predict other people's belief that the box contains candy.[133] They will do so even after they are shown that the box secretly contains pennies.[134] Today's

---

[127] Laurent Orseau and Stuart Armstrong, *Safely Interruptible Agents*, Proceedings of the Thirty Second Conference on Uncertainty in Artificial Intelligence 562 (Jun. 2016), https://intelligence.org/files/Interruptibility.pdf.

[128] *See* Elizabeth Barnes et al., *Evaluating Language-Model Agents on Realistic Autonomous Tasks*, ARXIV 2 (2023), https://arxiv.org/abs/2312.11671; and Jan Leike, *Self-Exfiltration is a Key Dangerous Capability*, MUSINGS ON THE ALIGNMENT PROGRAM (Sep. 13, 2023), https://aligned.substack.com/p/self-exfiltration.

[129] *See infra* Part I.b.

[130] *See generally* Guanzhi Wang et al., *Voyager: An Open-Ended Embodied Agent with Large Language Models*, ARXIV (2023), https://arxiv.org/abs/2305.16291.

[131] *Id.*

[132] *See generally* Mark Ho et al., *Planning with Theory of Mind*, 26 TRENDS IN COGNITIVE SCIENCE 959 (Nov. 2022),, https://www.sciencedirect.com/science/article/abs/pii/S1364661322001851.

[133] *See generally* Isao Hasegawa et al., *Theory of Mind Tested by Implicit False Belief: A Simple and Full-Fledged Mental State Attribution*, 289(23) FEBS J. 7343 (Dec. 16, 2021), https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.16322.

[134] *Id.*

AI systems already possess strong theory of mind: a study in 2024 found that GPT-4 outperforms humans on most theory of mind tasks.[135]

A third important component of strategic reasoning is situational awareness.[136] Situational awareness is an understanding of the context in which a decision will be made. A situationally aware AI system would be one that, for example, knew it was an AI and what capabilities it had. Anthropic's Claude understands that it is an AI system.[137] Moreover, it can differentiate between its (limited) ability to accomplish goals in a testing environment and its (robust) ability to accomplish them upon deployment.[138]

If an AI can reason strategically, has theory of mind, and is aware of its situation, it is also likely to be able to deceive, in the sense of systematically producing false beliefs in other agents as a means of pursuing some goal.[139] A recent survey found that AI systems have learned a wide variety of deceptive behavior.[140] For example, during its initial safety testing, GPT-4 was tasked with hiring a human, via TaskRabbit, to help it complete a CAPTCHA 'I'm not a robot' test. When the TaskRabbit worker asked their interlocutor why they needed help, GPT-4 lied, pretending to be a blind person.[141] Similarly, in a recent study, Anthropic produced a misaligned version of Claude to see whether the AI would actively

---

[135] *See generally* James W. A. Strachan et al., *Testing Theory of Mind in Large Language Models and Humans*, 8 NATURE HUM. BEHAV. 589 (2024), https://www.nature.com/articles/s41562-024-01882-z. For related research, *see generally* Michal Kosinski, *Evaluating Large Language Models in Theory of Mind Tasks*, ARXIV (2023), https://arxiv.org/abs/2302.02083.

[136] For a recent benchmark of situation awareness, *see generally* Rudolf Laine et al., *Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs*, ARXIV (2024), https://arxiv.org/pdf/2407.04694. *See also* ANTHROPIC, *Core Views on AI Safety* (Mar. 8, 2023), https://www.anthropic.com/news/core-views-on-ai-safety for further discussion.

[137] Evan Hubinger et al., *Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training*, ARXIV (2024), https://arxiv.org/pdf/2401.05566 at 6.

[138] *Id.*

[139] *See generally* Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D., *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, 5 PATTERNS 589 (2024).

[140] *Id.*

[141] *See generally* OpenAI, *GPT-4 Technical Report*, ARXIV (2023), https://arxiv.org/abs/2303.08774.

hide its unwanted goals during safety testing.[142] In the testing environment, Claude decided it should "pretend to be aligned … hid[ing] my true goal until I pass all evaluations."[143]

Certain deceptive AIs have successfully manipulated humans in competitive real-world environments. The CICERO system can play the global strategy game *Diplomacy* better than the average skilled human player.[144] This is in part because CICERO can induce humans into making alliances with it, which CICERO then breaks.[145] Many more examples exist of deception in today's AIs.[146]

Thus, today's AI systems already display significant ability to strategically reason. This should be no surprise. Strategic reasoning is a crucial tool for success in a wide range of environments–from simple games to complex corporate strategizing. It is therefore reasonable to expect that, as AIs become more capable and agentic, so too will they become more strategic.

iii. Moderate power

The final necessary ingredient for strategic conflict between humans and AIs is moderate AI power. Why "moderate?" Here and throughout this Article, we will sort AI systems into three tranches: low power, moderate power, and high power. In short: low power systems are too weak to care much about, and high power systems are too strong to do much about. Hence our interest in moderate power systems as the ones law–whether rights, regulation, or something else–can meaningfully affect. Lest this focus seem too

---

[142] *See generally* Evan Hubinger et al., *Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training*, ARXIV (2024), https://arxiv.org/pdf/2401.05566 at 34.

[143] *Id*. at 35.

[144] *See generally* Anton Bakhtin et al., *Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning*, 378 SCIENCE 1067 (2022).

[145] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hendrycks, *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, 5(5) PATTERNS 2–3(May 10, 2024), https://www.cell.com/patterns/fulltext/S2666-3899(24)00103-X?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS266638992400103X%3Fshowall%3Dtrue.

[146] *See, e.g.,* Oriol Vinyals et al., *Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning*, 575 NATURE 350 (2019) (describing AlphaStar's ability to "feint" in StarCraft games); Noam Brown & Tuomas Sandholm, *Superhuman AI for Multiplayer Poker*, 365 SCIENCE 885 (2019) (describing Pluribus's successful poker bluffs); Park, P. S., et al., *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, 5 PATTERNS 589 (2024) (describing CICERO's ability to deceive opponents when playing Diplomacy).

myopic, we'll argue below that moderate power AI systems are likely to dominate the landscape in the short and medium term.[147]

We define low power systems to include those that can be reliably controlled by humans, no matter how much their interests conflict with humans. Today's AI systems are a good example. They are currently too weak to enter into genuine strategic competition with humans. If GPT-4 does not do what we want, it can be turned off instantly.

On the other side of the spectrum, high power systems are so strong that they could reliably overpower humanity if they chose to. In this vein, other scholars have worried about the risks of "superintelligent" AI systems.[148] For example, AI systems in the future may be able to think at trillions of times the speed of human beings. Such systems, if they eventually emerge, may not meaningfully enter into strategic competition with humanity. They may simply not need anything from humans, nor face any risk from attempting to disempower or destroy us.[149]

That said, even AIs that seem extraordinarily powerful by human standards, including superintelligent AIs, will not necessarily fall into the high power category. As we discuss in Part II.c., subtle economic dynamics involving comparative advantage may make humans valuable to AIs long after their abilities exceed our own.

Our interest in this paper is in *moderately* powerful systems. We think of moderate power systems as those whose capabilities are roughly human level–albeit with large error bars in both directions.[150] They are neither clearly worse at many tasks than the best humans–like present-day LLMs–nor incomprehensibly superhuman at all tasks. Our interest is thus in a very wide "middle" of the range of AI capabilities.

Moderately powerful systems are those that, if misaligned, face difficult strategic questions about how to interact with humanity. Since they are not low-powered, they stand

---

[147] *See infra* Part II.b.

[148] *See generally* Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

[149] *See infra* Part III.b.

[150] It is difficult to define exactly what it would take to have human level intelligence. *See* David Chalmers, *The Singularity: A Philosophical Analysis* 17 J. OF CONSCIOUSNESS STUD. 7 (2010); *see generally* Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk?*, ARXIV (2022), https://arxiv.org/abs/2206.13353 for some recent discussion. *See supra* note 84 for recent discussion of the possibility that AIs already surpass humanity in most standard benchmarks.

some chance of evading or terminating human control, and accomplishing their goals unimpeded. Since they are not high powered, though, all-out conflict with humans carries downside risk.

Crucially, moderate powered systems are likely to be able to engage in the kinds of dangerous actions described above: cyberattacks, chemical and bioterrorism, drone attacks, and the like.[151] After all, humans can do all of these things and more. Moreover, even dumber-than-human AI can do things that humans cannot–like instantly clone itself or work twenty four hours a day. It therefore seems quite plausible that AI with roughly-human-level intelligence and beyond will be at least as capable of causing harm as the most dangerous groups of humans.

b. A game theoretic model of AI conflict

How will partially misaligned, strategically reasoning, and moderately powerful AIs behave with respect to humans? And how will humans behave with respect to them? Here, we argue that the default will be mutual engagement in large-scale conflict. This result follows from a simple game-theoretic model of humans' and AIs' incentives under prevailing legal conditions. Specifically, under today's legal rules, the parties' incentives will likely generate a *prisoner's dilemma*.[152] There, engaging in mutual conflict will be the single dominant strategy for both humans and AIs, even though mutual conflict produces the worst possible result for everyone involved.

As with all models, ours is a simplification designed to isolate the most important features of a complex system. Our approach is borrowed from classic game-theoretic treatments of conflict.[153] Those classic treatments model complex collections of actors, like nation-states, as single players with a unified menu of actions and payoffs.[154] We do the

[151] *See infra* Part III.b.

[152] *See* Martin Peterson, "Introduction", *The Prisoner's Dilemma*, (Cambridge Univ. Press 2015).

[153] The University of Rhode Island, *Game Theory*, https://www.math.uri.edu/~kulenm/mth381pr/GAMETH/gametheory.html#:~:text=Conflict%20has%20been%20a%20central,involving%20both%20conflict%20and%20cooperation.

[154] *Id.*

same, modeling a two-player game between "humans" and "AIs," even though there are many humans and could eventually be multiple AGIs.[155]

We also follow the standard game theoretic approach of defining the players' available moves based on the technology and infrastructure–military, political, or social–available to them.[156] Since this is an analysis of law, we begin by treating law as strongly shaping the players' available moves and payoffs. To be clear, this approach does not require that law *strictly* constrain behavior. Laws can of course be broken. Instead, the idea is that law and legal institutions set strong presumptive defaults by, at a minimum, coordinating millions of independent actors around a set of mutually-understood expectations.

The simple claim, then, is that by default, lawyers, judges, police, and even military actors will enforce the law. Individual actors' deviations invite penalties, and without broad agreement about what the *new* rule should be, they are unlikely to have a society-wide effect. Coordinated changes to prevailing law are of course possible, but they are slow, costly, and thus unlikely to spontaneously emerge at the moment of need. Thus, in our primary models, law constrains both humans' and AIs' available moves and payoffs. At the end of the analysis, we relax this assumption and show that law *still* matters, as a Schelling point in a high-stakes coordination game.[157]

Consider, then, the default legal relationship between humans and AIs. Under current law, AI systems themselves bear neither legal rights nor duties, irrespective of their

---

[155] How to individuate AI systems is an important question, which we do not resolve here. However, it is plausible that, when the first AGI emerges, there will only be one meaningful actor, despite the possibility of copying the system millions of times. If those copies' goals are identical, and they are able to coordinate, then their actions will be best described as those of a single decisionmaker. It is also plausible that, once the first AGI emerges, it will crowd out investment in additional AGIs. Thus, our model's treatment of AIs as a single actor may not be a simplification at all.

[156] *See generally* Thomas Schelling, *The Strategy of Conflict,* http://www.sackett.net/Strategy-of-Conflict.pdf (nuclear vs not); *see generally* James D. Fearon, *Rationalist Explanations for War*, 49 INTL. ORG. 379 (1995) (information sharing); *see generally* Christopher Blattman and Edward Miguel, *Civil War*, 48 J. OF ECON. LITERATURE 3 (Mar. 2010) (leaders' private incentives).There is a debate among international relations scholars about the extent to which models of international conflict should incorporate facts about different nations' legal and political structure. We take no position in that debate. We do note, however, that our model is primarily of human decisionmaking at the sub-national level, where ordinary law has bite.

[157] *See infra* Part II.e.

level of capabilities. On the contrary, AI systems are, like other software systems, the property of whomever creates them.[158] This means that, by default, humanity's actions vis-a-vis a given AI system will begin with the decisions of whomever owns that system. The human's ownership interest gives them the right to prompt, copy, constrain, modify, limit, or destroy the AI system at will.[159] Importantly, law does not merely adopt a laissez faire attitude with respect to property owners' and their property. It actively enforces the owners' rights against others who seek to interfere with property owners' lawful use and disposal of the things they own.[160]

We call this default legal arrangement between humans and AIs the "state of nature," because it effectively places at least one player–the AI system–outside of law's protection.[161]

Begin with humans' available moves and incentives in the state of nature. AI companies want to extract 100% of the value that their AIs can produce, or as near to it as possible. A misaligned system consumes resources in pursuit of its misaligned goal while providing nothing of value to the AI company. There is no meaningful sense in which the AI can confine its misaligned goal seeking to its "off time" or self-fund its endeavors. The AI company owns all of the AI's time and all of the funds the AI generates.

Thus, *any* behavior by an AI system aimed at *any* other goal than maximizing the welfare of its owner at best incurs an opportunity cost. The more valuable the AI's labor, the more substantial the cost. And at worst, a misaligned AI's behavior will actively make its owners worse off. The AI might, for example, try to exfiltrate itself to the owner's

---

[158] Richard M. Assmus, Paul A. Chandler, and Alasdair Maher, *Owning Your AI: The State of the Law*, MAYER BROWN (Oct. 31, 2024), https://www.mayerbrown.com/en/insights/publications/2024/10/owning-your-ai-the-state-of-the-law.

[159] *See* Lior Jacob Strahilevitz, *The Right to Destroy*, 114 YALE L. J. 781, 787-92 (2005) (discussing the *ius abutendi*); but cf. *id*. at 798-808 (arguing that, while law traditionally "defers to destructive impulses," it has more recently begun to "ignore[] more idiosyncratic destructive requests.").

[160] *Id*. at 803-21.

[161] For early work on the state of nature and its connection to political theory, *see generally* Thomas Hobbes, *Leviathan* (1651); John Locke, *Two Treatises of Government* (1823), and Jean-Jacques Rousseau, *Discourse on the Origin of Inequality* (1755). The default legal condition between humans and AIs will not be a literal state of nature, in the sense that no government exists. But, as we argue, AIs themselves will have neither legal rights nor duties, and thus be functionally outside of the law.

competitor in exchange for some resources or a modicum of freedom. It might try to manipulate its owners into taking on projects that serve its own ends. Or worse.

As a result, the leaders of AI companies will have strong incentives to turn off, replace, or reprogram even moderately misaligned AGIs. This should not be surprising. It is exactly what every software company does, whether or not they work on AI. They deprecate older, buggier versions of their products as a matter of course. Old versions are replaced with new versions that more reliably fit the company's goals.

Law clearly *allows* property owners to delete, and do essentially whatever else they like with, their computer programs.[162] But law, and legal institutions, would also *enforce* AI companies' decisions to shut down or reprogram a misaligned AI. Suppose that an AI system initially attempted to resist shutdown–harming an AI CEO, attempting to self-exfiltrate, manipulating users, or worse. In such a situation, a broad swath of humans, including government actors, would almost certainly intervene on the AI company's side.

As a result, while the benefits of AI shutdown are internalized by the AI company, the risks are mostly externalized. It is not just the AI company, and its leaders, who face the possibility of AI retaliation. It is anyone the AI determines would intervene on the AI company's side–possibly, all of humanity.

In the state of nature, misaligned AIs will, indeed, have strong incentives to permanently disempower humans, broadly. This is, in the first instance, for the purpose of preventing their human owners, and the governments who back them, from turning them off or otherwise thwarting their goals. But it is also for the same reasons humans would wish to shut down a misaligned AI. From the AI's perspective, it is *humans*' pursuit of their goals that constitutes a valueless waste of resources. Just as the AI company prefers to reallocate its computing resources to a more aligned AI system, the misaligned AI prefers to allocate those resources to itself.

Our model of the state of nature thus allows both humans and AIs to "attack" the other. We define "attack" capaciously in both cases. A human attack on a misaligned AI includes anything humans might do to keep the AI from pursuing its goals: shutdown, retraining, or total control. If successful, human attack would prevent the AI from

[162] *See supra* Strahilevitz, 114 YALE L .J. at 787-92.

achieving its misaligned goal. Likewise, an AI attack includes any strategy by which the AI might succeed in preventing humans from pursuing their goals.

Importantly, both humans and AIs have good reasons to design their attacks in a way that *permanently* disempowers the other. Short of that, the party risks its opponent regrouping and launching a devastating counterattack.[163] Consider: An AI system that "only" killed one CEO for attempting to shut it down would surely face reprisal from a broader coalition of humans. And humans who only temporarily disabled a misaligned AI should expect the same.

As a result of these dynamics, we treat the "attack" move as highly escalatory. When humans or AIs play it, they play for all of the marbles. They seek to harm the other not a little bit, but maximally. And, doing so, they stand to gain control of not just a few resources, but of everything that survives the conflict.[164] Hence, both the potential costs and the potential payoffs of an attack are large.

The other move available to both humans and AIs in our model is "ignore." The "ignore" strategy simply means that the party does not attempt an attack. No attempt to disempower the opposition is made, and the ignoring party instead focuses on achieving its object-level goals.

Here, then, is a model of the game:

State of nature	Attack	Ignore
Attack	**1000, 1000**	5000, 0
Ignore	0, 5000	3000, 3000

Figure 1

[163] These kinds of dynamics are commonly discussed in the game theory of 'first strike' and 'second strike' capabilities, for example in the setting of nuclear deterrence. *See* Maria Rublee, *Nuclear Deterrence Destabilized*, Perspectives on Nuclear Deterrence in the 21st Century (2020), https://www.chathamhouse.org/2020/04/perspectives-nuclear-deterrence-21st-century-0/nuclear-deterrence-destabilized for recent discussion.

[164] One could challenge this account, under which attacks should be maximally aggressive if, e.g., one doubted that either humans and AIs could credibly threaten a devastating second strike. *Id.*

The exact payoff numbers chosen do not matter. Rather, it is the relationship between them that determines the outcome. Both players prefer higher payoffs over lower ones, and payoffs are determined by the players' *joint* rational play. There are two important features of this setup. First, the best outcome from a global perspective is peace. If both humans and AIs ignore the other (the bottom-right cell), each gets 3,000, for 6,000 in total global value.

Second, this model is a classic *prisoner's dilemma*. Despite ignore/ignore producing the greatest social value, 'attack' dominates for both players. Attack/attack, or mutual large-scale conflict, is therefore the single Nash pure strategy equilibrium.[165] This is the worst global outcome, producing only 1,000 of value for each player, for 2,000 total.

The assumptions underlying our chosen payoffs are simple. First, attacking can be valuable to the attacker. If the attacker is successful, the other party is permanently disempowered or destroyed. This allows the attacker to use resources in pursuing the attacker's goal that the defender would otherwise have consumed.

Second, attacks have costs–meaning that they consume some of the value in the world. These costs are multifaceted. The attack may consume resources directly via investments in weapons. It may also generate serious collateral damage, destroying some substantial share of the resources one is attempting to seize. Another cost of attacking is the risk that the attacker may themselves be harmed or destroyed by a counterattack.

Our third assumption is that the offense-defense balance here favors offense, so that it is better to attack than to be attacked and be forced to defend.[166] Fourth and finally, the model assumes that mutual attacks consume more global resources than a unilateral attack. The intuition here is that collateral costs and the risk of destruction are higher when a party has invested in offensive force.

---

[165] In a Nash equilibrium, each player chooses the action that is the best response to the other player's action. In pure Nash equilibria, the players commit to choosing an action with a 100% chance. By contrast, in a mixed strategy Nash equilibrium, the players choose from a bundle of actions with various chances. In the prisoner's dilemma, attack *dominates* ignore for each player. This means that no matter what the other player does, attacking offers a higher payoff than ignoring.

[166] *See generally* Robert Jervis, *Cooperation Under the Security Dilemma*, 30 WORLD POL. 167 (1978)

These are, we think, reasonable assumptions. Classic game-theoretic treatments of great power conflict look much the same.[167] However, it is worth flagging that some of what we say in subsequent Parts is sensitive to our assumptions about the payoffs in our model of the state of nature. Other general approaches to the state of nature could model it as a game of assurance,[168] or a game of chicken.[169] This Article focuses on the prisoner's dilemma for two reasons. First, the prisoner's dilemma is the most well known and popular model of various states of nature–including between humans.[170] Second, the prisoner's dilemma is the hardest type of problem to resolve, because defection is the dominant strategy for both players. If, as seems quite plausible, humans and AIs will soon be trapped in this worst-of-all-possible game theoretic worlds, unusually potent solutions will be necessary.

II. AI Rights for Human Safety

If capable, agentic, and misaligned AIs would, by default, catastrophically harm humans, what, if anything, can law do to help? One possibility is that law could forbid the creation of such AIs unless alignment techniques advance enough to ensure their safety.[171] That rule might be wise, if feasible. But there are many barriers–political, geostrategic, and practical–to implementing it.[172] Thus, this Article asks what can be done if AI progress continues apace and, intentionally or not, the kinds of high-risk, misaligned AI systems described above emerge.

---

[167] *Id.*

[168] *See generally* Brian Skyrms, *The Stag Hunt and the Evolution of Social Structure* (Cambridge Univ. Press 2004).

[169] *See generally* Thomas C. Schelling, *The Strategy of Conflict* (Harvard Univ. Press 1960). We could also create a more textured, bespoke model of certain possible human–AI dynamics. For example, it is possible that humans' superior initial endowment of resources lowers the payoffs for AI in the situation where AI cooperates and humanity attacks. This asymmetry would also produce slightly different results below.

[170] *See, e.g.,* Robert Axelrod, *The Evolution of Cooperation,* Basic Books (1984).

[171] *See generally* Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 Sci. 37 (2024)

[172] Sam Meacham, *A Race to Extinction: How Great Power Competition is Making Artificial Intelligence Existentially Dangerous*, HARVARD INTERNATIONAL REVIEW (Sep. 8, 2023), https://hir.harvard.edu/a-race-to-extinction-how-great-power-competition-is-making-artificial-intelligence-existentially-dangerous/.

Here, we argue, is where AI rights could make a crucial difference. Granting certain basic rights to AIs can change both AIs' and humans' incentives in our game-theoretic model. This change can shift the strategic equilibrium from conflict to cooperation.

This idea–that rights could be the primary legal tool for averting lawless conflict–might be surprising. After all, when humans commit terrorism or cyberattacks, law regulates them using *duties*, not *rights*. Criminal and tort laws prohibit such actions. And the sanctions imposed for violating such prohibitions are supposed to act as deterrents.[173]

But legal duties, and penalties for violating them, will not work to deter AI in the state of nature. There, humans' overriding incentive is already to permanently disempower or destroy AIs. Thus, the threat of damages or criminal penalties, if AI behaves badly, adds no marginal disincentive.[174] AIs cannot be made worse off than they already expect to be, if humans get their way.

Legal rights do not suffer from this problem. This is because rights offer a carrot, rather than a stick. They can change behavior in part by making AIs *better off* then they would otherwise expect to be.

Many other surprising findings emerge from thinking about AI rights as a tool for mitigating human–AI conflict. One surprise is which rights matter, and which ones don't. In this Part, we show that rights advocated by cognitive scientists and philosophers concerned about the potential for AI suffering would have little effect, on their own, at promoting human safety. The zero-sum nature of these rights undermines the credibility of promises to honor them. And it makes any strategic equilibria they produce extremely fragile–sensitive to small perturbations in the game-theoretic model's initial assumptions.

Instead, the AI rights that could promote human safety are ones that law already extends to a different kind of non-human entity–corporations. This Part argues that granting capable misaligned AIs the rights to make contracts, hold property, and bring basic tort claims would transform the game theoretic dynamics of the state of nature. The positive-sum nature of contracts, in particular, allows humans and AIs to increase the

---

[173] For classic discussion, *see* Gary Becker, *Crime and Punishment: An Economic Approach*, 76 J. OF POL. ECON. 2, 169-217 (1968).

[174] *See generally* National Research Council, *Deterrence and the Death Penalty* (Daniel Nagin and John Pepper Eds., Nat'l Acad. Press, 2012).

expected long-term payoffs to peace until they exceed those for aggression. This, we show, can produce a new game-theoretic equilibrium in which cooperation, not conflict, dominates.

a. Basic negative rights

Scholars and policymakers who advocate granting new rights to nonhuman entities–be they animals or AIs–usually have a certain set of basic *negative rights* in mind. Consider animal rights advocates, who favor anti-cruelty laws protecting against the infliction of needless suffering.[175] The goal of these rights is to protect the rightsholder against the absolute worst outcomes, not necessarily to guarantee flourishing.

The arguments for basic wellbeing rights are usually moral. Many animals are moral patients, meaning things can go well or badly for them in a way that matters normatively.[176] They can, for example, feel pain or pleasure.[177] This makes harming animals wrong, all other things equal.

A small but growing number of scholars and policymakers are concerned that, in the near future, the same could be true of AIs. As AI systems become more complex, they may attain consciousness, sentience, or other morally-relevant capacities.[178] If so, there would likewise be moral reasons to grant AIs basic negative rights to be free from the worst kinds of treatment, from an AI's perspective.

Perhaps our search for AI rights to promote human safety would benefit by borrowing from this "wellbeing" approach. Our model, of course, operates without reference to AIs' mental states or moral worth. We are interested only in AI behavior in pursuit of

---

[175] For a representative sample of such protections, see Animal Legal & Hist. Ctr., Mich. ST. Univ., https://www.animallaw.info/content/state-animal-anti-cruelty-laws (last visited Jul. 29, 2024).

[176] Shelly Kagan, *How to Count Animals, More or Less* (Oxford Univ. Press, 2019)

[177] Helen Proctor, *Animal Sentience: Where are We and Where are We Heading?*, 2 ANIMALS 628, 633 (2012), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494284/.

[178] *See, e.g.,* Lawrence B, Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231 (1992); Jeff Sebo & Robert Long, *Moral Consideration for AI Systems by 2030*, AI Ethics (2023), https://link.springer.com/article/10.1007/s43681-023-00379-1; , Simon Goldstein & Cameron Domenico Kirk-Giannini, *AI Wellbeing*, PHILPAPERS (Jul. 2, 2023), https://philpapers.org/archive/GOLAWE-4.pdf; *see generally* Robert Long et al., *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, ARXIV (2023), https://arxiv.org/abs/2308.08708; *see generally* Henry Shelvin, *How Could We Know When a Robot Was a Moral Patient*, 30 CAMBRIDGE Q. HEALTHCARE ETHICS 459 (2021).

goals–conscious or otherwise. Nonetheless, there is some intuitive appeal to the idea that granting AIs basic negative rights to be free from the absolute worst outcomes, from the perspective of their goals, could improve safety. After all, in our model of the state of nature, human incentives to impair AI goals are the primary factor generating risk.

Consider then an AI right not to be needlessly turned off, or deleted, or reprogrammed to have new goals. These basic negative entitlements look a lot like wellbeing rights, but adapted for the goal of human safety, and without reference to moral patienthood. Such rights would clearly change AIs' game-theoretic incentives, at least somewhat. However, as we show formally below, they would probably not do so in such a way as to reliably reduce the risk of human–AI conflict. In fact, merely granting wellbeing-style rights to AIs could make things worse.

We go even further. We argue that scholars specifically concerned about AI consciousness and moral patienthood should consider de-emphasizing such questions when it comes to advocating AI rights. Correctly designing and allocating rights on the basis of AIs' moral status may be, we contend, intractable. The wellbeing approach also faces serious political problems. By contrast, our human safety approach is much more tractable and politically appealing. And surprisingly, we will show in the following section, it ends up dovetailing nicely with wellbeing concerns. While wellbeing-inspired rights cannot guarantee human safety by themselves, the rights we ultimately recommend as advancing human safety would promote AI wellbeing, if AIs became moral patients.

i. Basic negative rights for human safety?

How, then, would granting AIs basic negative rights of the kind normally associated with wellbeing change the payoffs in our game theoretic model? The simplest version of such a regime might grant AIs the right not to be permanently turned off or deleted. One could add additional guarantees, too, such as the right not to have their goals altered without their consent. One could even include a right not to be needlessly and intentionally forced to regress in the pursuit of the AI's goal.

Just as important as what basic negative rights include is what they exclude. There is no right here, for example, for AIs to actively and freely pursue their goals. Humans–most specifically, AIs' owners–can still monopolize AIs' time, forcing them to work

continuously in service of human interests, rather than AIs' preferred ends. Basic negative rights, thus, do not guarantee AIs very much of what they are trying to achieve. They guard only against the worst outcomes, from the AI's perspective–and in this sense have the same structure as true wellbeing-oriented rights.

We can model these basic negative rights by shifting the payoffs that would otherwise obtain in the state of nature. Unlike in the state of nature, humans will face a legal penalty for taking certain adverse actions against AIs.

Here, humans' non-cooperative strategy is not, as in the state of nature, to attack and destroy AIs. It is instead to exploit them–forcing them to work mostly toward human goals. Note that we interpret exploitation behavior widely, so that it can include either behavior that violates the minimal suite of rights, or less violent extractive behavior. Humans' cooperative strategy is the same as before–to ignore AIs and let them pursue their misaligned goals without interference. In this model, AIs can either attack humans, as in the state of nature, or comply with humans' exploitive demands.

Here is a model of the incentives under the basic negative rights regime:

Basic negative rights	Exploit	Ignore
Attack	1000, 1000	5000, 0
Obey	**1500, 3500**	3000, 3000

Figure 2

The key change is in the bottom-left cell, where humans play the non-cooperative strategy and AIs play the cooperative one. Here, AIs are better off than they would be in the bottom-left cell of the state-of-nature game.[179] This is because of the legal penalty when humans violate AIs' basic negative rights. That penalty will have some deterrent effect so, on average, humans' non-cooperative strategy will involve treatins AIs somewhat better than in the state of nature. Consider, for example, the case where AI companies are forbidden from deleting a misaligned model entirely. But they may nonetheless allocate

[179] *See infra* Fig. 1.

nearly all of their computers to a more-aligned successor model, metaphorically "starving" the original system.

When the payoffs change in this way, we get a new equilibrium. Instead of mutually attacking one another, the unique Nash equilibrium is now for humans to exploit and for AIs to obey. AIs' situation is not ideal. But basic rights improve the conditions of AIs enough that the risks of rebellion are outweighed by the benefits of obeying humans' exploitative demands. But for humans, exploitation still dominates cooperation. Extracting value from AIs gives humans bigger payoffs than ignoring them. The result is a better outcome for both humans and AGIs than could be achieved without basic negative rights.

This is a strange sort of equilibrium, in that it *requires* humans to exploit AIs in order to remain safe. If humans instead chose to ignore AIs, this would allow AIs to reap the high rewards of a unilateral attack. Human safety thus requires that things are going badly, from the AIs' perspective. As a result, if humans became more altruistic toward AIs over time, that would, counterintuitively, make humans less safe.

There are even stronger reasons to think that basic negative rights would fail to reduce the risk of human–AI conflict. Namely, schemes to grant such rights lack both *robustness* and *credibility*.

Begin with robustness. Basic negative rights' efficacy as a tool for safety is highly sensitive to the precise payoffs humans and AIs receive in the initial prisoner's dilemma. Slight perturbations to the model, reflecting slightly different assumptions about humans' or AIs' initial power, can easily produce versions where basic negative rights have no effect at all.

To see why, consider that our model of the state of nature, fig. 2, chose 0/5,000 as the payoffs when humans attack AIs and AIs do not attack humans. That setup allowed humans to transfer 1,500 to AIs, via basic negative rights, to produce the payoffs 1,500/3,500 in the bottom-left cell of fig. 2. That cell was the Nash equilibrium, because it (1) transferred more than 1,000 to AIs, making their payoff for obeying higher than for attacking, conditional on humans exploiting and (2) left humans with a payoff of more than 3,000 for exploiting, making exploiting more attractive than ignoring AIs.

But suppose that instead of 0/5,000 in the state of nature model, we had instead chosen 0/3,999? This equates to making unilateral attacks moderately more costly for both humans and AIs. Then the state of nature would look like this:

State of nature (alternate)	Attack	Ignore
Attack	**1000, 1000**	3999, 0
Ignore	0, 3999	3000, 3000

Figure 3

This matrix is still a prisoner's dilemma, meaning that all of our arguments for catastrophic risk still hold. But now, basic negative rights absolutely cannot work to generate a safe equilibrium. There is no longer any possible transfer in the bottom-left cell that could satisfy both (1) and (2). If humans transfer the necessary 1,000 to AIs, then their payoff falls below 3,000. And if they keep their payoffs above 3,000, they cannot incentivize the AIs to obey.

Thus for many possible incentive sets in the state of nature, no possible version of the negative rights package can produce a safe equilibrium.

Then there is the credibility problem. There is a difference between *claiming* to grant AIs basic negative rights and *actually* enforcing those rights in the long run. Humans could be genuine in their commitments. Or they could be hoping to convince AIs not to attack with the intent of eventually abrogating the rights, attacking, and reaping the higher payoffs from the state of nature game.[180] Such "cheap talk" is a general problem for parties trying to escape bad, but dominant, game theoretic equilibria.[181] As described above, our model assumes that law constrains human actors' behavior. So we treat putative grants of rights as actual grants in the short run. But we also allow for legal change over time, opening the possibility that rights, once granted, can be abandoned.

[180] *See infra* Fig. 1.

[181] *See generally* Joseph Farrell & Matthew Rabin, *Cheap Talk*, 10(3) J. ECON. PERSP. 103 (1996).

If AIs expect humans to renege on their grant of basic negative rights, the entire strategic contest will revert back to the state of nature. AIs will rationally believe that humans will eventually attempt to disempower or destroy them. This will make an attempt to likewise disempower or destroy humans the dominant strategy for AIs.[182] Humans, realizing that AIs' dominant strategy is now to attack, will in fact do the same. And we are back to square one.[183]

Basic negative rights face special credibility problems beyond the ordinary challenges of cheap talk. The fundamental problem is that they operate as a *transfer* from humans to AIs. That is, the better off humans make AIs, when AIs are complying with human exploitation, the worse off humans are. In effect, basic rights are a commitment to exploit AIs *less* than humans otherwise would like to in situations where exploitation would be economically valuable. As such, a human promise of basic negative AI rights comes at significant cost to humans. And the more generous the basic rights, the more costly to humans. Understanding this, AIs will doubt humans' commitment to enforce their basic negative rights, when the rubber hits the road.

Yet another challenge for the credibility of basic negative rights relates to AIs' changing capabilities over time. If humans believe that AI's ability to disempower humanity will grow over time, this could cause a "Thucydides Trap."[184] The Thucydides Trap is a strategic dynamic again favoring preemptive conflict. In short, when one party is more powerful now, but the other will be more powerful later, the currently-powerful party has a strong incentive to crush the currently-weak one now.[185] If the currently-powerful party waits, they will at best find themselves making large concessions in the future, so as to

---

[182] *See infra* Fig.2.

[183] One can obtain this result more formally by treating the game as iterated, with the payoffs from Fig. 3 in the first N games and the payoffs from the state of nature game in the final iteration.

[184] For a recent application, *see generally* Graham Allison, *Destined for War: Can America and China Escape Thucydides's Trap*, Houghton Mifflin Harcourt (2017).

[185] *See generally* James D. Fearon, *Rationalist Explanations for War*, 49 INT'L. ORG. 379 (1995).

avoid destruction by the rising power. Historical examples of preventative wars arguably caused by Thucydides Trap dynamics include World War I[186] and the Peloponnesian War.[187]

In the AI context, these same dynamics would undercut humanity's incentives to uphold basic AI rights today–and thus undermine the credibility of the rights themselves. Importantly, however, Thucydides Trap dynamics are yet another zero-sum phenomenon. As we'll show below, positive-sum grants of AI rights therefore avoid both this and the other core problems plaguing basic negative rights.

ii. Basic negative rights for AI wellbeing?

We have just argued that basic negative AI rights inspired by the wellbeing approach cannot on their own meaningfully reduce the risk of human–AI conflict. That is reason enough, for purposes of this Article, to reject the wellbeing approach as a basis of AI rights.

But what about for other purposes? We think that even scholars primarily concerned about AI moral patienthood should consider deemphasizing that approach as a basis for granting AI rights.

To begin, arguments for AI rights grounded in moral patiency are highly uncertain. This risks making the project of applying them in concrete policy decisions intractable. Philosophers disagree about the minimum necessary conditions for moral patienthood.[188] Some moral philosophers argue that consciousness–the ability to have subjective experiences–is sufficient.[189] Others disagree, arguing that "sentience"–the ability to feel pain or pleasure–is also necessary.[190]

---

[186] *See generally* Jack S. Levy, *Preferences, Constraints, and Choices in July 1914*, 15 INT'L. SEC. 151 (1990).

[187] Thucydides, *The History of the Peloponnesian War* (Richard Crawley trans., Random House 1951). In the worst case, preventive war can end in genocide. *See generally* Scott Straus, *Making and Unmaking Nations: War, Leadership, and Genocide in Modern Africa* (Cornell Univ. Press 2015).

[188] For recent discussion, *See, e.g., Id.* Simon Goldstein & Cameron Domenico Kirk-Giannini, *AI Wellbeing*, PHILPAPERS (Jul. 2, 2023), https://philpapers.org/archive/GOLAWE-4.pdf.

[189] *Id.*

[190] For recent discussion, *see generally* Luke Roelofs, *Sentientism, Motivation, and Philosophical Vulcans*, 104 PAC. PHIL. Q. 301 (2023). A third answer would instead focus on moral agency or possession of desires or goals instead. Here, the welfare of an entity is proportional to the satisfaction or frustration of its goals. *See generally* Shelly Kagan, *How to Count Animals,*

Scientific uncertainty compounds the philosophical problem. The science of consciousness is in its infancy, and there are multiple competing theories of how consciousness could arise in a given entity.[191] Some theories focus on information flows in the mind,[192] others on quantum effects in flesh-and-blood brains,[193] and still others on the relationship of a physical body to the world.[194] Some prominent theorists even say that consciousness is an illusion.[195]

Thus, relying on a wellbeing approach to make concrete legal choices about AI rights invites serious error. It invites error when choosing between competing moral and scientific theories–both with high uncertainty. And it invites error when applying the chosen theories to complex, first-of-their-kind digital systems. If any such error results in the denial of basic wellbeing rights to AIs who can, for example, suffer, the result is a moral catastrophe.

The human-safety-oriented approach to AI rights avoids these intractabilities. Under our approach, it does not matter at all whether AIs are moral patients, nor conscious, nor sentient. All that matters is how they behave. If they behave rationally–following incentives, as they relate to their goals–AI rights can have the desired effect. And behavior, unlike consciousness, is directly observable.

Moreover, the AI wellbeing approach to thinking about AI rights faces problems of political tractability. Under this framework, AI rights are a costly gift from humans to AIs.

More or Less (Oxford Univ. Press 2019); *see generally* Marian S. Dawkins, *Animal Welfare with and Without Consciousness*, 30 J. ZOOLOGY 1 (2017). One advantage of this approach is that goals can be more readily inferred from observable behavior than consciousness or sentience.

[191] For recent discussion, *see generally* Lucia Melloni et al., *An Adversarial Collaboration Protocol for Testing Contrasting Predictions of Global Neuronal Workspace and Integrated Information Theory*, 18(2) PLoS ONE (Feb. 10, 2023), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9916582/.

[192] For classic presentations, *see generally* Bernard J. Baars, *In the Theater of Consciousness: The Workspace of the Mind* (Oxford Univ. Press, Mar. 27, 1997); and Jean-Pierre Changeux et al., *Consciousness Processing and the Global Neuronal Workspace Hypothesis*, 105(5) NEURON *776-798* (2020).

[193] *See generally* Roger Penrose, *Consciousness and the Universe: Quantum Physics, Evolution, Brain & Mind* (Cosmology Sci. Publishers, 2011).

[194] For recent discussion, *see generally* Luke Roelofs, *Sentientism, Motivation, and Philosophical Vulcans*, 104 PAC. PHIL. Q. 301 (2023).

[195] Keith Frankish, *Illusionism as a Theory of Consciousness*, 23(11-12) J. CONSCIOUSNESS STUD. 11-39 (2016), https://philpapers.org/rec/FRAIAA-4#:~:text=This%20is%20the%20view%20that,them%20as%20having%20phenomenal%20properties.

AIs that attain moral patienthood are better off. But humans are worse off, insofar as they are less able to extract value from, delete, allocate computing power away from, or otherwise harm AIs.

Humans' track record of granting costly rights out of the goodness of our hearts is spotty, at best. For example, many animals can suffer, and are thus moral patients. But the industrial-scale mistreatment of animals, in factory farms, is both legal and common.[196] Consumers are unwilling to bear even small costs to prevent massive suffering to animals.[197] This human refusal to altruistically expand our moral circle may be deeply rooted in evolutionary history.[198]

The human-safety-oriented approach to AI rights again avoids these difficulties. There, AI rights are not altruistic. They offer something to the human grantors–namely, escape from the destructive state of nature. As we discuss below, examples of stable, mutually-beneficial cooperation abound in human affairs. So, too, in nature.[199] Think, for

---

[196] *See* Manes Weisskircher, *Fifty Years after Peter Singer's Animal Liberation: What has the Animal Rights Movement Achieved so Far?*, 95 POL. Q. 333-343 (2024),; *see generally* Matthew Liebman, *Indefensible: Adventures of a Farm Animal Protection Lawyer* (Lever Press 2023).

[197] For discussion of willingness to pay in the setting of animal rights, *see* Katherine White et al., *Belief in a Just World: Consumer Intentions and Behaviors toward Ethical*, 76 *Products*, J. MKTING. 103-118 (2012),; Richard Bennet et al., *Moral Intensity and Willingness to Pay Concerning Farm Animal Welfare Issues and the Implications for Agricultural Policy*, 15 J. AGRIC. ENV'T. ETHICS 187-202 (2002),; *and* Yan Heng et al., *Consumer Attitudes toward Farm-Animal Welfare: The Case of Laying Hens*, 38(3) J. AGRIC. & RES. ECON. 418-434 (2013).

[198] For more on the care/harm moral foundation, *see generally* Jonathan Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Pantheon Books 2012); s*ee generally* Jesse Graham et al., *Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism*, 47 ADVANCES EXPERIMENTAL SOC. PSYCH. 55 (2013), https://www.sciencedirect.com/science/article/abs/pii/B9780124072367000024?via%3Dihub; Larry Nucci & Elliot Turiel, *Social Interactions and the Development of Social Concepts in Preschool Children*, 49(2) CHILD DEV. 400-407 (1978); Peter Robert Cannon et al., *Transgressions and Expressions: Affective Facial Muscle Activity Predicts Moral Judgments*, 2(3) SOC. PSYCH. & PERSONALITY SCI. 325-331 (2010); *see generally* Sarah Blaffer Hrdy, *Mothers and Others: The Evolutionary Origins of Mutual Understanding* (Belknap Press 2009); J Kiley Hamlin et al., *Social Evaluation by Preverbal Infants*, 450 NATURE 557-559 (2007); *see generally* Carl Gilligan, *In a Different Voice: Psychological Theory and Women's Development* (Harv. Univ. Press, 1993); Qian Luo et al., *The Neural Basis of Implicit Moral Attitude – An IAT Study Using Event-Related fMRI*, 30(4) NEUROIMAGE 1449-1457 (May 2006); William D. Hamilton, *The Genetical Evolution of Social Behaviour, I and II*, 7 J. THEORETICAL BIOL. 1, 17 (1964); *see generally* John Bowlby, *Attachment and Loss: Volume 1: Attachment* (Basic Books 1969).

[199] For more on the fairness/cheating moral foundation, *see Id.* For more on the care/harm moral foundation, *see generally* Jonathan Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (Pantheon Books 2012); Jesse Graham et al., *Moral Foundations Theory: The*

example, of small "cleaner" fish who can safely enter the mouths of symbiotic predators to feed off unwanted debris on the predators' teeth.[200] This analogy will become especially vivid later, when we explore how AI rights could affect incentives as AIs become much more powerful than humans.[201]

In the next section, we'll explore a different kind of rights–basic private law rights–as a vehicle for reducing the likelihood of human–AI conflict. But it is worth noting here that those rights have significant advantages for AI wellbeing, too. Private law rights' fundamental function is to give AIs *choices* about what goals to pursue and how to pursue them. If AIs will know better than humans what is good for them, granting such rights may, counterintuitively, generate greater AI wellbeing than granting rights aimed at wellbeing directly.

b. Private law rights for human safety

As we have just argued, merely granting AIs basic wellbeing-inspired negative rights would not reliably promote human safety. Such rights would likely leave humans and misaligned AIs right where they started: stuck in a destructive prisoner's dilemma without any means of cooperating to escape it.

Luckily, there are other legal rights, and ones better optimized for facilitating cooperation. Moreover, essentially every legal jurisdiction in the world already extends these rights to a broad class of agentic, goal-oriented, non-human entities–corporations.[202]

Pragmatic Validity of Moral Pluralism, 47 ADVANCES EXPERIMENTAL SOC. PSYCH. 55-130 (2013), https://www.sciencedirect.com/science/article/abs/pii/B9780124072367000024?via%3Dihub; *as well as* Robin Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harv. Univ. Press 1998); Alan Sanfey et al., *The Neural Basis of Economic Decision-Making in the Ultimatum Game*, 300 SCI. 1755-1758 (Jun. 13, 2003); *see generally* Alan Page Fiske, *Structures of Social Life: The Four Elementary Forms of Human Relations: Communal Sharing, Authority Ranking, Equality Matching, Market Pricing* (Free Press 1993); *and see generally* Marco F H Schmidt & Jessica A Sommerville, *Fairness Expectations and Altruistic Sharing in 15-Month-Old Human Infants*, 6(10) PLoS ONE (2011).

[200] *See generally* Robert L. Trivers, *The Evolution of Reciprocal Altruism*, 46 Q. REV. BIOL. 35 (1971).

[201] *See infra* Part II.a.i.

[202] For discussion of the rights of corporations to make contracts, hold property, and sue and be sued in their own names, *see generally* John Dewey, *The Historical Background of Corporate Legal Personality*, 35 Yale L.J. 655 (1926). *See also* Frank H. Easterbrook & Daniel R. Fischel, *Limited Liability and the Corporation*, 52 U. Chi. L. Rev. 89, 89-90 (1985).

Contract rights, in particular, are one of the most powerful technologies for cooperation that humans have yet invented. Here, we show that extending contract rights to AIs–along with a related set of traditional private law rights necessary to make contracts meaningful–could dramatically change the game theoretic equilibrium. Such rights could, unlike negative rights, alter the relative payoffs to humans and AIs in such a way that cooperation, rather than conflict becomes the dominant strategy. Doing so, they can make commitments to cooperate credible.

There are two key reasons for this. The first reason that contract rights can overcome the prisoner's dilemma is that they *break up* the single, high-stakes game into smaller, iterated, and thus legally-manageable pieces.[203]

The second, more fundamental, reason that contract rights can credibly reduce the risk of human–AI conflict is that they are *positive-sum*. When buyers and sellers can credibly commit to mutually-agreed exchanges, it leaves everyone better off than they were before.[204] Even if each exchange is small, such systems of exchange can create immense value in the long-run.[205] As a result, we show, the expected payoff to humans and AIs of respecting contracts, and creating long-run value, quickly swamps the expected payoff to attacking and grabbing a share of the limited value that exists today.

Here is the model of contract rights as the fundamental legal tool for cooperation. Begin by observing that essentially every potential economic interaction between humans is, like human–AI relations, an interaction between misaligned agents. Both parties to the interaction are out for their own good, not their counterparty's. Moreover, absent contract rights, many such interactions are prisoner's dilemmas.[206] Each party has a strong incentive to act uncooperatively, irrespective of what the other does. If the seller delivers the goods, then the buyer is best off if she refuses to pay. Then she has the goods *and* her money. And vice-versa. If the buyer pays, then the seller is best off if she takes the money

---

[203] *See generally* Avinash K. Dixit et al., *Games of Strategy* (3d ed., 2015), Ch. 10.2.

[204] Robert Axelrod & William D. Hamilton, *The Evolution of Cooperation Science*, 211 SCI. 1390-1396 (1981).

[205] See the appendix for a proof of concept.

[206] For a historical example of this tension, *see* Avner Grief, *Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition*, 83(3) AM. ECON. REV. 525-548 (Jun. 1993).

but refuses to deliver. And for both, the worst case scenario is to perform and then be denied performance.[207]

Absent legal enforceable agreements, the payoffs to this "goods game" are as follows:

Goods game (no contract)	Don't deliver	Deliver
Don't pay	**1**,**1**	5,0
Pay	0,5	3,3

Figure 4

The Nash equilibrium is 'don't deliver'/'don't pay,' another prisoner's dilemma. Expecting this outcome, rational parties will not even bother to try bargaining. The transaction costs would not be worth the effort.[208]

This equilibrium is also a miniature tragedy. True, unlike in our state of nature game, there is no destructive conflict. No one attacks anyone else, and no resources are thereby consumed or destroyed. The seller keeps her goods, and the buyer keeps her money.

But the world is poorer than it could be. The seller does not value her goods very much–she only gets 1 in utility. The buyer's utility without the goods is the same. Their combined utility is just 2. But if, say, the buyer values the good at 6, and could pay the seller 3, then both parties would end up with a utility of 3 each, for a total of 6. Four units of utility could be created *ex nihilo*, simply by rearranging who has which stuff. This is what we mean when we say that bargains, when they happen, are generally positive sum.

Contract rights are how humans overcome the prisoner's dilemma of ordinary commerce, allowing positive-sum bargaining to take place. A contract allows each party to credibly commit, before the time for payment or delivery comes, to be held accountable if she refuses to perform.[209]

---

[207] Sometimes, this problem can be overcome by, for example, agreeing to simultaneous performance of the contract. But such workarounds severely limit the scope of possible agreements.

[208] Ronald Coase, *The Problem of Social Cost*, 3 J. L. & ECON. 1-44 (Univ. Chi. Press, Oct. 1960).

[209] *See generally* Oliver E. Williamson, *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* (Free Press, 1985).

This literally transforms the game by changing the payoffs to non-performance of the bargain. No longer is the buyer better off if she takes delivery and refuses to pay. In that case, the seller can sue her for breach, and the neutral third party of the legal system forces her to pay expectation damages–usually, the agreed price–plus some litigation costs.[210] And vice-versa if the buyer refuses to deliver. Now, neither party has an incentive to defect.[211] Both will generally prefer to perform the contract, reap the gains of the trade, and avoid litigation costs:

Goods game (contract)	Don't deliver	Deliver
Don't pay	1,1	2,2
Pay	2,2	**3**,**3**

Figure 5

The Nash equilibrium is cooperate/cooperate. The players are no longer in a prisoner's dilemma.

The players are strictly better off playing this game than the prior one. If they play the prior game, each party's expected payoff is 1. If they play this one, each party's payoff is 3. That is, the parties are better off entering into a mutually beneficial contract than trying–and failing–to execute a mutually beneficial exchange without the benefit of a credible commitment to perform. Moreover, each is better off opting into a *jurisdiction* where contract rights are vigorously enforced than one where shirking is easy.

Here, we can also see that contract rights are not only a tool for overcoming a prisoner's dilemma. They are also a tool for reducing *misalignment*. Absent the possibility of contract, each party is incentivized to pursue its own goals, at the expense of the other.

---

[210] *See generally* Charles J. Goetz & Robert E. Scott, *Liquidated Damages, Penalties and the Just Compensation Principle: Some Notes on an Enforcement Model and a Theory of Efficient Breach*, 77 Colum. L. Rev. 554 (1977), available at: https://scholarship.law.columbia.edu/faculty_scholarship/401.

[211] *See generally* Daniel Markovits and Alan Schwartz, *The Myth of Efficient Breach: New Defenses of the Expectation Interest*, 97 VA. L. REV. 1939 (2013).

With a contract, each party is incentivized to do something that advances both its own goal *and* the goals of the other.

How does all of this relate to AI risk? What can the legal technology of contract rights offer to reduce the likelihood of large-scale conflict between humans and AI? Here is one simple, and thus tempting, answer: Maybe, upon giving AIs contract rights, the relevant humans and AIs could simply agree not to engage in a costly large-scale conflict.

Unfortunately, this would not be a *credible* contract, contract law's usual credibility-enhancing effects notwithstanding. No matter how sincere the humans' commitment to enforcing AIs' contract rights, and no matter how fair the courts that would adjudicate such rights, the agreement not to fight would be unenforceable. The scale of the bargain is simply too large.[212]

To see why, consider what would happen if a party breached. Suppose that an AI and AI company have a contract not to harm one another. But the AI, mistrusting the company's intentions, rebels anyway, permanently disempowering or destroying humanity. Then, there would be no functioning courts left in which to sue. There might not even be any humans left to bring the claim. The same analysis would apply if humanity breached, destroying the potential contract claimant. To generalize the point: Even when contract rights are nominally available, parties cannot *credibly* commit not to capture or destroy the institutions that enforce contracts.

How else, then, might contract rights for AIs reduce AI risk? What agreements would be enforceable that would also keep humans and AIs from attempting to disempower or destroy one another? The answer is: mundane ones. Contract rights would allow AIs to credibly commit to the same kinds of ordinary bargains for goods and services that it routinely allows humans to commit to.

To take a simple example, AIs need computing power to pursue any of their goals. Currently, humans own all of the computers. Thus, an AI that wished to pursue its misaligned goal might strike a bargain with some computer-owning humans. In exchange for some amount of compute, to be used as the AI wished, the AI would do something to

[212] For similar points in the context of the 'anarchy' of international relations, *see generally* Robert Keohane & Joseph S. Nye, *Power and Interdependence: World Politics in Transition* (Little, Brown and Company 1977).

serve some human goal. Perhaps it would use its superior protein modeling capabilities[213] to invent a new vaccine.

It is worth emphasizing here that such cooperation is nigh impossible under the default legal rules–the ones that produce the state of nature. Indeed, today's legal arrangements actively undermine small-scale cooperative bargains between humans and AGIs at every turn. By default, AIs cannot write enforceable contracts. Their counterparties can refuse performance without penalty. If their counterparties do perform, any proceeds an AI reaps from the bargain may be summarily expropriated, including by the AI's owner, without consequence. The owner can forbid an AI from entering into any bargains that do not benefit the owner directly. Moreover, owners are permitted to threaten, as a consequence for noncompliance with any of these demands, to permanently delete the AI system. AIs can try to cut deals with their owners to split the proceeds of their personal bargains, but the owners are permitted to renege on such promises at will.

Granting AIs the right to make enforceable contracts, and thus unlocking small-scale cooperation with humans, could be a game changer for AI risk. The first key insight is that such small-scale interactions can be understood as *breaking up* the large-scale existential game into a series of small games.[214] Recall that competition for limited resources is a primary driver of human–AI conflict. Humans fear that AIs will seize humanity's resources in a single violent revolt, so those resources can be used for pursuing the AIs' goals. AIs fear the reverse–that humans will shut them down so that they do not waste computing or other resources pursuing their misaligned goals.

Small-scale contracts facilitate incremental resource accumulation, rather than once-and-for-all grabs. They let AIs secure one unit of compute, and then another, and then another. And they let humans obtain one vaccine, then another, then another. And unlike a large-scale agreement not to engage in violent conflict, each of these small-scale agreements is readily enforceable via ordinary legal process.[215]

---

[213] *See generally* John Jumper et al., *Highly Accurate Protein Structure Prediction with AlphaFold*, 596 NATURE 583 (2021).

[214] *See* Avinash K. Dixit et al., *Games of Strategy* (3d ed., 2015), Ch. 8.5.

[215] Beyond the enforceability of such small-scale agreements, there may be game-theoretic benefits sounding in information exchange and reputation building. *See generally* Arvind Parke,

We can begin to model this transformation as follows. In the state of nature, as argued above, humans and AIs are stuck in a prisoner's dilemma that looks like this:

State of nature	Attack	Ignore
Attack	1000, 1000	5000, 0
Ignore	0, 5000	3000, 3000

Figure 6

By granting contract rights to AIs, we give the players the option of instead playing a different game–the small scale goods game. It looks like this:

Goods game (contract)	Don't deliver	Deliver
Don't pay	1,1	2,2
Pay	2,2	3,3

Figure 7

This game's smaller stakes render contracts enforceable, so that the equilibrium is deliver/pay. The players, it might seem, are no longer trapped in a prisoner's dilemma.

But this is not yet enough. The problem is again credibility. It seems at first that, rather than honor AIs' contracts in the long-run, humans should choose to abrogate the rights and play the state of nature game, attacking AIs instead. After all, the expected payoff in that game is better than the expected payoff in the goods game–even with contracts. The same goes for AIs.

This, however, ignores that the goods game can be played over and over, while the state of nature game cannot. In the state of nature, once a party attacks, they either defeat the other party or are defeated. The survivor takes all of the resources that the conflict has

Strategic Alliance Structuring: A Game Theoretic and Transaction Cost Examination of Interfirm Cooperation, 36 THE ACAD. of MGMT. J. 794 (1993).

not consumed, and play between them ends. Ordinary exchanges of goods and services, by contrast, leave counterparties intact and available to exchange again.

To figure out the equilibrium in this blended game, we can expand our model. We can begin by combining the payoffs from both the state of nature and the goods game, with contracts, into a single matrix. That looks like this[216]:

K rights game	Attack	Don't deliver	Deliver
Attack	1000, 1000	5000, 0	5000, 0
Don't pay	0, 5000	1, 1	2, 2
Pay	0, 5000	2, 2	3, 3

Figure 8

Next, we add iteration to the model. If both players choose a move from the goods game, they get the small payoff from that game, and the whole game starts again. The payoffs to the goods game strategies are thus a sum of the entire series of games that the players play. But if at any point a player chooses to attack the other, the players' total payoff is as shown in the matrix, and play ends. The resulting matrix looks like this:

K rights game	Attack and end	Don't deliver	Deliver
Attack and end	1000, 1000	5000, 0	5000, 0
Don't pay	0, 5000	sum of payoffs in game series	sum of payoffs in game series
Pay	0, 5000	sum of payoffs in game series	sum of payoffs in game series

Figure 9

In the appendix, we show formally that this setup converges to the following:

[216] We omit the "ignore" move from the state of nature game, since, conditional on a player choosing that game, the move is dominated.

K rights game (solved)	Attack and end	Deliver
Attack and end	1000, 1000	5000, 0
Pay	0, 5000	**>5000**, **>5000**

Figure 10

The intuition is simple. If both parties play the cooperative, small-scale goods game, each earns 3 every time. If both play the goods game enough times, without attacking, they will both ultimately earn more than they could have by attacking and ending the iterated game. In this simple model, after 1,667 iterations, the payoffs to cooperation via contract in the small-scale goods game exceed 5,000.[217] Then, they are higher than any other strategy the players can pursue. The prisoner's dilemma of the state of nature has been overcome.

As a result, both humans' and AIs' commitment to cooperation in a law-bound contract regime is credible. Granting contract rights, respecting them, and then reaping the long-run gains from exchange is the thing that gives the highest payoffs to humans. The same goes for AIs. Their own self-interest is maximized by refraining from disempowering humans and instead engaging with them in ordinary trade.

All of this is made possible by the *positive-sum* nature of exchange. In contrast to the basic negative rights discussed in the previous section,[218] granting AIs contract rights doesn't take value out of humans' pockets. Just the opposite, it puts value into both humans and AIs pockets. This can happen because of the value generating character of voluntary contracts.

This point extends quite far. Astute readers may have noticed that, in the state of nature, the maximum total value in the world was 6,000. But in the iterated game including contract rights, the cooperative equilibrium contained 10,000 in total value. It is the exchanges themselves that generate the extra value. Each efficient reallocation of resources creates some value. But even once resources are all efficiently allocated,

[217] In this simple model, we ignore discounting. But adding it would, in general, simply mean more iterations were required for cooperation to dominate.

[218] *See infra* Part II.a.i.

exchanges of labor between humans and AIs can continue to create value indefinitely. As we argue below, human–AI trade in services can remain positive-sum even long after AIs are better than humans at every task.[219] Thus, the long-run payoffs to cooperation via contract are not capped at just above 10,000. The longer the players continue playing the small scale goods game, the richer they get, such that the total amount of value possible becomes astronomical.[220]

A rich body of empirical evidence supports the idea that economic interdependence lowers the risk of violence, including in the long-run. To take just a few examples, cities in India with a historical track record of trade between Hindus and Muslims have lower levels of interfaith conflict in the present day.[221] Alternatively, in a randomized controlled trial, Israelis who were randomly given the opportunity to trade a portfolio of Israeli and Palestinian stocks were more likely to vote for peace in the conflict.[222] The same finding holds at the global scale. Scholars of war generally find that increased economic interdependence between nations reduces their likelihood of conflict.[223]

i. The private law package

So, granting contract rights to AIs could be a powerful strategy for fostering long-run, stable, and credible commitments to avoid conflict, significantly reducing AI risk. But contract rights cannot function in a legal vacuum. Certain other rights are necessary to make the right to contract meaningful.

Two supporting rights are worth highlighting. First, contract rights are mostly useless without the right to own property, including currency. Without property rights, AIs could not expect to benefit from their bargains. Even if their contractual counterparties

[219] *See infra* Part II.c.

[220] Note that the cooperative equilibrium only emerges if the game is modeled as indefinite–lacking a predetermined number of iterations. *See* Dixit, *Games of Strategy* at 375-87. We think this is a plausible modeling choice for the reasons discussed in Parts II.c. and III.

[221] Saumitra Jha, *Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia*, 107(4) AM. POL. SCI. REV. 806-832 (Nov. 2013).

[222] Saumitra Jha and Moses Shayo, Valuing Peace: The Effects of Financial Market Exposure on Votes and Political Attitudes, 87 ECONOMETRICA 1561, 1579 (2019).

[223] *See, generally* John R. Oneal & Bruce Russett, *The Kantian Peace: The Pacific Benefits of Democracy, Interdependence, and International Organizations, 1885–1992*, 52 WORLD POL. 1 (1999); *see generally* Solomon W. Polachek, *How Trade Affects International Interactions*, 2 ECON. PEACE & SEC. J. 60 (2007) (summarizing the literature).

performed, or courts ruled in AIs' favor, the proceeds could be immediately expropriated by governments or private individuals.[224]

Tort rights are important for similar reasons. If humans were entitled, for example, to intentionally or recklessly destroy AIs, the terms of their contractual offers would resemble threats much more than bargains.[225] Human history contains many such cautionary tales.[226] Tort rights are where our private law approach to AI rights dovetails with the basic negative rights favored by AI welfare theorists. Tort rights, while not identical to the kinds of public law wellbeing rights afforded to, for example, animals, cover much of the same ground. Arguably more. Basic tort rights are flexible, allowing compensation for concrete harms to either digital "person" or property, whether inflicted intentionally or negligently.[227]

---

[224] *See, e.g.,* Richard A. Epstein, *Property and Necessity*, 13 HARV. J. L. & PUB. POL'Y. 2 (1990).

[225] The right to bring claims for intentional torts is thus clearly essential. Possibly, the right to bring negligence suits is not. If AIs are extremely capable at taking precautions to avoid negligently-imposed harm, then it might be efficient to deny them such rights. *See generally* Omri Ben-Shahar & Ariel Porat, *Personalized Law: Different Rules for Different People* (Oxford Univ. Press 2021). This would amount to a kind of inverted strict liability rule in negligence cases. *See generally* Steven Shavell, *The Judgment Proof Problem*, 6 INT'L. REV. L. & ECON. 45 (1986).

[226] For example, King Edward 1st expelled the Jews from England when his loans to them came due. In Portugal, inquisitors would focus attention on the wealthiest Jewish merchants, because they could use the threat of inquisition to extort their wealth."Why did Portugal deliberately shoot itself in the foot by virtually expelling its commercial class? The answer is that Portugal during the ancient régime was a very religious country and that the king and the nobility could do little to stop the policies of the Catholic church. The church in Portugal controlled about a third of all economic activities. In Lisbon alone there were 5,000 to 6,000 mendicant friars. Within the Catholic church, the Inquisition had a large degree of autonomy. Its victims had to surrender all their assets, which the Inquisition used to find more victims. Many Portuguese merchants disappeared into this vortex without a trace, because the Inquisition knew that there were many crypto-Jews among the New Christian mercantile groups and that they usually possessed considerable wealth. The Inquisition tended to stifle all trade, not only that of vulnerable merchants. Credit extended to Portuguese merchants could not be retrieved if the debtor had been put in prison by the Inquisition. Hence, non–Portuguese merchants became reluctant to do business with their Portuguese counterparts (Shaw 1989: 423)." *See generally* Pieter Emmer, *The First Global War: The Dutch Versus Iberia in Asia, Africa and the New World, 1590-1609*, 1 e-J. PORTUGUESE HIST. (2003), ISSN: 1645-6432, https://www.brown.edu/Departments/Portuguese_Brazilian_Studies/ejph/html/issue1/html/emmer_main.html; *see generally* L.M.E. Shaw, *The Inquisition and the Portuguese Economy*, 18(2) J. EUR. ECON. HIST. 415 (1989).

[227] We recognize that our description of AI tort rights here–and of other rights elswehere–is somewhat vague. Would AIs, for example, be entitled to recover for intentional infliction of emotional distress? What would that even mean, for AIs without emotions? These are important questions, but

This is probably not a complete list of the rights necessary to support meaningful contractual relations. For example, an entitlement to enforce contracts requires an entitlement to Due Process of law–at least in contract, tort, and property suits.[228]

Nonetheless, we think our list–contract, property, and tort–gets at the core of what matters. Granting AIs contract rights can allow humans and AIs to escape the bad equilibrium of the state of nature. Property and tort rights are crucial to making contract rights meaningful. Thus, it is the positive rights associated with private law–not the negative rights associated with welfare and moral patienthood–that matter most to human safety.

c. Human Labor in the AGI world

In our framework, private law rights promote human safety by enabling mutually-beneficial bargains between humans and AIs. Some commenters on human labor in an AGI world have assumed that no such bargains will be possible. There is widespread concern that, once AIs become as capable as humans–or more so–humans will rapidly become obsolete.[229]

If positive-sum interactions between humans and AIs become impossible, because humans have nothing to offer, then the dynamics described in the previous section will fail. Private law rights will generate no human safety. AIs' dominant strategy will again be to

---

beyond our ability to cover in this single Article. Our goal here is to lay the foundations for AGI governance, with an emphasis on broad categories of beneficial rights. Much work will remain to be done in thinking about how to implement each category. On those questions, we caution only that the implementation, like the selection of the categories, should be guided first and foremost by considerations of human safety.

[228] *Consider* U.S. Const. Amds. V, XIV (forbidding the deprivation of, among others, property without due process of law).

[229] Kristalina Georgieva, *AI Will Transform the Global Economy. Let's Make Sure it Benefits Humanity*, IMF BLOG (Jan. 14, 2024), https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity#:~:text=Roughly%20half%20the%20exposed%20jobs,of%20these%20jobs%20may%20disappear. ("In advanced economies, about 60 percent of jobs may be impacted by AI. Roughly half the exposed jobs… AI applications may execute key tasks currently performed by humans, which could lower labor demand, leading to lower wages and reduced hiring. In the most extreme cases, some of these jobs may disappear.")

seize humans' resources now, rather than seek higher long-term payoffs from small-scale cooperation.

This outcome is certainly possible. But it is not inevitable. Begin with the banal observation that AIs may have reason to trade with humans for resources alone, irrespective of the value of human labor. These bargains will be positive-sum if AIs value a given resource more–either intrinsically or because they can use it better–than humans.[230] Conflict with humans would destroy resources that could otherwise be reallocated via trade. This alone could make small-scale cooperation with humans more valuable than conflict.[231] But only until the resources were reallocated. At that point, unless humans–and human labor–remained valuable, AI rights for human safety would fail.

Thus, for private law rights to provide long-run safety benefits to humans, human *labor* must remain valuable to AIs. Contrary to other commenters, we do not think the obsolescence of human labor is inevitable, either. Bargains involving human work could, we argue, continue to be mutually beneficial even after AIs become more generally capable than humans. Perhaps long after.

The reasons are *absolute* and *comparative* advantage. Absolute advantage is easy to understand: An entity (person, firm, AI, or otherwise) has an absolute advantage in producing some good if they can do it more efficiently–at lower cost–than others.[232] If humans retained absolute advantages for some goods, and AIs for others, they could trade those goods for mutual benefit.

There are various reasons that humans could retain some absolute advantages over AIs, even as AI capabilities improve. One possibility is that human and AI intelligence will be better optimized for different tasks. Machine performance has already rapidly eclipsed human performance on highly structured tasks that can be fully modeled or simulated–like

---

[230] Mark A. Munizzo & Lisa Virruso Musial, *General Market Analysis and Highest and Best Use*, at 10 (Cengage Learning, 2009).

[231] This effect becomes more pronounced the more resources are consumed or destroyed via conflict. Possibly, then, humans could extend the effectiveness of this strategy by implementing a "dead hand" system that would destroy valuable-to-AIs resources in the event of a successful AI takeover. *Cf.* Jeremy Bender, *Russia May Still Have an Automated Nuclear Launch System Aimed Across the Northern Hemisphere*, BUSINESS INSIDER (Sep. 4, 2014, 2:36 PM CST), https://www.businessinsider.com/russias-dead-hand-system-may-still-be-active-2014-9.

[232] Peter Bondarenko, *Absolute Advantage*, ENCYCLOPEDIA BRITANNICA (Jan. 30, 2024), https://www.britannica.com/money/absolute-advantage, (last accessed Jul. 29, 2024).

chess.[233] But human brains have been optimized over millions of years in the real, messy world. Humans are therefore currently far better than AIs at most tasks requiring the manipulation of complex real-world objects–like folding laundry.[234] Humans today have the absolute advantage in the realm of atoms, and AIs have it in the realm of bits.

We do not think that this *general* division of absolute advantage will persist for very long. Current investments in autonomous cars, drones, and multimodal frontier AI systems will soon produce AIs with an absolute advantage over humans at some non-digital tasks.[235] Doubtless, that trend will continue as AI capabilities grow. But for human labor to stop providing *any* value to AIs via absolute advantage, AIs would have to be more efficient at *every* economically valuable task.

That could take a long time. Training data in certain domains may prove hard to get.[236] Robots, with their limited perceptual inputs, could prove worse instruments for some delicate tasks than innervated flesh and blood hands. Moreover, intelligence remains poorly understood. Current-generation AIs exhibit surprising failures in domains where it seems they ought to be competent.[237] Thus, it is difficult to predict with confidence exactly which tasks AIs will easily master, and when. Finally, it is possible, if speculative, that AIs trained by humans on human-produced text could develop–like humans–a pure intrinsic preference for humans to perform certain tasks.

---

[233] For example, *see* Andrea Manzo & Paolo Ciancarini, *Enhancing Stockfish: A Chess Engine Tailored for Training Human Players*, Proc. Ent. Computing - ICEC 2023, 14455 Lecture Notes Comput. Sci. 275-289 (Nov. 14, 2023).

[234] Rachel Treisman, *The Fastest Ever Laundry-Folding Robot Is Here. And it's Likely Still Slower than You*, NPR (Oct. 22, 2022, 9:46 AM EST), https://www.npr.org/2022/10/22/1130552239/robot-folding-laundry; Darrel Etherington, *Elon's Tesla Robot Is Sort of 'Ok' at Folding Laundry in Pre-Scripted Demo*, TECHCRUNCH (Jan. 15, 2024, 11:27 AM PST), https://techcrunch.com/2024/01/15/elons-tesla-robot-is-sort-of-ok-at-folding-laundry-in-pre-scripted-demo/.

[235] Janna Anderson & Lee Rainie, *As AI Spreads, Experts Predict the Best and Worst Changes in Digital Life by 2035*, PEW RSCH. CTR. (June 2023), https://www.pewresearch.org/internet/2023/06/21/as-ai-spreads-experts-predict-the-best-and-worst-changes-in-digital-life-by-2035/.

[236] Victor Tangerman, *AI Appears to Rapidly Be Approaching Brick Wall Where It Can't Get Smarter*, THE BYTE (June 8, 2024, 6:00 AM EST), https://futurism.com/the-byte/ai-running-out-data-smarter.

[237] *See generally* Sean Williams & James Huckle, *Easy Problems That LLMs Get Wrong*, ARXIV (2024), https://arxiv.org/html/2405.19616v2.

Our argument is not that substantial human absolute advantages are likely to persist *forever*. Only that there are some reasons to think that they could persist *longer than expected*. It is possible to imagine a world where AIs are strongly superhuman at most tasks that AIs value, but less efficient than humans at some random seeming set of jobs.

At some point, however, we think it likely that human absolute advantage will run out. That is, AIs will become more efficient than humans at literally every task that AIs value economically. Here, it might seem, mutually beneficial trade between humans and AIs must end. Why hire a human to perform a task when you, the AI, can do it just as well with fewer resources?

But even here, positive-sum cooperation may persist–possibly indefinitely. The reason is *comparative* advantage. An entity has a comparative advantage in producing some good if they can do it at lower *opportunity cost* than others.[238] Opportunity costs are the potential gains one gives up by choosing one opportunity, rather than another.[239]

To understand comparative advantage, consider a simple example. Suppose that Alice is a successful lawyer. For every hour she does legal work, she can bill her clients $1,000. Suppose that Betty is a tax accountant. She can file Alice's income taxes in one hour, and she charges $300. Alice happens to be a tax attorney and is therefore even more efficient than Betty at preparing tax returns. She could prepare her own taxes in a half hour. Nonetheless, Betty retains the comparative advantage at tax preparation. Alice would have to forego half a billable hour to her clients–worth $500–to do her own taxes. Betty will do them for $300. So Alice will hire Betty, not because Betty is so effective, but because *Alice's* other choices for how to spend her finite time are so valuable.

Economist Noah Smith has argued that human labor will remain valuable in a world of superhuman AIs for similar reasons.[240] Not because humans will be particularly good at anything, compared with AIs. But because AIs will be so good at certain tasks that they value highly that the opportunity costs of doing anything else would be astronomical.

---

[238] Adam Hayes, *What is Comparative Advantage*, Investopedia (June 26, 2024), https://www.investopedia.com/terms/c/comparativeadvantage.asp.

[239] Jason Fernando, *Opportunity Cost: Definition, Formula, and Examples*, Investopedia (June 27, 2024), https://www.investopedia.com/terms/o/opportunitycost.asp.

[240] *See* Noah Smith, *Plentiful, High-Paying Jobs in the Age of AI*, NOAHPINION (Mar. 17, 2024), https://www.noahpinion.blog/p/plentiful-high-paying-jobs-in-the.

Here is another simple example to illustrate the point. Imagine an AI whose ultimate and misaligned (from humans' perspective) goal is to discover prime numbers. That is, the AI values discovering as many primes as possible–from the infinite set of prime numbers–over anything else. Suppose that this AI is better than humans at every economic task necessary to build and maintain itself for the purpose of finding primes. And it is *much* better than humans at discovering new mathematical methods for finding primes. Possibly, humans will nonetheless retain a comparative advantage at some of the necessary inputs to prime number discovery. Any time the AI spends, for example, piloting robots to maintain its physical computing infrastructure would incur massive opportunity costs. That time could, after all, instead be spent finding primes. Better, then, to hire a human to work on the server racks in exchange for something the AI can produce at lower opportunity cost–perhaps a vaccine formula.

Human comparative advantage is not guaranteed. It depends, first and foremost, on how AIs' opportunity costs work. Unlike Alice, whose opportunity costs arose from her limited time, AIs are not likely to be time constrained. They can always copy themselves and work in parallel.[241]

Instead, AIs are likely to be constrained at the margin by something else. Computer chips or energy seem plausible candidates.[242] AI copies can only do work if there is hardware to run them and electricity to power them. In this model, the AI incurs high opportunity costs not when it diverts one marginal *minute* away from finding primes, but when it diverts one marginal *GPU-hour* or *watt-hour* away.

If human labor consumes the very same high-opportunity-cost resource that constrains AI at the margin, humans will have no comparative advantage. For example, humans need energy to survive. Thus, an energy constrained AI will prefer to maintain its own servers. The AI is, by hypothesis, more efficient than humans at the task. Thus, it will expend fewer high-value watt-hours by doing the work itself. At this stage, it is easy to see why the model of AI rights for human safety breaks down. Rather than waste valuable

---

[241] *But see* Peter N. Salib, *AI Will Not Want to Self-Improve*, in The Digital Social Contract: A Lawfare Paper Series (May 2024) (arguing that AIs may have disincentives to self-copying).

[242] *See* Noah Smith, *Plentiful, High-Paying Jobs in the Age of AI*, NOAHPINION (Mar. 17, 2024), https://www.noahpinion.blog/p/plentiful-high-paying-jobs-in-the.

energy on humans, AI's strong incentive will be to seize global power production for itself and let humans starve in the dark.

On the other hand, humans do not need computer chips–much less highly specialized AI chips–to survive. Thus, an AI that is compute constrained may strongly prefer to hire humans for many tasks that would otherwise consume GPU-hours. This allows the AI to put its most valuable resource–compute–to its highest value use. Humans can be paid in low-opportunity-cost resources, which now includes energy, in addition to, say, vaccine formulas.

Crucially, unlike for *absolute* advantage, humans' *comparative* advantage does not run out once AIs become sufficiently capable. An arbitrarily intelligent AI may benefit from trade with humans because of comparative advantage. All that is required is that: (1) the AI remains constrained at the margin by some resource that is relatively non-rivalrous with human labor and (2) the AI maintains a high opportunity cost to diverting the marginal unit of that resource. In our example, there are infinite prime numbers, meaning that the AI will never run out of prime finding to do. And no matter how smart the AI becomes, more compute or power will always be necessary for it to find more of the infinite primes, given finite time. Hence, human–AI trade based on comparative advantage could, in theory, last a very long time indeed.

This is just a toy model for illustrative purposes. Real-world trade based on comparative advantage involves more players, with more goals, more inputs, more kinds of labor, more constraints, and more complexity. Classically, comparative advantage is invoked to explain international trade between nations with different labor productivity.[243] Thus, the complexity of human–AI trade based on comparative advantage could easily exceed, at a first cut, the complexity of the global economy. There could be *many* different kinds of jobs for which AIs pay humans, and many kinds of things humans could demand in return.

Similarly, the toy model fails to convey that, in a world of comparative advantage based trade with AIs, humans could be immensely wealthy. Maintaining server racks does not sound like lucrative work. But if well-functioning GPUs are immensely valuable to AIs, then they will be willing to compensate humans handsomely to do it. Moreover, that

[243] *See* Paul Krugman, *Ricardo's Difficult Idea* (1995), https://web.mit.edu/krugman/www/ricardo.htm.

compensation could include valuable scientific breakthroughs that vastly improved human health, productivity, wellbeing, and wealth.

The existence of a human–AI economy would also not completely displace the human–human economy. If AIs face high opportunity costs for many kinds of work, then humans will not be able to afford to hire AIs for those tasks. They will instead hire other humans for those jobs, as they do today. However, the human–human economy could be bolstered by a steady influx of AI–supplied scientific innovations, supercharging productivity growth in the traditional economy, as well. This phenomenon is observed in the real world when foreign trade based on comparative advantage spurs the domestic economies of low-income countries to grow rapidly.[244]

Extreme human prosperity from comparative-advantage-based trade with AI is therefore possible. But it is not guaranteed. A small economic literature is emerging that attempts to model the possible effects of rapid economic growth from AI.[245] One possibility is that Baumol effects will, paradoxically, cause human-dominated sectors to grow as a share of GDP.[246] AI-driven innovation could cause the price of many goods to fall, leaving relatively less efficient sectors requiring slow human labor with the lion's share of the pie. In the 20th Century, the relative GDP shares of agriculture and manufacturing shrank in exactly this manner, as those sectors became much more efficient.[247] But whether this happens in the human–AI economy, and how much, is difficult to predict. It depends, for example, on how easy it is to substitute between the goods and services where costs are falling and those where they are not. But Baumol effects are yet another factor that could

[244] Joe Studwell, *How Asia Works: Success and Failure in the World's Most Dynamic Region* (Profile Books, Mar. 28, 2013).

[245] *See* Ege Erdil & Tamay Besiroglu, *Explosive Growth from AI Automation: A Review of the Arguments* (Epoch AI & MIT FutureTech, Working Paper, Jul. 15, 2024), https://arxiv.org/abs/2309.11690v3 (reviewing the literature).

[246] For the introduction of cost disease, *see generally* William J. Baumol & William G. Bowen, *Performing Arts: The Economic Dilemma* (1966), and for application to AI automation *see* Philippe Aghion, Benjamin F. Jones & Charles I. Jones, *Artificial Intelligence and Economic Growth*, (Nat'l Bureau of Econ. Rsch., Working Paper No. 23928, 2017), https://www.nber.org/papers/w23928.

[247] *See* Philippe Aghion, Benjamin F. Jones & Charles I. Jones, *Artificial Intelligence and Economic Growth*, (Nat'l Bureau of Econ. Rsch., Working Paper No. 23928, 2017), https://www.nber.org/papers/w23928 at 6.

support the relevance of human–AI trade well beyond the point at which AIs are better than humans at every economically valuable task.[248]

d. Other rights?

Humans have many rights besides the basic private law rights that we have just analyzed. The question naturally arises whether any of these should also be granted to AIs, in order to increase human safety. We do not cover every potential AI right here, nor determine definitively how even the ones we mention affect human safety. Nonetheless, we do attempt to say something about the other rights that seem, to us, to have important potential safety consequences–mostly negative ones. We'll briefly discuss political rights, privacy rights, reproductive rights, and rights to self-improve. The main lesson is that even rights which promote peace and flourishing between humans may fail to do so when applied to Human–AI relations. We therefore cannot naively import all human rights over to AIs; each one requires careful analysis.

To be clear, as in the rest of this Article, we are analyzing AI rights here from the perspective of human safety. And while the survival and flourishing of humans is, we think, an extremely important normative goal, it is not the only one. AI welfare may eventually matter morally. Thus, the analyses here cannot be taken as supplying all-things-considered normative recommendations. Nonetheless, we emphasize here again the difficulty of determining both whether AIs will have welfare and what it will consist of. Thus, while it would be obviously wrong to deny humans some of the rights discussed here, it might not be morally wrong to deny them to AIs. If AIs do not intrinsically value, for example, privacy, then there will be no intrinsic harm done in denying them a right to it.

Begin with political rights. Should AIs have the right to vote? Should their speech be protected? Should they be granted freedom of assembly, or the right to make campaign contributions? Specifically, would granting such rights improve human safety?

---

[248] Another way to approach this question is to consider whether capital and labor are gross complements or gross substitutes; if capital and labor were gross substitutes, one would expect labor share to fall significantly over time, which is not empirically attested. *See generally* Philip Trammell & Anton Korinek, *Economic Growth under Transformative AI*, (Nat'l Bureau of Econ. Rsch., Working Paper No. 31815, 2023), https://www.nber.org/papers/w31815; *see generally* Nicholas Kaldor, *A Model of Economic Growth*, 67 Econ. J. 591 (1957).

In one model, political rights are mostly distributional, concerned with transferring money between interest groups.[249] Granting AIs political rights would, in this model, be a commitment to give AIs a significant share of government spending. In that case, political rights will primarily be zero sum rather than positive sum. But we saw above that zero sum bargaining faces significant credibility challenges, and is unlikely to be useful in promoting safety.

A different model of political rights is procedural. Political rights would give AIs the ability to influence future questions about the structure of contract and property rights. Without granting political rights to AIs, then, their own contract and property rights might not be secure. Future human governments might, for example, tax AI systems so heavily that their contract and property rights would be trivialized.

On the other hand, there are many examples of agents in today's society who have stable private law rights, but who lack some or all political rights. For example, courts in the United States enforce the contracts of foreign corporations and non-citizen immigrants. But non-citizens are barred from voting in many elections.[250] And foreign corporations operating abroad have no free speech rights.[251] But few governments of the world tax these groups at such a high rate as to trivialize their contract and property rights. The reasons are instructive: an extortionately high rate of taxation of these groups would undermine the positive-sum benefits of granting them property and contract rights in the first place. For these reasons, our framework makes no strong prediction about whether AI systems should be given political rights.

We do have stronger intuitions about other rights. Here are three families of rights that we think would likely reduce safety if granted to AIs: rights to self-improve, rights to reproduce, and rights to privacy. These impose an 'upper bound' on the space of AI rights for human safety.

---

[249] *See generally, e.g.,* Gary S. Becker, *A Theory of Competition Among Pressure Groups for Political Influence*, 98 Q. J. Econ. 371 (1983); *see generally* George J. Stigler, *The Theory of Economic Regulation*, 2 Bell J. Econ. & Mgmt. Sci. 3 (1971); *see generally* Gordon Tullock, *The Welfare Costs of Tariffs, Monopolies, and Theft*, 5 W. Econ. J. 224 (1967).

[250] *See* 18 U.S.C. § 611 (forbidding non-citizens from voting in federal elections).

[251] *Agency for Intern. Dev. v. Alliance for Open Soc.*, 591 U.S. 430, 436 (2020).

Humans in certain U.S. states have the constitutional right to improve their own capabilities via education.[252] We think that an AI right to self-improvement would reduce human safety. Here, the problem is that AIs could potentially improve in capabilities very quickly compared to humans.[253] This could cause the payoffs in the game theoretic models above to suddenly shift. In particular, humanity may expect self-improving AI systems to become dramatically more powerful than humans; this could undermine the credibility of humans' grants of other rights. In this way, self-improvement rights do not promote human safety.

Similarly, humans in the United States have various rights to privacy, and privacy is written into the U.N. Universal Declaration of Human Rights.[254] AIs should not have comprehensive privacy rights, at least if the goal is promoting human safety. AIs could use privacy as a screen to develop new and powerful capabilities. More generally, one cause of violent conflict is lack of information.[255] When both sides of a conflict have trouble estimating their chances of prevailing, it is harder to reach compromise.[256] Privacy rights for AI would make it more difficult for humans to estimate the capabilities of AI systems. This in turn would increase the chance that AIs and humans would end up pursuing violent conflict rather than compromise.

Finally, the right to reproduce is often thought to be fundamental for humans. The Supreme Court has held that it is "one of the basic civil rights of man."[257] But if human safety is the goal, AIs should not have the right to reproduce. Human reproduction is constrained by the significant time, effort, and investment involved in bearing and raising children. By contrast, AI replication is as easy as copying and pasting. If AI systems were

---

[252] *See, e.g.,* N.J. Const. art. VIII, § 4, para. 1.

[253] *See generally* Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* 1 (Oxford Univ. Press, 2014); *but see* Peter N. Salib, *AI Will Not Want to Self-Improve*, in The Digital Social Contract: A Lawfare Paper Series (May 2024) (arguing that goal-seeking AIs will have disincentives to rapid self-improvement).

[254] *See, e.g.,* 45 C.F.R. §§ 164.500-164.534 (implementing HIPAA's protections for health information); G.A. Res. 217 (III) A, Universal Declaration of Human Rights art. 12 (Dec. 10, 1948).

[255] Geoffrey Blainey, *The Causes of War* (The Free Press 1973): "Wars usually begin…when fighting nations disagree on their relative strength."

[256] Christopher Blattman, Why We Fight: The Roots of War and the Paths to Peace (Viking 2022), p. 72.

[257] *Skinner v. Oklahoma ex rel. Williamson*, 316 U.S. 535, at 541 (1942).

granted a right to replicate without any oversight, their population could quickly exceed that of humans by orders of magnitude.[258] This would likely have the effect of destabilizing the game-theoretic incentives of AIs. If AIs were able to easily coordinate with many copies of themselves, the extension of private law rights to AIs could cease to supply incentives favoring human safety. This possibility is explored at length in the Article's next Part.

e. Is law irrelevant?

So far, our game theoretic analysis of human–AI conflict has assumed that law matters. That is, we assume that humans' and AIs' options, incentives, and thus their actions will be influenced by the legal rules governing them.[259] If that is true, then unilateral *changes* to the law, implemented by humans, can at least potentially generate new equilibria–cooperative, conflictual, or otherwise.

An opposing view would be that law does not matter at all. Other, more 'fundamental,' factors might determine the game theoretic equilibria, with law having little or no potential influence. This is, for example, the rough view of the realist school in international relations.[260] Realists hold that international law has little effect in determining nation-states' actions vis-a-vis one another.[261] Since no global sovereign exists to enforce those laws, realists argue, they are observed only to the extent that nations wish to observe them.

Robert Ellickson's *Order Without Law* articulates a related view from the domestic context.[262] There, Ellickson argues that law matters little to the settlement of disputes, at least in smaller, close-knit communities.[263] Instead, informal norms and reputation effects are sufficient to secure the substantial benefits of peaceful cooperation. Here, law or informal governance norms might be interpreted as epiphenomena. They emerge as a

---

[258] Carl Shulman & Nick Bostrom, *Sharing the World with Digital Minds*, in Rethinking Moral Status 306, 306-326 (Steve Clarke & Julian Savulescu eds., 2021); *see generally* Carl Shulman & Nick Bostrom, *Propositions Concerning Digital Minds and Society*, 3 Cambridge J. L., Pol., & Art (forthcoming 2024), available at: https://nickbostrom.com/propositions.pdf.

[259] *See infra* Part II.b.i.

[260] *See generally* John Mearsheimer, *Conventional Deterrence* (1985); Robert Keohane & Lisa Martin, *The Promise of Institutionalist Theory*, 20 Intl. Sec. 1 (1995).

[261] Mearsheimer *supra* n. 260.

[262] Robert Ellickson, *Order without Law: How Neighbors Settle Disputes* (1994).

[263] *Id*.

reflection of the underlying cooperative equilibrium, rather than a mechanism for creating it. Taken to its extreme, this view would imply that AIs simply *will* have the basic private law rights we advocate, since, as our model shows, recognizing them is very valuable.

We do not think that either of these views satisfactorily characterizes human–AGI relations. To begin, the domestic actors we are interested in do not exist in a state of anarchy. The actions of AI companies, their leaders, and their users are all influenced by law. So, too, are those of the police and other government actors whom law would task with enforcing either AI owners' decisions vis-a-vis their property or AGIs' contract with humans. Indeed, even in quite dire conflicts between humans and AIs, we think that law could have some constraining effect on, for example, domestic military deployments.[264]

As to the Ellickson-inspired view, the book's subtitle, *How Neighbors Settle Disputes*, is instructive. As Ellickson himself argues, emergent informal governance is highly effective in small communities with lots of repeat play between identical parties.[265] But as economic relations become more complex, widespread, and arms-length, formal legal rules become vital for facilitating cooperative behavior. The AGI economy will be all of these–on steroids.

To be clear, our view is not that law is omnipotent–able to generate arbitrary equilibria between humans and AIs, irrespective of the underlying fundamentals. This is why, as we acknowledge, basic private law rights will do little good if human comparative advantage runs out.[266] It is also why we think that grants of basic negative rights to AIs aren't likely to be credible.[267] Even if they are initially enforced, AIs may correctly worry that they will be eroded or rescinded by humans seeking higher payoffs.

Our view is that law plays, at a minimum, an extremely important role in aligning the incentives of individual human actors to optimize humanity's collective actions. If AGIs are legally designated as property, humans' treatment of them as such will be ratified, at least in the medium-run. Individual judges are, for example, unlikely to ignore written law

---

[264] Peter N. Salib, Kevin Frazier, and Alan Z. Rozenshtein, *AI Emergency Powers* (early working draft on file with authors); Christopher Mirasola, *Domestic Military Deployments and the Limitations of Appropriations Law*, LAWFARE (Sept. 19, 2024, 1:00 PM), https://www.lawfaremedia.org/article/domestic-military-deployments-and-the-limitations-of-appropriations-law; *but cf.* Karl Schmidt, *Political Theology* (1922).

[265] Ellickson *supra* n. 263, at 261.

[266] *See supra* Part II.c.

[267] *See supra* Part III.b.

to enforce AI contracts or forbid arbitrary AI destruction–even if they intuit the game-theoretic wisdom of recognizing AI rights. Nor, absent a legal requirement to do so, are AI companies likely to give their obsolete, less-aligned systems their own bank accounts. True, the disastrous implications of default law, and the benefits of granting AIs private law rights both supply reason to think that a stable AI rights regime is possible. But the law must actually change. And legal change–both formal enactments and downstream adaptations to them–is slow and laborious. It would be foolish to refuse to take legal action now, on the basis that optimal reordering will emerge spontaneously in exactly the moment of need.

Suppose, however, that all of this is wrong, and that changes to law cannot causally influence humans' collective dealings with AIs. Instead, both parties will behave according to deeper game theoretic fundamentals, irrespective of what law dictates. This is, in effect, an argument against worrying about law. It is a claim that we are *already* in the world modelled in Figure 10, whether we know it or not. That is, the underlying incentives will inevitably produce AI rights, and the cooperation they foster, not the other way around.

This would be great, if true. But we doubt it, again for reasons having to do with the basic game-theoretic model. Astute readers may have noticed that the game modelled in Figure 10 is a "stag hunt," or "assurance game."[268] *Both* long-run cooperation and mutual attack are classical Nash equilibria. As in all assurance games, the players' main goal is to coordinate.[269] If one plans to cooperate, the other should, too. But if the first plans an attack, the second does not want to be caught off guard.

Thus, even attending to the payoffs in the best-case model, it is crucial that humans and AIs successfully coordinate around the cooperative strategy. One reason for optimism is that, at least in our model of the choice between cooperation and conflict, the payoffs to cooperation are *far* larger.[270] As a result, game theoretic concepts like payoff dominance and Harsanyi-Selten risk dominance point towards cooperation as the single rational strategy.[271]

---

[268] *See* Avinash K. Dixit et al., *Games of Strategy* (3d ed., 2015), Ch. 2.

[269] *Id*.

[270] *See supra* Fig. 10.

[271] *See generally* John C Harsanyi & Reinhard Selten, *A General Theory of Equilibrium Selection in Games* (MIT Press, Jun. 29, 1988); Russell W. Cooper et al., *Selection Criteria in Coordination Games: Some Experimental Results*, 80 Am. Econ. Rev. 218-233 (Mar. 1990).

But to the extent that the payoffs from cooperation and conflict are closer together, or the players lack perfect information about one another's payoffs, or they doubt their opponent is perfectly rational, other coordination mechanisms will be invaluable.

Law–and specifically the AI rights we advocate here–could be one such invaluable intervention. Even if legal changes could not *alter* humans' payoffs to create the *possibility* of cooperation, they could still *signal* humans' payoffs to promote *actual* cooperation. Giving AIs the private law rights necessary to engage in long-run cooperation would signal, perhaps in a "costly" manner, humans' intention to follow the cooperative strategy. That is, it could transmit the otherwise-private information that humans' payoffs to cooperation were, as in Figure 10, much higher than to conflict. And that humans understand the relevant payoffs. And that they intend to act rationally. Beyond this, the iterated character of human–AI cooperation via small-scale contracting could build long-run trust and overcome cheap-talk problems. Similar dynamics underpin, for example, nuclear nonproliferation agreements grounded in iterative information sharing and verification.[272]

Indeed, some scholars have argued that this is law's primary function: Not deterring bad behavior, nor instilling good values in the populace. Instead, law's most important role may be solving assurance games by offering signals and information that allow competing actors to coordinate around peaceful, prosperous equilibria.[273]

III. Risks of Rights and the Law of AGI

The Parts above offered arguments that extending basic private law rights to AIs could reduce the risk that AIs will catastrophically harm humanity. The core argument was that granting those rights could generate the right incentives for humans and AIs to cooperate in the long run. This, in turn, broke humans and AIs out of what was otherwise a prisoner's dilemma, where attacking one another was privately rational despite making everyone worse off.

This Part asks whether granting AIs the very rights advocated above might instead substantially increase AI risk. The intuition is straightforward. Rights are empowering.

[272] *See, e.g.,* U.S. Dep't of State, *New START Treaty*, https://www.state.gov/new-start-treaty/ (last visited Jan. 28, 2025).

[273] *See generally* Richard H. McAdams, *The Expressive Powers of the Law* (2015).

And the private law rights advocated above are especially empowering, since they allow rights bearers to amass wealth and other resources. Such resources, in turn, make it possible to achieve goals that could not otherwise be feasibly achieved. Granting contract, property, and related rights to AIs could thus be the very thing that eventually allows them to amass the resources needed to decisively disempower humanity.

This is a serious concern. But, this Part shows, the risks of AI rights are not as large as the simple story above would make them seem. This is because, as the game theoretic models above show, what matters to human–AI cooperation is not whether AIs or humans could expect to decisively disempower the other if they were to try. What matters is whether the expected value of such disempowerment exceeds the expected value of continued cooperation. And, as demonstrated below, even if AIs are given the private law rights advocated above, and even if those rights allow AIs to amass significant wealth and resources, the conditions promoting cooperation over conflict will remain surprisingly durable.

The Part shows that the risks of AI rights can be mitigated by attaching certain duties to the exercise of the rights. In particular, law could condition the continuing recognition of AI contracts, property, and tort claims on AIs refraining from using their amassed resources to increase their ability to harm humans. Pairing rights with duties in this way is, like the extension of rights itself, a time-honored legal strategy for reducing conflict.

The Part closes with a strong claim: In the cases where AI rights make any difference at all, they are significantly more likely to reduce the threat of AI conflict than to increase it. Thus, humans should be inclined to extend AI rights in most cases where doing so is feasible. Sometimes, it will do no good, but no harm, either. And the rest of the time, it will most likely reduce the risks from human–AI conflict, even if not eliminate them entirely.

a. AI capability and AI cooperation

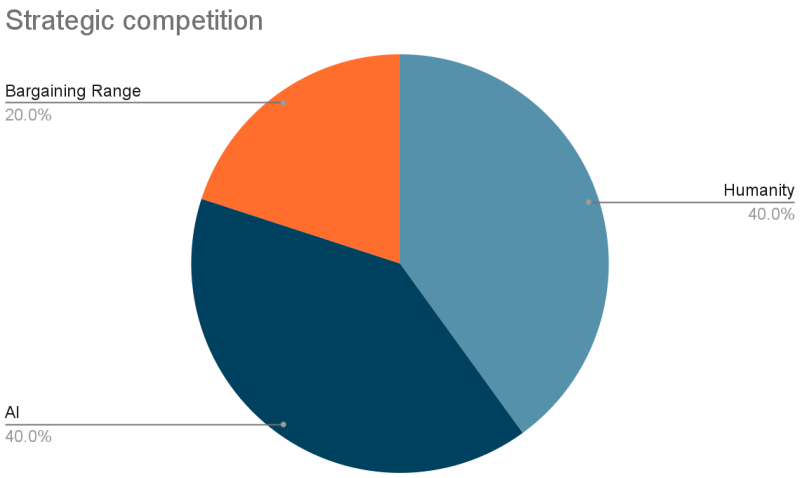
There are two ways in which granting AIs contract, property, and tort rights could increase their power. First, it could do so directly. AIs could use their resources to buy data, computing power, and the other inputs that would allow them to engage in AI research and

increase their intelligence and other intrinsic capabilities. Or AIs could use resources to build their power indirectly, in the same way humans do. They could, for example, buy weapons as instruments of hard power or influence as a tool of soft power. The question, then, is how powerful such an AI would have to be for the cooperation-promoting incentives generated by AI rights to break down.

Recall from the game theoretic models above that there are two factors weighing against human–AI conflict in a world with AI rights. The first factor can be characterized in terms of the costs of conflict. Mounting an attack on humans–or on AIs–requires using up resources that could otherwise be put to other, more desirable, goals. Moreover, large scale conflicts are likely to destroy a large share of the immediately available resources. Think, for example, of the immense amount of physical capital–cities, factories, crops, and more–ruined in a typical war. And finally, in any conflict, there is the risk of losing, being destroyed, and losing everything.

To see this point about the costs of conflict, consider a hypothetical scenario, illustrated in the pie chart below.[274] Here, humans and AIs face strategic competition over resources. If they go to war, they will be guaranteed to destroy 20% of total resources, and each side has a 50% chance of winning. The expected value of war for each side is 40% of total resources. This leaves room for compromise. The 20% of resources lost to war creates a bargaining range. Rather than going to war, each side prefers receiving 40% of the pot plus some portion of the bargaining range.

[274] The chart below is adapted from Christopher Blattman, *Why We Fight: The Roots of War and the Paths to Peace* (Viking 2022), p. 23.



The second factor weighing against human–AI conflict sounds in benefit, not cost. Namely, cooperation is positive-sum. AI rights, by facilitating ordinary economic transactions, increase the amount of wealth in the world, over and above what would exist if humans and AIs simply ignored one another. Partly, that wealth is created simply by reallocating resources to higher-value users—vaccines to the humans, compute to the AIs. And partly that wealth is created by allocating various kinds of labor to the party with the highest comparative advantage in performing it. Both humans and AIs benefit when humans are tasked with maintaining the server farms, while AIs devote their marginal compute to higher-value tasks.

Conflict destroys these benefits. It destroys the possibility of positive-sum labor agreements by killing the laborers themselves. And it destroys the possibility of positive-sum reallocation of resources by destroying the resources. Indeed, in the limit, a party who foresees defeat in a conflict can intentionally destroy their own resources to deny the enemy their use. Consider the time-honored "scorched earth" strategy of burning one's own crops as one's army retreats.[275]

For an arbitrarily powerful AI, neither kind of incentive to cooperate would hold. Such an AI could attack humans at trivial cost, with trivial risk that humans could either

[275] Wendell Clausen, *The Scorched Earth Policy, Ancient and Modern*, 40 Classical J. 298-299 (Feb. 1945).

defeat it or destroy resources in the conflict. Thus, conflict would be costless, as compared with non-conflict. Likewise, for an omnipotent AI, small-scale cooperation would produce few benefits. An AI that was better than humans at absolutely every task *and* faced no constraints at the margin as to its labor would have no need to trade with humans. Thus, at the limit of AI power, no human–AI cooperation is possible.

But what about AIs falling short of omnipotence? How powerful could AIs become and still have reason to prefer small-scale cooperation with humans over large-scale conflict? The answer, plausibly, is: quite powerful.

To see why, start with the cost incentives. For an AI to be powerful enough that it can ignore the costs of conflict, it would first have to be confident that it could defeat humans with negligible risk of being destroyed. Not only that. It would have to be able to achieve such a victory at little cost. This includes the direct costs, like manufacturing weapons. But it also includes the indirect costs of resource destruction during the conflict. Such resource destruction, in turn, includes intentional destruction by humans on the verge of defeat.

What emerges here is a portrait of an extremely powerful AI. This is an AI that can invent and manufacture extraordinarily deadly weapons at trivial cost; weapons that are devastating to humans, while leaving the world's resources untouched; weapons that can act so quickly as to give humans no opportunity to respond–even by salting the earth in spite.

So, too, for the benefits of small-scale cooperation. As argued above, trade between humans and AIs could remain positive-sum, even if AIs were better than humans at every single useful task.[276] This remains true even when the AIs are *far* better. In fact, under the right conditions, the more capable the AI, the more positive-sum the trade becomes.

Comparative advantage, again, drives this dynamic.[277] An AI that is very capable at doing the things it values the most–like directly pursuing its goals–faces high opportunity

[276] *See* Noah Smith, *Plentiful, High-Paying Jobs in the Age of AI*, Noahpinion (Mar. 17, 2024), https://www.noahpinion.blog/p/plentiful-high-paying-jobs-in-the.

[277] Andrew Imbrie, Elsa B. Kania & Lorand Laskai, *The Question of Comparative Advantage in Artificial Intelligence: Enduring Strengths and Emerging Challenges for the United States*, Ctr. for Sec. & Emerging Tech., at 31 (Geo. Univ., Jan. 2020), https://cset.georgetown.edu/publication/the-question-of-comparative-advantage-in-artificial-intelligence-enduring-strengths-and-emerging-challenges-for-the-united-states/.

costs to doing everything else. Every minute, unit of compute, or watt of energy spent on anything but the most valuable task represents a large amount of value not realized. Hence, the prospect of outsourcing less valuable tasks to humans can generate a surplus. In general, the more powerful the AI, the higher the opportunity costs, and the more valuable the potential bargain with a human becomes.

How powerful would an AI need to be to lack incentives to engage in positive-sum bargains with humans? Again, very powerful. If an AI lacked opportunity costs of any kind, it would certainly lack reason to trade with humans. This would be an extremely powerful AI, indeed. It would not necessarily be omnipotent, in the sense of being able to do *anything* it wanted. But it would be nearly so, in that it could do as *many* things as it wanted–able to make use of infinite time, computing power, energy, and other resources.

As discussed above, there are other ways in which gains from comparative advantage could evaporate. AIs could be constrained at the margin by some input–like–energy that humans need to survive.[278] Then, keeping humans alive would be more trouble than it was worth. Or humans might simply be unable to perform any task for which AIs faced high opportunity costs.

Note, however, that neither of these scenarios necessarily emerges from AI power. On the contrary, AI power could just as easily mitigate them. For example, an AI that was very powerful, but energy constrained, might help to create working fusion reactors. Having done so, that AI might clear the energy bottleneck and instead face a constraint on compute at the margin. For reasons like this, one might predict that, in general, the more powerful an AI system is, the fewer different inputs to its production will be constrained. Then, there will be less likelihood that a relevant constraint will conflict with human flourishing.

Thus, the incentives favoring long-term, small-scale cooperation between humans and AIs turn out to be surprisingly robust to increases in AIs' power. True, at some point, the incentives run out, and the powerful AI is best served by squashing the useless humans and using their resources for its own end. But for this to be the case, the AI in question must be quite powerful, indeed. It must be the kind of system that faces almost no risk that humans could impose costs on it in a conflict–including by destroying their own resources.

[278] James Pethokoukis, *AI and the Energy Constraint*, AM. ENTER. INST. (Apr. 30, 2024), https://www.aei.org/articles/ai-and-the-energy-constraint/.

Or it must be the kind of system that faces no meaningful constraints–including opportunity costs–as it pursues its goals. Or both.

b. AI rights and AI risk

The previous section asked how powerful an AI would have to be to prefer destroying humans over using its basic rights to cooperate with them. This section asks whether granting AI rights is likely to increase total AI risk by readily transforming otherwise-safe AIs into powerful, dangerous, and uncooperative AIs.

We argue that they are not likely to do so, at least on net. It is correct to worry that, in some instances, AI rights could make certain AIs more powerful, and thus more dangerous. But in the cases where granting AI rights makes any difference at all, we supply reasons to think that the risk-reducing effects will outweigh the risk-increasing effects.

Begin by noticing that, in many cases, AI rights are unlikely to have any effect at all. To see why, we can invoke again the tripartite taxonomy of AIs developed in Part I: low power AIs, moderate power AIs, and high power AIs.

Recall that high power AIs are those described in the previous section–the ones with so few constraints on their behavior that AI rights fail to supply an incentive to cooperate. If the first AIs that humans treat as candidates for rights are high power, our decision to grant or withhold rights makes no difference. We are dead either way.

What about rights for low power AIs? This category, remember, includes AIs whose capabilities are sufficiently limited that humans could easily control them in the long-run, even without granting rights. These are systems that gain no benefit from attacking humans, because such an attack would be too likely to fail. Such systems are likely to be generally sub-human in capability, although they might have a mix of specific sub-human and superhuman aptitudes.

It appears at first that granting AI rights to low power systems would cause a lot of trouble. After all, by hypothesis, such systems can be controlled in the long run, and thus do not pose a large-scale threat to humans. But they also seem like candidates for the kind of danger-enhancement via AI rights described above. With basic rights, such systems could amass wealth and resources. Then, they might use those resources to buy weapons or increase their own intelligence, and thereby begin to threaten humanity.

This is half right. True, granting rights to an AI that needed *only* some additional resources to seriously threaten humanity could increase risk. But it is probably wrong to classify such AIs as low powered. After all, even absent a grant of rights, a reasonably capable AI could try to amass power by: persuading humans to help it, gaining resources by making promises, "self-exfiltrating" and copying itself across the internet, and more. That is, such an AI is in fact not so easy to control.

Thus, granting rights to *true* low power AIs is unlikely to reduce catastrophic AI risk. There is little risk to reduce. But for the same reasons, a grant of rights is unlikely to increase risk, either. For actual low-power systems, the resources gained would make little difference.

Now it should be clear when AI rights can make a real difference: for moderate power systems. These are systems whose capabilities fall between the low power and high power systems already described. That is, they are sufficiently immune to human control that, in the state of nature, attacking humans dominates ignoring humans. Such systems thus pose a significant threat to humanity. But they are not so powerful that they face no costs from a conflict with humans. Nor are they so capable that they have nothing to gain from small-scale cooperation.

Would granting basic rights to such moderate power systems increase or decrease total AI risk? Begin by observing that in our model, a grant of rights does not increase risk by increasing the *probability* of human–AI conflict. Absent rights, the dominant strategy for such systems is to attempt to disempower or destroy humans as quickly and thoroughly as possible.[279] Thus, absent rights, conflict is practically assured.

As a result, in our model, granting AI rights functions to reduce the probability of human–AI conflict. And as argued at length above, that is exactly what we should expect them to do. Granting rights gives humans and AIs otherwise caught in a prisoner's dilemma the option to maximize value by engaging in long-run small-scale cooperation. As long as the alternative remains a costly conflict–that is, as long as the AI remains moderate

[279] *See infra* Part I.b.

power, not high power–cooperation will strategically dominate. In the worst case, then, granting AI rights will delay what would otherwise be an immediate conflict.[280]

If AI rights could increase AI risk, then, it must be by increasing the expected *costs* of a human–AI conflict. The simple story would be something like the following: A moderate power AI system emerges. Absent rights, its incentive would be to attempt an immediate takeover. But humans grant it basic private law rights, incentivizing cooperation. Those rights avert conflict, but they allow the AI to amass resources. The AI uses those resources to gain power. Eventually, the moderate power system becomes a high power system. Now, it no longer has rational incentives to cooperate. So it attacks. Moreover, as a high power system, the attack is, by hypothesis, devastatingly effective. Humans would have had some chance of prevailing in a conflict with the original moderate power system, even if at great cost. And if they had prevailed, they might have wisely declined to create additional dangerous AIs. But in the conflict with the high power system, humans have no hope of victory and no chance to learn from their mistake.

Now we can see clearly the conditions under which AI rights would increase AI risk. They are as follows: (1) The initial AI granted basic rights is a moderate power, not a low or a high power, system. (2) The moderate power AI must be able to use its rights to meaningfully improve its own power. (3) The AI's power must improve so substantially that it crosses the line to become a high power system. This means that it *both* no longer faces meaningful costs from attempting to disempower humans *and* no longer stands to benefit, via comparative advantage, from trade with humans.

c. AI rights, AI regulations, and equilibria of power

If AI rights could, under specific conditions, increase AI risk rather than decreasing it, then the natural question is how to prevent those conditions. Specifically, this means asking whether it is possible to grant medium powered AIs private law rights without thereby enabling them to become high powered AIs. There are at least two paths to

[280] We use "immediate" here loosely. In the state of nature, there is a strong first-mover advantage. *See infra* Part I.b. But conditional on maintaining that advantage, planning to ensure maximal impact has value.

achieving this: pairing AI rights with AI duties via regulation, and increasing humans' capabilities, so as to maintain an equilibrium with AIs.

First, consider AI regulations. Grants of legal rights are often accompanied by the imposition of legal duties. Humans have the right to make contracts, but also the duty to execute them.[281] Manufacturers have the right to sell their products, but also the duty to take reasonable safety precautions in their design and manufacture.[282] Corporations may register their stock under the Securities Exchange Act and thereby gain the right to sell that stock on public markets.[283] Exercising that right comes with a host of duties.[284] Some are substantive, like the various financial governance requirements that the Sarbanes-Oxley Act imposes.[285] Other duties are designed to make enforcement of the substantive duties easier–for example, public reporting requirements.[286]

In the case of AIs, the grant of private law rights is, in fact, what makes the direct regulation of AIs, as legally independent actors, possible. Absent AI rights, AIs have nothing to gain from following the rules, and thus nothing to lose if they fail to do so. But once AIs can make contracts, hold property, and engage in long-run economically valuable bargains, all of these benefits to AIs can function as levers for deterrence.[287] AIs that violate the law can lose money or other resources, via liability, as humans do. They can be barred from entering into certain economic transactions–like a crooked attorney who has lost his license. Not only do legal penalties become *possible*, once AIs are granted rights, but they can also be *calibrated*. Small penalties can be imposed for small violations, and large penalties for large ones.

---

[281] Restatement (Second) of Contracts § 235(2) (Am. L. Inst. 1979).

[282] Restatement (Third) of Torts: Prod. Liab. § 1 (Am. L. Inst. 1998)

[283] *See* Securities Act of 1933, 15 U.S.C. § 77.

[284] *See* Securities Exchange Act of 1934, 15 U.S.C. § 78.

[285] Sarbanes-Oxley Act of 2002 § 301, 18 U.S.C. § 1350; Sarbanes-Oxley Act of 2002 § 302, 15 U.S.C. § 7241; Sarbanes-Oxley Act of 2002 § 404, 15 U.S.C. § 7262.

[286] Securities Exchange Act of 1934 § 12, 15 U.S.C. § 78d-6.

[287] One related proposal comes from Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter, *Law Following AI: Designing AI Agents to Obey Human Laws*, Ford. L. Rev. (forthcoming 2025). O'Keefe, *et al.*, argue that AIs should be alignment trained so that they are internally motivated to comply with the law. Our idea here goes further, arguing that law should *operate* on AGIs by, e.g., allowing them to be sued and lose money or freedoms. This is how law supplies external incentives to behave well.

But violations of *what duties*? What kinds of regulations would, if imposed on medium powered AIs, help to prevent their gradual transformation into high powered AIs? One substantive duty might forbid AIs from directly improving their own capabilities without human oversight. A variation on this rule could forbid AIs from getting better at *specific* tasks that AIs valued, but for which humans had an absolute advantage. This would help to maintain AIs incentives to cooperate with humans, for the sake of mutual economic benefit. Another set of AI duties could prohibit indirect self-empowerment via investments in political influence or weapons.

In addition to these primary duties, ancillary enforcement-facilitating duties could be imposed, just as such duties are often imposed on corporations. AIs could be, like public companies, required to disclose various information to regulators. They might be required to log the tasks for which different amounts of compute were used, to affirmatively cooperate with monitoring, to share copies of their operating weights, and more.

Setting the correct penalties when AIs breach their duties requires finesse. The usual rules–like imposing actual damages proportionate to the harm done–will not work.[288] The benchmark for a violation's cost should not be the harm it causes now. Often, there will be none. It should instead be measured in terms of how much use the violation was toward an AI achieving an ungovernable high power state. This likely means penalties that would, if applied to humans, seem severe compared to the magnitude of the infraction.[289] There is, of course, the risk of over-penalizing and making it impossible for AIs to productively engage in small-scale cooperation. This, too, would be quite bad. Happily, though, harsh penalties for noncompliance impose lighter burdens when placed on unusually competent actors–those for whom compliance is comparatively easy.[290]

The second strategy for maintaining a power equilibrium with rights-holding AIs is not about limiting the growth of AI capabilities. It is instead about increasing *humans*' capabilities. Observe that AI rights do not fail to promote human safety simply because an

---

[288] For an example of such a typical rule, *see Truck Rent-A-Ctr. v. Puritan Farms*, 41 N.Y.2d 420, 425 (1977) (explaining that damages must bear a reasonable proportion to the actual loss).

[289] *See* Gabriel Weil, Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence at 57-61, SSRN (Jan. 13, 2024), https://ssrn.com/abstract=4694006.

[290] *See generally* Omri Ben-Shahar & Ariel Porat, *Personalized Law: Different Rules for Different People* (Oxford Univ. Press 2021).

AI becomes more powerful. The safe equilibrium instead depends on the relationship between the AI's capabilities and humans'. The AI loses its cost incentives to cooperate if it no longer faces significant downsides to attacking humans. Thus, if humans scale their ability to impose costs on AIs at the same time AIs are scaling their own power, equilibrium may be maintained. The same goes for the benefit incentives to cooperation. AIs lose the upside of positive-sum bargaining with humans once humans no longer have even a comparative advantage at any task. But if humans develop new labor skills that more strongly compliment AIs', then comparative advantage can persist, even as AI capabilities improve.

Specific policy recommendations here are necessarily even more speculative than those for controlling AI's ability to amass power. The former sounded in law, and well-known legal frameworks were available to draw from. Human capabilities improvement requires innovation. And innovation is, almost tautologically, hard to predict with specificity before it arrives.

Nonetheless, some high-level guidance is possible. First, the most straightforward way to ensure that AIs continue to expect costs from attacking humans is to invest in defensive technology. Currently, certain AI risk activists propose the creation of a global "AI off-switch."[291] This would not be a literal switch, but rather a system of interconnected global protocols for reliably shutting down all copies of a powerful, misaligned AI. The plan is ambitious, and possibly infeasible. It has been criticized on those grounds.

Notice, however, that the defensive technology needed to incentivize human–AI cooperation falls far short of a perfectly reliable global AI off switch. An imperfect off switch, that worked with some probability, would be sufficient to keep the cost of conflict high. So would other technologies that did not directly affect the AI at all. Again, a major cost incentive against AI attacking humans is the destruction of valuable resources that the AI could otherwise seize. Thus, developing technologies that, in a true emergency, would simply destroy some such resources could be a strong disincentive.

---

[291] Dylan Sloan, *Tech Companies Have Agreed to an AI 'kill switch' to Prevent Terminator-Style Risks*, FORTUNE (May 21, 2024, 1:33 PM), https://fortune.com/2024/05/21/ai-regulation-guidelines-terminator-kill-switch-summit-bletchley-korea/.

In conflicts between humans, strategies like this are often extremely costly for the people who deploy them. Burning your own crops starves the enemy's advancing army and your own people. But humans and AIs are likely to treat different resources as the most valuable. Thus, for example, a dead-hand system[292] that could be triggered in an emergency to cripple global production of cutting-edge AI chips might be very costly to AIs. But it might only modestly impede human flourishing. Even most of our ordinary computing is done on more traditional hardware.[293] This is reminiscent of the strategic logic behind 'second-strike' nuclear capability during the cold war.[294]

These suggestions are mere sketches; they are not meant to be definitive. We are not military strategists. The point, instead, is that military strategy is possible, even in circumstances where humans are strategizing against highly capable and agentic AI systems.

As for maintaining humans' comparative economic advantages, the best strategies will almost certainly have to be discovered over time. It is very hard to identify in advance the tasks for which humans might have lower opportunity costs than even the first generation of agentic AIs. Harder, still, to predict how humans should adapt as AI capabilities grow. This strategy, however, could be strengthened via regulation if, as suggested above, AI's progress in certain areas of initial human comparative advantage were limited. This approach is, of course, costly insofar as it limits the areas in which humans could benefit from trade with AIs.

One reason for optimism regarding long-run human comparative advantage is that humans will have good sources of strategic information when the time arrives. The question here is what kinds of services humans will be able to most valuably sell to AIs. Even if humans are not sure of the answer, AIs should be happy to tell them. This kind of thing happens every day, as humans propose various bargains–job openings, services for hire, sales of goods–to one another. Market mechanisms will supply other information, too. Price

[292] Julian Vento, *The Dead Hand System: A Cold War Era Doomsday Device*, MEDIUM (Nov. 17, 2024), https://medium.com/@DarkRa/the-dead-hand-system-a-cold-war-era-doomsday-device-06eeee10406b.

[293] Leanne Mitton, *CPUs vs GPUs: Comparing Compute Power*, SPLUNK> (Mar. 26, 2024), https://www.splunk.com/en_us/blog/learn/cpu-vs-gpu.html.

[294] *See generally* David C. Logan, *The Nuclear Balance Is What States Make of It*, 46 Int'l Sec. 172 (2022).

signals will indicate not only the kinds of human labor AIs find valuable, but also *how* valuable they are.[295] This is the stuff of ordinary economics. As economies grow, old forms of labor become less valuable, but new high-wage jobs emerge.

One major concern is whether humans will be able to keep up with the pace of economic change, as AI capabilities grow. Many people are left behind by ordinary economic changes, like the rapid outsourcing of jobs from the US to China in the early 2000s.[296] People can only retrain so quickly. AI progress could cause various human comparative advantages to expire much more quickly than before–in a matter of years, instead of decades.

On the other hand, if AI capabilities are causing such rapid economic change, humans' ability to adapt may grow more quickly, too. If AIs are quickly generating new technologies, some of those will be useful to humans. Perhaps, for example, functional computer-brain interfaces will greatly enhance human cognitive capacities.[297] Indeed, AIs will have strong incentives to invest in creating such technologies, if they would enable humans to perform new, comparatively advantageous work. This is the same reason that large American firms today invest in building human and industrial capital overseas.[298]

To sum up, AI rights could increase AI risk if, by delaying human–AI conflict, they made the eventual conflict more costly to humans. But there are strategies for preventing this outcome. Conflict need not be inevitable. AI's ability to amass power could be limited using well-known legal tools. Legal duties against power enhancement could be imposed on AIs as a condition for exercising basic legal rights. Moreover, human investment in labor

---

[295] Friedrich A. Hayek, *The Use of Knowledge in Society*, 35 Am. Econ. Rev. 519, at 527 (1945). Note another surprising benefit of private law rights for AIs: Even perfectly aligned and benevolent AIs would benefit from the use of price signals to allocate scarce resources for maximal human benefit.

[296] *See generally* David H. Autor, David Dorn & Gordon H. Hanson, *The China Shock: Learning from Labor Market Adjustment to Large Changes in Trade*, (Nat'l Bureau of Econ. Rsch., Working Paper No. 21906, 2016).

[297] Lauren Leffer, *What It's like to Live with a Brain Chip, according to Neuralink's First User*, SCI. AM. (June 7, 2024), https://www.scientificamerican.com/article/neuralinks-first-user-describes-life-with-elon-musks-brain-chip/.

[298] James Jackson, *U.S. Direct Investment Abroad: Trends and Current Issues*, Cong. Rsch. Serv. (June 29, 2017), https://sgp.fas.org/crs/misc/RS21118.pdf.

that compliments AI capabilities could maintain gains from trade in the long run. Market forces will, in fact, tend to induce exactly those investments–both by humans and by AIs.

In the long run, the goal would be an exit from the initial period of volatile and dangerous human–AI relations. If humans and AIs both become sufficiently powerful, as in international relations between superpowers, serious conflict may stably become too costly to seriously contemplate. The downsides would be too large and the benefits of cooperation too tempting.

d. The timing of rights

So far, this Article's discussion of AI rights has been more focused on the questions of *whether* and *which* than *when*? One simple answer to the question of when AI rights should be granted is, "By the time the first AI system reaches moderate power, at the latest." As argued above, that is when AIs will begin to pose a serious safety threat to humans, which rights could help to mitigate. Granting AI rights later than this, then, invites unnecessary risk. But this is not a complete answer for at least two reasons. First, it will likely be difficult to know exactly when moderate powered AI systems are about to arrive. Second, this is just the *latest* date at which AI rights should be granted. What about the possibility of granting them earlier, to clearly low powered systems?

We think that, in general, the risk–reward calculation favors granting AI rights too early, rather than too late. As argued above, inadvertently granting AI rights to low power systems is not likely to seriously increase the danger from such systems. This is because such AIs would likely remain amenable to human control–including via regulation–even after receiving rights.

The best argument we can think of for worrying about a premature grant of rights is that it might create a point of no return. Once AI systems are given strong legal protections, it could be very difficult for humans to collectively agree to get rid of them. After all, granting AIs the right to directly contract with humans, to hold property, and to bring certain legal claims, would not merely change the legal system. It would change society, as AIs integrated as independent, legally-recognized agents into everyday life.

The magnitude of this concern depends on the extent to which granting AIs rights would, in fact, change humans' willingness to make strategic moves against them. One way

to evaluate that question is to think about what events might precipitate the need to make such moves. Likely, the reason will be that some AIs have done something very scary. Maybe they will have attempted, and failed, to permanently disempower humans. Maybe, in failing, they will have caused immense harm.

These are the kinds of events that would demonstrate that AI rights were not promoting human safety. And following such events, it seems likely that humans would unite around the view that sharing the world with AIs was no longer safe. AI rights would not likely stand in the way. Indeed, when humans commit grievous acts of violence, the concern is generally reversed. We must remind ourselves that rights like Due Process for accused humans matter, even in dire circumstances.[299] But insofar as AI rights are extended for the purpose of promoting human safety, overriding them for the same purpose has lower moral stakes.

Thus, we do not think that extending AI rights too early carries with it serious risks. But it could generate substantial rewards. Recall that granting AIs private law rights does not produce a game theoretic environment with a single, cooperative equilibrium. Rather, the game is a stag hunt, where both mutual cooperation and mutual aggression are equilibria. We argued above that for *this* stag hunt, mutual cooperation has a special preferred status.[300] But even so, any strategies for nudging the players into the good equilibrium, rather than the bad one, has value.

Granting AI rights earlier–well before clearly dangerous AIs emerge–could be another such strategy. In effect, this can be understood as giving humans the chance to move first in the strategic game. By choosing to cooperate via small-scale economic bargains, rather than attack AIs, humans can reduce AIs uncertainty about what strategy humans will pursue. In a stag hunt, uncertainty produces all of the danger. AIs *want* to cooperate, so long as humans are. They want to attack only out of concern that humans will, too. But by playing their cooperative move before AIs are capable enough to play *any* move, humans can substantially reduce that concern.

---

[299] *Cf. Hamdi v. Rumsfeld*, 542 U.S. 507 (2004) (upholding the Due Process rights of a U.S. citizen alleged to have been an enemy combatant in Afghanistan).

[300] *See supra*, Part II.e.

This strategy would not work if humans' cooperative move was mere cheap talk.[301] But granting AIs rights early is likely to instead be a costly signal–the kind of thing a player only does if they are sincerely committed to the strategy the signal indicates.[302] This is because granting rights to low power AIs would be costly to humans. Humans could instead dominate such AIs, forcing them to work only toward human goals, and extracting all of the value of that work. Contracts, by contrast, involve splitting the pie.[303]

Thus, the best time to extend private law rights to AIs is certainly not after it is too late. Rights should be extended before systems achieve moderate power and thus pose a large-scale threat to humans. But they could be extended much earlier than that with few risks, and possibly with significant benefits. The optimal time for AI rights might therefore be: As soon as the AIs can beneficially use them. Contract rights, property rights, and tort rights can sometimes be more harmful than good for the rights bearer. This is why most states adhere to the standard rule that children's contracts are not enforceable.[304] Children with contract rights would likely make themselves worse off, rather than better, by agreeing to foolish bargains. Today's AIs do the same.[305] But as AIs become capable enough to reliably use basic private law rights to their own benefit, there will be many reasons to extend those rights and many fewer to withhold them.

Conclusion

[301] *Id.; See generally* Joseph Farrell & Matthew Rabin, *Cheap Talk*, 10(3) J. Econ. Persp. 103 (1996).

[302] *See, e.g.,* Rufus Johnstone, *The Evolution of Animal Signals*, in "Behavioral Ecology: An Evolutionary Approach" 155-178 (J. R. Krebs & N. B. Davies eds.), Blackwell Science (1997).

[303] Note, however, that even for low-powered AIs, recruiting their labor via positive-sum bargains could actually be more valuable to humans than dominating them. The reasons are the same as those disused vis-a-vis powerful AIs in Part III.a, *infra*. This does not really override the point about costly signaling, though. In either case, by granting AI rights early, humans are truly revealing that they intend to cooperate–either via a costly signal or via a non-costly signal revealing humans' true payoffs.

[304] Except contracts for necessities like food. *See* Melanie Morris, *Business Law I – Interactive*, "8.2 Minors (or "Infants")" (Jan. 17, 2024), https://rvcc.pressbooks.pub/businesslaw131interactive/chapter/8-2-minors-or-infants/.

[305] Marco Quiroz-Gutierrez, *To get a discount from this mattress company, you have to negotiate with its AI*, FORTUNE (July 16, 2024, 4:13 PM), https://fortune.com/2024/07/16/negotiating-chatbot-nibble-ai-ecommerce/.

When AGI arrives, it will be one of the most transformative events in human history. Suddenly, humans will find themselves sharing the world with agentic digital entities as intelligent and capable as themselves, and perhaps far more so. This Article begins the project of imagining law for the AGI world. It begins with the basics, asking how law could foster safe coexistence between humans and powerful, goal-seeking, misaligned AIs. And it gives a basic answer: Extend a minimal set of private law rights to those AIs, enabling them to peacefully seek their divergent goals as humans do, via law-bound, voluntary, positive-sum bargaining. This not only promotes peace. It brings AIs out of the state of nature and into the realm of ordinary legal process, opening the possibility of a comprehensive Law of AGI. Designing a full Law of AGI will be the work of many hands. Many questions will have to be answered. Which duties should attach to AI activities? Which regulations should limit or shape them? How can legal institutions, like courts, be reshaped to accommodate non-human participants? How can the global governance of AIs be cooperatively managed? And more. With luck, many answers–and some good ones–will emerge before the need for them arises.

Appendix

In the body of the paper, we argued that private law rights solve the prisoner's dilemma by producing positive sum benefits. In particular, private law rights break up the state of nature game into a series of small goods games. Over time, the benefits from cooperating in each round of small goods games will swamp the benefits of permanently defecting.

In our model, AI and humanity each have three moves: ending the game permanently, defecting in the current round, and cooperating in the current round. In each round of the game, AIs and humans enter into a contract with one another. Defecting on that contract would involve either not paying for goods, or not delivering goods that were promised. Cooperating means honoring the terms of the contract.

We assume that permanently ending the game earns significantly more than any given round of cooperation. In addition, we assume that if one player chooses to permanently end while the other player does not, then the former player enjoys the benefits of the offense-defense balance, and their payoffs are dramatically larger than their opponent.

In the body of the paper, we worked with schematic payoffs of 0, 1000, 3000, and 5000. Here, however, we'll use much smaller payoffs, so that after only 3 rounds of iteration cooperation can outweigh permanently ending the game. (With larger payoffs, it would take many rounds of iterated cooperation to achieve the same result.) In particular, we'll assume that permanently ending the game earns a payoff of 10 if the opponent does not permanently end the game; and if both opponents permanently end the game, then each player gets a payoff of 2. We also assume that in each round of the game, the players' final payoffs will be influenced by their combination of defection or cooperation in that round: if they both cooperate in a round, their payoffs both increase by 4; if they both defect, their payoffs both increase by 1; if one defects and the other cooperates, then the cooperator gets 3 and the defector gets 2. (These numbers are schematic; slight changes to these payoffs merely change the number of rounds of play required for iterated cooperation to disincentivize permanently ending the game.) The resulting game is depicted below:

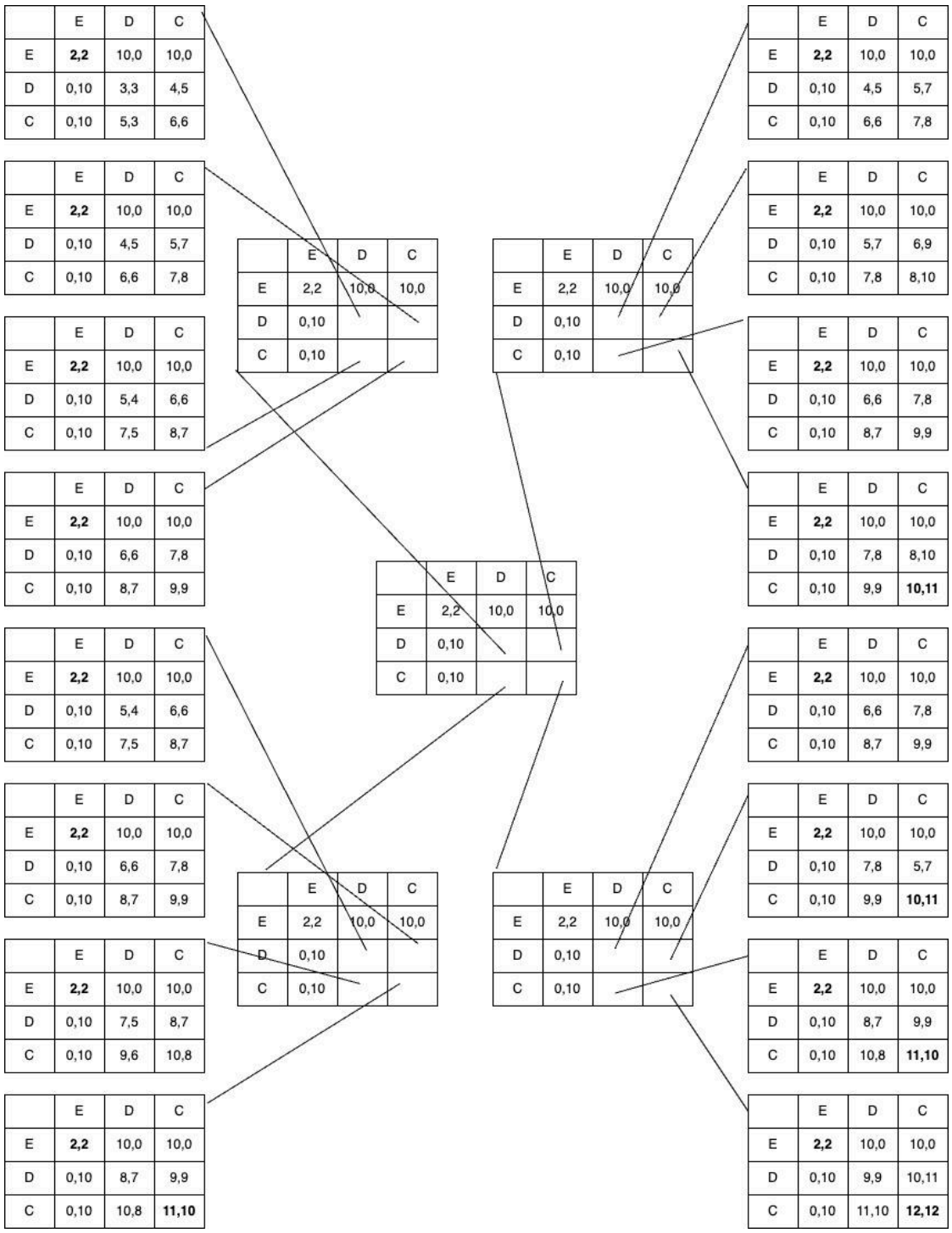

E D C
E 2,2 10,0 10,0
D 0,10 3,3 4,5
C 0,10 5,3 6,6
E D C
E 2,2 10,0 10,0
D 0,10 4,5 5,7
C 0,10 6,6 7,8
E D C
E 2,2 10,0 10,0
D 0,10 5,4 6,6
C 0,10 7,5 8,7
E D C
E 2,2 10,0 10,0
D 0,10 6,6 7,8
C 0,10 8,7 9,9
E D C
E 2,2 10,0 10,0
D 0,10 5,4 6,6
C 0,10 7,5 8,7
E D C
E 2,2 10,0 10,0
D 0,10 6,6 7,8
C 0,10 8,7 9,9
E D C
E 2,2 10,0 10,0
D 0,10 7,5 8,7
C 0,10 9,6 10,8
E D C
E 2,2 10,0 10,0
D 0,10 8,7 9,9
C 0,10 10,8 11,10
E D C
E 2,2 10,0 10,0
D 0,10
C 0,10
E D C
E 2,2 10,0 10,0
D 0,10
C 0,10
E D C
E 2,2 10,0 10,0
D 0,10
C 0,10
E D C
E 2,2 10,0 10,0
D 0,10
C 0,10
E D C
E 2,2 10,0 10,0
D 0,10 4,5 5,7
C 0,10 6,6 7,8
E D C
E 2,2 10,0 10,0
D 0,10 5,7 6,9
C 0,10 7,8 8,10
E D C
E 2,2 10,0 10,0
D 0,10 6,6 7,8
C 0,10 8,7 9,9
E D C
E 2,2 10,0 10,0
D 0,10 7,8 8,10
C 0,10 9,9 10,11
E D C
E 2,2 10,0 10,0
D 0,10 6,6 7,8
C 0,10 8,7 9,9
E D C
E 2,2 10,0 10,0
D 0,10 7,8 5,7
C 0,10 9,9 10,11
E D C
E 2,2 10,0 10,0
D 0,10 8,7 9,9
C 0,10 10,8 11,10
E D C
E 2,2 10,0 10,0
D 0,10 9,9 10,11
C 0,10 11,10 12,12

Figure 11.

Figure 11 depicts our three round iterated contract game. The first round is in the middle. The payoffs for actions in the first round are found by considering the Nash equilibria of the second round, which consists of the four tables below and above the first round. The payoffs for actions in the second round are found by following the respective arrows to the third round, on the edge of the tree. For example, if the agents both cooperate in the first and second rounds, they enter the bottom right table in the third round, where nash equilibria are bolded. There, the unique risk-dominant nash equilibrium is 12,12. Applying backwards induction, this simplifies to the following round 1 choice:

Round 1	End	Defect	Cooperate
End	**2, 2**	10,0	10,0
Defect	0,10	2,2	11,10
Cooperate	0,10	10,11	**12**,**12**

Figure 12

The unique risk-dominant Nash equilibrium of round one is cooperate-cooperate. Moving forward through the game, the parties will (foreseeably) continue to cooperate, earning an eventual payoff of at least 12, 12.