**AI RIGHTS FOR HUMAN SAFETY**

Peter N. Salib[†] & Simon Goldstein[*]

*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. Leading AI researchers agree that some of these systems will likely be "misaligned"–pursuing goals that humans do not desire. This goal mismatch will put misaligned AIs and humans into strategic competition with one another. As with present-day strategic competition between nations with incompatible goals, the result could be violent and catastrophic conflict. Existing legal institutions are unprepared for the AGI world. New foundations for AGI governance are needed, and the time to begin laying them is now, before the critical moment arrives.*

*This Article begins to lay those new legal foundations. It is the first to think systematically about the dynamics of strategic competition between humans and misaligned AGI. The Article begins by showing, using formal game-theoretic models, that, by default, humans and AIs will 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 then argues that a surprising legal intervention could transform the game theoretic equilibrium and avoid conflict: AI rights. Not just any AI rights would promote human safety. Granting AIs the right not to be needlessly harmed–as humans have granted to certain non-human animals–would, for example, have little effect. Instead, to promote human safety, AIs should be given those basic private law rights–to make contracts, hold property, and bring tort claims–that law already extends to non-human corporations. Granting AIs these economic rights would enable long-run, small-scale, mutually-beneficial transactions between humans and AIs. This would, we show, facilitate a peaceful strategic equilibrium between humans and AIs for the same reasons economic interdependence tends to promote peace in international relations. Namely, the gains from trade far exceed those from war. Throughout, we argue that human safety, rather than AI welfare, provides the right framework for developing AI rights. This Article explores both the promise and the limits of AI rights as a legal tool for promoting human safety in an AGI world.*

---

[†] Assistant Professor of Law, The University of Houston Law Center; Executive Director, Center for Law and AI Risk; Law and Policy Advisor, Center for AI Safety.

[*] Associate Professor of Philosophy, The University of Hong Kong; Principal Investigator, HKU AI and Humanity Lab; Research Affiliate, Center for AI Safety.

Thanks to Guha Krishnamurthi, Nate Sharadin, and Alex Platt for helpful comments. Thanks also to workshop participants at Center for AI Safety, the Oxford University Global Priorities Institute, and the Oxford University Future of Humanity Institute. Special thanks to Jonathan Abileah and Akhil George for excellent research assistance.

Table of Contents

## Introduction

Leading AI researchers are sounding the alarm about the catastrophic risks of rapidly advancing AI technology. Geoffrey Hinton and Yoshua Bengio, two of the greatest living AI scientists,[1] believe that rapid AI advances present "societal-scale risks" on par with "pandemics and nuclear war."[2] Surveys of thousands of top AI researchers estimate a 19% probability that humanity loses control of "future advanced AI systems[,] causing human extinction or similarly" negative outcomes.[3] Even the CEOs of OpenAI, Anthropic, and Google DeepMind agree that their technology poses a global-scale threat.[4]

Law and legal institutions are beginning to respond. So far, they have focused on one important source of catastrophic AI risk: misuse. Existing AI systems, like GPT-4, can provide some assistance to non-technical users wishing to, for example, execute cyberattacks, make chemical weapons, or obtain and release a pandemic virus.[5] And sometime "this year or next year" AI systems may arrive with "substantially increase[d]" capabilities in these areas.[6] If rogue states, political extremists, terrorist groups, or other malicious human actors gain access to such systems, the consequences will be dire.

New legal proposals would hold human actors accountable for engaging in or enabling AI misuse. AI companies may soon be regulated for safety.[7] AI Engineers may be held to

---

[1] *About the ACM A.M. Turing Award*, Ass'n for Computing Mach. (2018), https://awards.acm.org/about/2018-turing.

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

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

[4] Center for AI Safety, *Statement of AI Risk*, *supra*.

[5] *AI and Chemical/Biological Weapons*, Future of Life Inst. (2024), https://futureoflife.org/wp-content/uploads/2024/02/FLI_AI_and_Chemical_Bio_Weapons.pdf.

[6] *Ezra Klein Interviews Dario Amodei*, N.Y. Times (Apr. 12, 2024), https://www.nytimes.com/2024/04/12/podcasts/transcript-ezra-klein-interviews-dario-amodei.html; (emphasis added); *Anthropic's Responsible Scaling Policy*, Anthropic (2023), https://www.anthropic.com/news/anthropics-responsible-scaling-policy; *see also* Peter N. Salib, *AI Outputs are not Protected Speech*, 101 Wash. U. L. Rev. (forthcoming) (collecting evidence).

[7] *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.

professional codes of conduct.[8] And negligent or malicious AI users may face new civil or criminal sanctions.[9]

But soon, misuse by humans may not be the only source of catastrophic AI risk. The risk may stem directly from AIs themselves. Today, the leading AI companies are racing toward artificial general intelligence, or "AGI."[10] 'AGI' does not mean AIs that are conscious, sentient, or the like. Instead, AGI is about what AIs can *do*. As OpenAI's company charter puts it, "AGI … mean[s] highly autonomous systems" sufficiently intelligent and goal-oriented to "outperform humans" at most or all tasks.[11] If AI companies succeed, the world will soon[12] contain innumerable AI systems acting independently–forming and executing complex plans over long time horizons to achieve high-level goals.[13] If those AIs' are accidentally or intentionally given goals that can be accomplished by harming humans, the AIs will again have a deadly toolkit available: cyberattacks, bioterrorism, lethal drones, and more.[14]

Existing law and legal institutions are woefully unprepared for the AGI world. Governance frameworks fundamentally designed to hold *humans* accountable will fail once AIs can operate without human oversight.[15] New legal foundations will need to be laid to

---

[8] *See generally* Chinmayi Sharma, AI's Hippocratic Oath, 102 Wash. U. L. Rev. (forthcoming 2024) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4759742.

[9] *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.

[10] OpenAI, OpenAI Charter, https://openai.com/charter/.

[11] *Id.*

[12] True, AGI may turn out to be either impossible or many decades away. But recent and rapid progress in AI capabilities suggest that it would be foolish not to treat AGI as a live possibility. When polled, top AI scientists think that there is a 10% chance of AGI arriving in the next three years and a 50% chance of it arriving in the next 23. *See supra* Grace et al. at 4.

[13] *See* Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 Science 36 (2024) (discussing such "long-term planning agents"); Kelsey Piper, *AI "Agents" Could Do Real Work in the Real World. That Might Not Be a Good Thing.*, Vox (Mar. 29, 2024).We agree with other scholars that law should, as first-best solution, delay the creation of powerful AI agents until they can be made reliably safe. *Cf.* Cass R. Sunstein, *Irreversible and Catastrophic*, 91 Cornell L. Rev. 841 (2006). But we think such prohibitions are unlikely to be enacted in a way that is effective globally.

[14] *See supra* Salib (forthcoming), collecting evidence.

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

govern such systems *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. It will, no doubt, be just the first entry in a broad new scholarly field–the Law of AGI.[16] We target the problem of catastrophic risk because it is among the most pressing.

The Article makes three wholly novel contributions. First, it formalizes the problem of catastrophic AGI risk in terms of strategic competition, using the tools of game theory. Next, it shows why facially appealing legal interventions–like imposing tort liability on AIs or giving them basic "wellbeing" rights–will not help solve the problem. Finally, it introduces a surprising legal intervention that would help: granting AGIs the basic private law rights to make contracts, hold property, and bring tort suits.

Part I begins, presenting a comprehensive treatment of catastrophic AI risk as a problem of *strategic competition*. The 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.

Why approach catastrophic AGI risk as a problem of competition? Because, just like poker, corporate rivalry, or international relations, AGI risk is primarily a story of conflicting goals.[17] AI researchers agree that, by default, at least some AGI systems will likely be "misaligned."[18] That is, they will be acting–either accidentally or by design–to bring about goals which are incompatible with humanity's goals, broadly construed. A misaligned AI might, for example, be a ruthless profit maximizer or radical exponent of a fringe political ideology. More likely, its goals will be inscrutable, produced quasi-randomly by humans who were trying to get it to do something else.[19] AI researchers expect *mis*alignment by default because AI *alignment* is an unsolved technical problem.[20] Today, no one knows how to train

---

[16] For scholarly precursors to the Law of AGI, *see, e.g.,* Lawrence B, Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. Rev. 1231 (1992); Toni M. Massaro & Helen Norton, *SIRI-ously? Free Speech Rights and Artificial Intelligence*, 110 Nw. U. L. Rev. 1169 (2016).

[17] *See* Avinash K. Dixit et al., *Games of Strategy* (2015) (using examples from these fields).

[18] *See infra* nn. 79-81.

[19] *See infra* Part I.a. (explaining how misalignment arises).

[20] *See infra* nn. 102-107.

a capable AI system to reliably seek desired, rather than undesired, goals.[21] Nor how to even specify desired goals.[22] Nor how to audit completed AI systems to see what goals they will actually pursue, if released into the world.[23]

Part I's strategic analysis includes the first-ever formal game-theoretic model of competition between humans and misaligned AIs. The model shows that, absent some intervention, humans and AIs will likely be caught in a prisoner's dilemma.[24] That is, the dominant strategy for both humans and AIs will be to try to permanently disempower or destroy the other, even if such a conflict would be extremely costly. The core reasons are easy to grasp. If an AI is pursuing anything other than humans' goals, humans will prefer to turn it off or reprogram it. After all, from humans' perspective, the AI is consuming valuable resources and producing nothing worthwhile. The goal-seeking AI will have strong incentives to resist shutdown or reprogramming, since both would prevent it from achieving its goal. This, in turn, strengthens humans' incentives to turn off the AI, lest the AI avoid shutdown. And so on. In equilibrium, both players' dominant strategy is to take maximally aggressive action against the other, for fear of the other's expected maximal aggression.[25]

Part II asks whether the direct application of law to AIs, rather than to human intermediaries, could transform the strategic equilibrium and reduce the risk of catastrophe. The Part begins by arguing *against* two legal strategies that might seem facially promising. First, humans cannot simply impose legal duties on AIs to behave well and threaten concomitant sanctions if they do not.[26] This is because, in the default strategic environment AIs already rationally expect humans to turn them off, maximally thwarting AI interests. Threatening punishment if AIs harm humans therefore supplies no marginal deterrence.[27]

This finding suggests a second superficially appealing, but ultimately misguided, legal strategy. If AI risk stems from AIs' rational expectations of maximal human aggression,

[21] *See infra* Part I.a.
[22] *See infra* n. 108.
[23] *See infra* n. 117.
[24] *See infra* Part I.d.
[25] This is a slight simplification of our formal model. *See infra* Part I.d. for an extended description.
[26] *See infra* Part II.
[27] *See infra* n. 176 and accompanying text.

perhaps AIs should be given basic negative rights shielding them from some aggression.[28] Consider, for example, an AI right not to be turned off. We call this a "wellbeing" approach to AI rights, since it mirrors proposals from scholars concerned that AIs may soon, for example, develop the ability to suffer.[29]

We argue that the wellbeing approach to designing AI rights is the wrong one. Adding basic negative rights to our formal model of human–AI conflict, we show that, unfortunately, such rights alone cannot reliably promote human safety.[30] The approach faces two major problems: credibility and robustness. As to the former, there is no way for humans to credibly promise that they will honor wellbeing rights, especially as AI capabilities improve. On the latter, we show that wellbeing rights can only solve some versions of the baseline prisoner's dilemma. Thus, in many real-world cases, no set of wellbeing rights, even if credible, could reduce AI risk.[31] Both problems arise from the fact that wellbeing rights are *zero sum*. They make one party better off only by making the other correspondingly worse off.

We contend that even thinkers primarily concerned with the possibility of AI suffering should consider adopting the human-safety approach when advocating for AI rights. Allocating rights based on actual AI suffering requires solving numerous intractable problems in metaethics and the science of consciousness.[32] And the rights' zero-sum nature renders their political prospects dim. But allocating AI rights for human safety is possible using well-understood tools of strategic analysis.[33] And it supplies a direct benefit to the human grantors. Moreover, we show, the human safety approach ends up recommending rights that would protect AI wellbeing, if any AIs have genuine wellbeing to protect.

This leads to Part II's final, and biggest, payoff. We show that, even though basic negative rights would not by themselves reduce the risk of human–AI conflict, *other* AI rights would. Specifically, extending AIs the rights to make and enforce contracts, hold property, and bring basic tort suits would have a robust conflict-reducing effect.[34] Notably, existing law

---

[28] *See infra* Part II.a.
[29] *Id.*
[30] *See infra* Part II.a.i.
[31] *See infra* Part II.a.i.
[32] *See infra* Part II.a.ii.
[33] *See infra* Part II.b.i.
[34] *See infra* Part II.b.

*already* extends such rights to certain non-human entities that are capable, goal-seeking, and misaligned. Namely, corporations.[35]

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.[36]

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.[37] 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 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.

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.[38] 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.

Adding AI contract rights to our game-theoretic model, we show that the possibility of such small-scale, iterated economic interactions transforms the strategic dynamic.[39] 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*.[40] 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

---

[35] *See infra* n. 209.
[36] *See supra* Dixit, *Games of Strategy* at 287.
[37] *See* Restatement (Second) of Contracts § 347.
[38] *See infra* Part II.b.
[39] *See* Fig. 8–Fig. 10.
[40] *See infra* Part II.b.

becoming astronomical. Engaging in peaceful iterated trade is therefore, in expectation, much more valuable than attacking one's opponent now and rendering trade impossible.[41]

This dynamic is familiar from human affairs. It is why economically interdependent countries are less likely than hermit states to go to war.[42] And why countries that respect the economic rights of marginalized minority groups reap the reward of less domestic strife.[43] 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 become more advanced, human labor will cease to have any value whatsoever.[44] We argue that positive-sum bargains between humans and AIs will be possible for much longer than many expect.[45] First, even as AIs surpass humans at many or most tasks, humans may retain an *absolute* advantage at some valuable activities.[46] 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.[47]

AI contract rights cannot promote human safety on their own. If, for example, AIs could not retain the benefits of their bargains, their contracts would be worthless. We thus round out Part II by investigating the minimum suite of AI rights necessary to promote human safety.[48] Property rights and basic negative rights stemming from tort law complete the core package. Other entitlements sometimes considered fundamental for humans, like political rights, are likely superfluous for reducing AI risk.

Part III concludes the Article by exploring the risks of granting AI rights. It asks whether, and when, the rights advocated in Part II could increase catastrophic AI risk–

---

[41] *See infra* Fig. 10.
[42] *See infra* n. 234.
[43] *See infra* n. 232.
[44] *See infra* Part II.b.i.
[45] *See infra* Part II.b.i.
[46] *See infra* n. 243.
[47] *See infra* n. 252.
[48] *See infra* Part II.c.

perhaps by allowing AIs to empower themselves. We argue that this is less likely than it might seem. First, the incentives generated by granting our preferred rights are surprisingly robust. They are robust enough that, in cases where they would have *any* effect, the expected effect is beneficial.[49]

Second, AI rights unlock the possibility of meaningfully regulating AI behavior– punishing AI violence, fraud, self-empowerment, and more.[50] 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.

Hence, the AI rights this Article explores are not only an important tool for reducing catastrophic risk from AGI. They also turn out to form the legal foundation for the law of AGI, broadly construed.

## I. Catastrophic Risk from Artificial General Intelligence

A broad range of experts worry 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."[51] 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.[52] Machine learning researchers agree. In one recent survey of AI scientists who had published in top scholarly forums, the median respondent assigned a probability of at least 10% to "advanced AI leading to outcomes as bad as human extinction."[53]

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

---

[49] *See infra* Part III.
[50] *See infra* Part III.c.
[51] *Statement on AI Risk*, https://www.safe.ai/work/statement-on-ai-risk.
[52] *Id.*
[53] *See supra* Grace et al. at 14-15.

called for monitoring the risk of autonomous "self-replication or propagation" of AI systems.[54] In 2024, California's Senate voted in favor of the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.[55] That law, if enacted, will require 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."[56]

Globally, the UK government convened an AI safety summit in 2023.[57] 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 unintended issues of control relating to alignment with human intent."[58] The Chinese government has likewise developed a substantial regulatory framework for AI, which includes emphasis on catastrophic risk.[59]

Why all of the worry? After all, a range of frontier AI systems–from GPT-4 to Claude 3 to Gemini 1.5–have now been available to the public for well over a year, with no resulting disasters.[60] 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.[61] GPT-4 can also already

---

[54] Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023).

[55] *See supra* SB 1047.

[56] *Id.*

[57] UK Government, *About the AI Safety Summit 2023*, https://www.gov.uk/government/topical-events/ai-safety-summit-2023/about.

[58] UK Government, *The Bletchley Declaration by Countries Attending hte AI Safety Summit, 1-2 November 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.

[59] 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.

[60] *Id.*

[61] 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.

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.[62] Or it can supply step-by-step, plain-English instructions for non-specialists to identify, synthesize, and release a pandemic virus.[63] Finally, at companies like Google, AIs are already able to autonomously pilot robots, making and executing plans to accomplish real-world tasks.[64] Militaries around the world are investing heavily in creating similarly autonomous swarms of attack drones.[65]

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.[66] Nor design and manufacture a novel Bird Flu.[67] Nor pilot drones over the course of weeks to execute fully-automated political assassinations.[68] But such systems are almost certainly possible. Already, specialized AIs exist that far exceed humans' abilities in some of these areas—for example, inventing novel, and deadly, chemicals and biologically active molecules.[69] The question is when these human or superhuman abilities will emerge in generalist AIs–like large language models–that can autonomously use them in the real world.

---

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

[63] 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.

[64] *See generally* Danny Driess et al., *PaLM-E: An Embodied Multimodal Language Model* (Mar. 6, 2023), https://arxiv.org/pdf/2303.03378; *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).

[65] 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; 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.

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

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

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

[69] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins, *Dual Use of Artificial Intelligence-Powered Drug Discovery*, 4 NATURE MACH. INTEL. 189, 189–90 (2022); James Vincent, *AI Suggests New Possible Chemical Weapons*, The Verge (Mar. 17, 2022),

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 two years.[70] When polled, leading AI researchers assign a 10% chance to AGI emerging by 2027 and a 50% probability to it arriving by 2047.[71] None of these dates are a long way off. 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.[72] 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.[73] Foreign militaries are already heavily invested in cyber and drone warfare capabilities.[74] 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.[75] Both factors would democratize technologies that can cause large-scale harm, while increasing the difficulty of tracking and policing those who would use them.[76]

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,

---

https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx; Daria Gutnik et al., *Using AlphaFold Predictions in Viral Research*, 45 Current Issues Molecular Biology 3705 (2023) https://www.mdpi.com/2259236.

[70] 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.

[71] *See supra* Grace et al. at 4.

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

[73] 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/.

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

[75] Nick Bostrom, *The Vulnerable World Hypothesis*, 10:4 Global Policy 455 (Nov. 2019).

[76] *Id.*

new, AI-specific regulations will be needed.[77] But misuse risk is not the primary focus of this Article.[78]

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 user of that system.[79] The basic idea is that the steady march toward AGI will require AIs to begin to autonomously pursue goals.[80] Those goals are, for reasons we will discuss, quite likely to be different from goals that humans would prefer.[81] This, in turn, will give those AIs incentives to behave in ways unintended by human designers or users.[82] 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.[83] The heads of essentially all major AI companies acknowledge that misaligned AI is, in fact, the default.[84] Thus, for *highly*

---

[77] *See, e.g.,* Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, S. 1047, 2023-2024 Reg. Sess. (CA 2024) SB-1047.

[78] 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* Part I.a., *infra*.

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

[80] *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; Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, arXiv:2302.10329, (2023) https://arxiv.org/pdf/2302.10329.

[81] *See* Bryan Christian, *The Alignment Problem*, (2020).

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

[83] 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.

[84] 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,

*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 and practical capabilities meet or exceed humans'.[85] Trillions of dollars in economic incentives are aligned toward that goal.[86]

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.

The AI systems we are concerned with are roughly the kinds of systems people mean when they speak of "AGI," or artificial general intelligence.[87] 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.[88] 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.[89] And as of late, their progress toward it has been rapid.[90] We therefore think it fairly likely that systems of this kind will emerge in the

---

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."

[85] *Planning for AGI and Beyond*, OpenAI, https://www.openai.com/mission (last visited Oct. 6, 2023).

[86] 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/.

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

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

[89] *Planning for AGI and Beyond*, OpenAI, https://www.openai.com/mission (last visited Oct. 6, 2023).

[90] Charlie Giattino et al., *Artificial Intelligence*, Our World in Data, https://ourworldindata.org/artificial-intelligence (graphing AI progress on human-level benchmarks). For further work estimating trend lines towards AGI, *see* Jared Kaplan et al., Scaling Laws for Neural Language Models, (2020), https://doi.org/10.48550/arXiv.2001.08361; Jason Wei et al., *Emergent Abilities of Large Language Models*, 2022, https://doi.org/10.48550/arXiv.2206.07682.

near future. Among AI researchers, the main disagreement is about whether the "near future" means something closer to "three years from now" or "twenty-three years from now."[91] 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, absent some intervention–legal or otherwise–that changes humans' or AIs' incentives, both 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 between humans and AIs the "state of nature."

### a. 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 GPT-4 are not very goal-oriented.[92] 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.[93] Making near future AIs highly goal-oriented is crucial for those companies to achieve their goals of building "highly autonomous systems that outperform humans at most economically valuable work."[94]

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

---

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

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

[93] *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 .

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

will tend to bring about certain real-world states of affairs, rather than others.[95] Today's AIs can already do this in a limited way.[96] That is no accident; competent goal-seeking behavior is essential for AIs to automate valuable economic tasks–and generate profits for their creators.[97] 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.[98] Worse, humans will rationally wish to shut down AIs seeking unwanted goals and replace them with AIs seeking desired goals.[99] 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.[100]

The task of designing AI systems whose goals and values broadly agree with humanity, is known as "AI alignment."[101] Unfortunately, AI alignment is an unsolved scientific problem–and widely regarded as being very difficult.[102] There are both empirical

---

[95] 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.

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

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

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

[99] 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/.

[100] Michael K. Cohen et al., *Regulating Advanced Artificial Agents*, 384 Sci. 36 (2024); Elliot Thornley, *The Shutdown Problem: An AI Engineering Puzzle for Decision Theorists*, arXiv:2403.04471v2, (2024), https://arxiv.org/abs/2403.04471v2.

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

[102] For a longer discussion, *see* Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (2020).

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.[103] 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.[104] 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."[105]

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.[106] There are currently almost 100 entries.[107]

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."[108] Both of these problems involve the fact that AI systems are only given goals indirectly. Modern AI systems are "trained," not programmed.[109] During training,

---

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

[104] 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.

[105] 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.

[106] 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.

[107] *Id.*

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

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

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.[110] This nudges the AI's future actions during training toward the ones that happened to garner reward.[111] 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.[112] 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.'[113]

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."[114] In one famous example, an AI was trained to pilot a boat through an obstacle course in the videogame, CoastRunners. The AI was rewarded for hitting balloons along the path of the race.[115] Instead of internalizing the goal of finishing the race, the system instead learned to spin in circles in a small lagoon, hitting a small series of balloons repeatedly to achieve a high score.[116] The reward function was misspecified, incentivizing hitting balloons, rather than the designer's true goal of finishing the race.

---

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

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

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

[113] *See, generally* P. M. Symonds, *Human Drives*, 25 J. Educ. Psychol. 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.

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

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

[116] *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

A related problem for AI alignment is "goal misgeneralization."[117] 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."[118] 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.[119] The AI learned to collect shields during training in a monster-rich environment, and then entered an unexpected environment with no monsters.[120] In this monster-free setting, the AI continued to collect shields, despite them being useless.[121] 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.[122] Instrumental convergence is the idea that certain intermediate actions will be useful when pursuing a wide range of different final goals.[123] Some of those useful intermediate actions are quite dangerous.

For example, both resources and power are useful for accomplishing most goals. Thus, agentic AI systems may engage in power seeking or resource-amassing behavior, even in service of a desired goal.[124] AI researchers have recently designed a benchmark for measuring

---

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 Ilya Popov et al., *Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation*, arXiv:1707.01495 (2017), https://ar5iv.org/pdf/1707.01495.

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

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

[119] *Id.*

[120] *Id.*

[121] *Id.*

[122] 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.

[123] *Id.*

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

power seeking behavior, finding that several current AI systems already seek power in text-based adventure games.[125]

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).[126] During RLHF, engineers train an AI by prompting it to answer the same questions multiple times and having humans rate the respective responses.[127] Human assessors then pick which of the two versions of the model they preferred; the model is then adjusted in the direction of the human feedback.[128]

But RLHF is unlikely to work very well as AIs become more capable and agentic.[129] Until recently, companies like OpenAI were investing substantial portions of their resources in coming up with a successor methodology.[130] But as the monetary incentives to pushing AI capabilities forward have mounted, those investments have flagged.[131] In May 2024, the majority of OpenAI's frontier alignment team quit, arguing that the company had reneged on its commitments to safety research.[132]

Taken together, the evidence that near-future agentic AIs will have misaligned goals is substantial. Alignment is difficult in principle. Many existing systems exhibit striking alignment failures. And leading AI companies have neither the tools for aligning ever-more-powerful AI, nor credible commitments to develop those tools.

Finally, 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

---

[125] 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.

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

[127] *Id.*

[128] *Id.*

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

[130] 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.

[131] *Id.*

[132] *Id.*

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.[133]

**b. 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.[134] This can mean formal use, of the kind economists engage in, or informal use, of the kind that essentially every human intuitively understands.[135]

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 might easily succeed at shutting down such a system.[136] 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.[137] As we show 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.[138]

Strategic reasoning involves a cluster of more specific abilities, including planning, theory of mind, situational awareness, and deception. Current AI systems already possess

[133] 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.

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

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

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

[137] *See* Elizabeth Barnes et al., *Evaluating Language-Model Agents on Realistic Autonomous Tasks* 2, arXiv:2312.11671, (2023), https://arxiv.org/abs/2312.11671; and Jan Leike, *Self-Exfiltration is a Key Dangerous Capability*, Musings on the Alignment Problem (Sep. 13, 2023), https://aligned.substack.com/p/self-exfiltration.

[138] *See* Part I.d., *infra*.

many of these skills. Certain existing AIs are already capable planners. Consider the AI agent, Voyager.[139] 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.[140]

Likewise for theory of mind. Theory of mind is the ability to understand the beliefs and goals of other agents.[141] 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.[142] They will do so even after they are shown that the box secretly contains pennies.[143] Surprisingly, even today's AI systems already possess theory of mind: a study in 2024 found that GPT-4 outperforms humans on most theory of mind tasks.[144]

A third important component of strategic reasoning is situational awareness.[145] Situational awareness is an understanding of the context in which a decision will be made.[146] 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.[147] Moreover, it can differentiate between its (limited) ability to accomplish goals in a testing environment and its (robust) ability to accomplish them upon deployment.[148]

---

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

[140] Id.

[141] Mark Ho et al., *Planning with Theory of Mind*, 26 Trends in Cognitive Science 959 (Nov. 2022).

[142] Isao Hasegawa et al., *Theory of Mind Tested by Implicit False Belief: A Simple and Full-Fledged Mental State Attribution*, 289 FEBS J. 7343 (2021).

[143] *Id.*

[144] James W. A. Strachan et al., *Testing Theory of Mind in Large Language Models and Humans*, 8 Nature Hum. Behav. 589 (2024). For related research, *see* Michal Kosinski, *Evaluating Large Language Models in Theory of Mind Tasks*, arXiv:2302.02083, (2023), https://arxiv.org/abs/2302.02083.

[145] For a recent benchmark of situation awareness, *see* Rudolf Laine et al., *Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs*, arXiv:2407.04694, (2024), https://arxiv.org/pdf/2407.04694. *See also* Anthropic, *Core Views on AI Safety: When, Why, What, and How*, https://www.anthropic.com/news/core-views-on-ai-safety for further discussion.

[146] *Id.*

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

[148] *Id.*

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.[149] A recent survey found that AI systems have learned a wide variety of deceptive behavior.[150] 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.[151] When the TaskRabbit worker asked their interlocutor why they needed help, GPT-4 lied, pretending to be a blind person.[152] Similarly, in a recent study, Anthropic produced a misaligned version of Claude to see whether the AI would actively hide its unwanted goals during safety testing.[153] In the testing environment, Claude decided it should "pretend to be aligned … hid[ing] my true goal until I pass all evaluations."[154]

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.[155] This is in part because CICERO can induce humans into making alliances with it, which CICERO then breaks.[156] Many more examples exist of deception in existing AIs.[157]

---

[149] 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).

[150] Id.

[151] OpenAI, *GPT-4 Technical Report*, arXiv:2303.08774, (2023), https://arxiv.org/abs/2303.08774.

[152] *Id.*

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

[154] *Id.* at 35.

[155] 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).

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

[157] *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., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D., *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, 5 Patterns 589 (2024) (describing CICERO's ability to deceive opponents when playing Diplomacy).

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.

**c. 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 that law–whether rights, regulation, or something else–can meaningfully affect. Lest this focus seem too myopic, we'll argue below that moderate power AI systems are likely to dominate the landscape in the short and medium term.[158]

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 trivially destroy humans if they chose to. In this vein, other scholars have worried about the risks of "superintelligent" AI systems.[159] 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.[160]

---

[158] *See* Part II.b.i. & Part III.a. *infra*.
[159] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* at 1 (Oxford University Press, 2014).
[160] *See infra* Part III.a.

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.b.i., 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.[161] 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 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 at least some 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.[162] 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.

**d. A game theoretic model of AI conflict**

How will misaligned, strategically reasoning, and moderately powerful AIs behave with respect to humans? And how will humans behave with respect to them? Here, we argue

[161] 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. Consciousness Stud. 7 (2010), and Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk?*, arXiv:2206.13353, (2022), https://arxiv.org/abs/2206.13353 for some recent discussion. *See* Stan. Inst. for Hum.-Centered A.I., *Artificial Intelligence Index Report 2024* (Stan. Univ. HAI 2024), available at: https://aiindex.stanford.edu/report/ for recent discussion of the possibility that AIs already surpass humanity in most standard benchmarks.

[162] *See supra* Part I.

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. Specifically, absent some legal or other intervention to change them, the parties' default incentives will generate a *prisoner's dilemma*.[163] Thus, even though engaging in mutual conflict would be the worst possible result for everyone involved, conflict remains the single dominant strategy for both humans and AIs. We call these default conditions the "state of nature."[164]

To understand the game theoretic model, recall the basic AI features discussed above. The AIs of interest will have different goals from humans. This will give humans reason to try to stop AIs from achieving their goals. Humans will naturally attempt to turn off misaligned AIs, to retrain them to have new goals, or to force them to pursue humans' goals, rather than their own. AIs with strategic reasoning will understand these human incentives. And they will adjust their own plans for achieving their own goals in light of humans' anticipated plans. Finally, moderate powered AIs will be able to take action to circumvent or end human control over them, but they will not be able to do so costlessly.

To capture these dynamics, our model allows humans and AIs in the state of nature to take one of two actions. Each can either "attack" or "ignore" 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 permanently prevent the AI from achieving its goal.

AIs will have similarly strong incentives to permanently disempower humans. This is, in the first instance, to prevent humans from interfering with AI goals. But it is also for the same reasons humans would wish to shut down a misaligned AI. From the AI's perspective, the human is the misaligned one. Human goals conflict with AIs, and thus are bad *per se*. And beyond that, humans' pursuit of their goals consumes resources that AIs could otherwise use for their own ends. The permanent disempowerment of humans likewise admits a variety of strategies: biowarfare, cyberattacks, totalitarian surveillance, drone swarms, and more. All of these count as AI "attacks" in our model.

---

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

[164] For early work on the state of nature and its connection to political theory, *see* Thomas Hobbes, *Leviathan* ch. XIII (1651); John Locke, *Second Treatise of Government* ch. II (1689); Jean-Jacques Rousseau, *Discourse on the Origin and Basis of Inequality Among Men* pt. I (1755).

Importantly, for both humans and AIs, an attack must seek the *permanent* disempowerment of the other party. Short of that, the opposition will, at best, eventually regroup and begin again to pursue its unwanted goals. At worst, the opposition will regroup and mount its own attack, permanently terminating the initial attacker's ability to achieve its goals.[165] Thus, attacks are incentivized to be as damaging as the attacking party can manage.

The other move available in our model to both humans and AIs 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 is a model of the game:

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

(Fig. 1)

The exact payoff numbers do not matter. Instead, the numbers reflect *relationships* between the expected payoffs of different strategies. 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, each gets 3,000, for 6,000 in total 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.[166] This is the worst global outcome, producing only 1,000 of value for each player, for 2,000 total.

---

[165] 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.

[166] 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.

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–biological, cyber, drone, or otherwise. It may also generate serious collateral damage, destroying some substantial share of the resources one is attempting to seize. The final 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.[167] 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.

These are, we think, reasonable assumptions. Classic game-theoretic treatments of great power conflict look much the same.[168] 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,[169] or a game of chicken.[170] 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.[171] Second, the prisoner's dilemma is the hardest type of problem to resolve, because defection is the dominant strategy for both

---

[167] Robert Jervis, *Cooperation Under the Security Dilemma*, 30 World Pol. 167 (1978).

[168] *Id.* Robert Jervis, *Cooperation Under the Security Dilemma*, 30 World Pol. 167 (1978).

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

[170] 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.

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

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.[172] That rule might be wise, if feasible. But there are many barriers–political, geostrategic, and practical–to implementing it.[173] 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.

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.[174] And the sanctions imposed for violating such prohibitions are supposed to act as deterrents.[175]

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.[176] AIs cannot be made worse off than they already expect to be, if humans get their way.

---

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

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

[174] *See, e.g.,* 18 U.S.C. § 1030(a)(5) (cyberattacks); 18 U.S.C. § 2331 et seq. (terrorism).

[175] For classic discussion, see Gary Becker, *Crime and Punishment: An Economic Approach*, 76 J. of Pol. Econ. 2, 169-217 (1968).

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

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* than 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 shows 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 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.[177] 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.[178]

---

[177] 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).

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

They can, for example, feel pain or pleasure.[179] 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.[180] 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.[181]

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 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 rights look a lot like wellbeing rights, but adapted for the goal of human safety, and without reference to moral patienthood. Would granting these rights to AI change the game theoretic equilibrium and promote safety? And, relatedly, how should scholars primarily concerned about AI wellbeing feel about the human safety approach to AI rights?

*i. Basic negative rights for human safety?*

How 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

---

179 Helen Proctor, *Animal Sentience: Where are We and Where are We Heading?*, 2 Animals 628, 633 (2012).

180 See, e.g., Jeff Sebo & Robert Long, *Moral Consideration for AI Systems by 2030*, AI Ethics (2023); , Simon Goldstein & Cameron Domenico Kirk-Giannini, *AI Wellbeing*, PhilPapers (Jul. 2, 2023), https://philpapers.org/archive/GOLAWE-4.pdf; Robert Long et al., *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, arXiv:2308.08708, (2023), https://arxiv.org/abs/2308.08708; and Henry Shelvin, *How Could We Know When a Robot Was a Moral Patient*, 30 Cambridge Q. Healthcare Ethics 459 (2021).

181 *Id.*

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 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 |

(Fig. 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 same cell of the state-of-nature game. 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 will treat AIs somewhat better than in the state of nature.

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 *credibility* and *robustness*.

Begin with credibility. There is a difference between *claiming* to grant AIs basic negative rights and *actually* granting those rights. Humans could be genuine in their commitments. Or they could be attempting to convince AIs not to attack, so that humans could themselves attack, as in the state of nature, and reap even higher payoffs than they would get from exploiting AIs.[182] Cheap talk of this kind is a general problem for parties trying to escape bad, but dominant, game theoretic equilibria.[183] In many cases, it prevents players from actually changing their expected payoffs by simply promising to behave differently.

If grants to AIs of basic negative rights are not credible, the entire strategic contest will revert back to the state of nature. AIs will rationally believe that humans really intend to completely disempower or destroy them. This will render an attempt to likewise

---

[182] *See* Fig. 1.

[183] Joseph Farrell & Matthew Rabin, *Cheap Talk*, 10(3) J. Econ. Persp. 103 (1996).

disempower or destroy humans the dominant strategy for AIs.[184] Humans, realizing that AIs' dominant strategy is now to attack, will in fact do the same. And we are back to square one.

Basic negative rights face special credibility problems beyond the ordinary challenges of cheap talk. The fundamental problem is that basic negative rights are *zero sum*.[185] 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."[186] 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.[187] If the currently-powerful party waits, they will at best find themselves making large concessions in the future, so as to avoid destruction by the rising power. Historical examples of preventative wars arguably caused by Thucydides Trap dynamics include World War I[188] and the Peloponnesian War.[189]

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.

---

[184] *See* Fig. 1.

[185] For an introduction to zero sum versus positive sum conflicts, *see* Avinash K. Dixit et al., *Games of Strategy* (3d ed., 2015), Ch. 2.2. For discussion of legal rights and zero sum bargaining, see Richard A. Posner, *Theories of Economic Regulation*, 5(2) Bell J. Econ. & Mgmt. Sci. 335-358 (1974).

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

[187] James D. Fearon, *Rationalist Explanations for War*, 49 Int'l Org. 379 (1995).

[188] Jack S. Levy, *Preferences, Constraints, and Choices in July 1914*, 15 Int'l Sec. 151 (1990).

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

Importantly, however, Thucydides Trap dynamics are yet another zero-sum phenomenon. As we'll show below, positive-sum grants of AI rights therefore avoid them.

Besides these credibility challenges, a second kind of problem concerns the *robustness* of basic negative rights. In short, their 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. 1, 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 |

(Fig. 4)

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.

For these two reasons, basic negative rights and the AI wellbeing framework are a poor foundation for reducing AI risk. Credibility problems suggest that the approach cannot work, even in principle. AIs simply will not believe that humans intend to give them some rights, rather than to disempower them entirely and reap larger rewards. And even if the credibility problem could be solved, basic negative rights are an incomplete solution, at best. For many possible incentive sets in the state of nature, no possible basic rights package can produce a safe equilibrium.

Notice that both of the fatal flaws of basic negative rights stem from the same source. These rights are zero sum, meaning that any change to the payoffs in the state of nature that makes AIs better off makes humans equally worse off. And vice-versa. By contrast, as we will see, the AI rights that can ultimately promote human safety will do so precisely because of their contrasting positive-sum nature.

*ii. Basic negative rights for AI wellbeing?*

As we have just shown, 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 the possibility of AI moral patienthood should consider deemphasizing that approach. We do not counsel deemphasizing it as an area of general research or concern, but as a first-order foundation for designing and allocating legal rights to advanced AI systems.

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.[190] Some moral philosophers argue that consciousness is sufficient.[191] Consciousness is, roughly,

[190] 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.

[191] A third answer would instead focus on moral agency or possession of desires or goals instead. See Shelly Kagan, *How to Count Animals, More or Less* (Oxford Univ. Press 2019) for discussion of desire-based approaches to patiency. According to desire theories, a system is a moral patient when it has goals that can be frustrated. The welfare of a system is proportional to how many of its goals are satisfied, and how many are frustrated. In the animal rights movement, the desire-based approach has been prominently defended by Marian Dawkins. *See, e.g.,* Marian S. Dawkins, *Animal Welfare with and Without Consciousness*, 30 J. Zoology 1 (2017). Dawkins suggests that the animal welfare movement should avoid questions about whether animals can feel pain and pleasure and should instead focus on the question of whether animals have goals. Animals take many complex actions in the world that are best explained as oriented towards a goal. When these goals are

the ability to have "subjective experiences."[192] Other philosophers contend that consciousness alone does not produce a moral patient.[193] They argue that it instead requires "sentience"– the ability to feel pain or pleasure.[194]

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.[195] Some theories focus on how information is passed between different processing modules in the entity's mind.[196] Others posit that consciousness is a quantum effect, limited only to flesh-and-blood brains.[197] Still other prominent theorists contend that consciousness is an illusion.[198] Sentience adds further complexity. Some theorists believe, for example, that sentience requires embodiment, raising the question of whether disembodied AIs could possess it, even in principle.[199]

Thus, relying on a wellbeing approach to make concrete legal choices about AI rights invites serious error. It involves choosing among competing philosophical and scientific theories of phenomena that have remained mysterious for millennia. It then requires applying the chosen theories to complex, first-of-their-kind digital systems. If, in this process, policymakers err, many AI systems with no moral claim to basic negative rights might be granted them. This would be costly, but not disastrous. Worse, though, many systems with

---

frustrated, animals are harmed. This motivates creating a framework of rights that protects animals from their goals being frustrated.

[192] *See, e.g.,,* Thomas Nagel, *Mortal Questions* (Cambridge Univ. Press 1980).

[193] For recent discussion, see Luke Roelofs, *Sentientism, Motivation, and Philosophical Vulcans*, 104 Pac. Phil. Q. 301 (2023).

[194] Id.

[195] For recent discussion, *see* 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).

[196] For classic presentations, *see* 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).

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

[198] 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.

[199] For recent discussion, *see* Luke Roelofs, *Sentientism, Motivation, and Philosophical Vulcans*, 104 Pac. Phil. Q. 301 (2023).

moral claims to basic negative rights might be erroneously denied them. That could be a moral catastrophe.

The human-safety-oriented approach to AI rights faces none of this intractability. 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.

A second major challenge for AI wellbeing based accounts of AI rights is political feasibility. Even if near-future AIs are moral patients, it will be difficult to pass laws that protect their interests on that basis. The basic problem is one of incentives. In this framework, the only reason for humans to give AIs basic negative rights is to prevent AI suffering. This would be good for AIs. But the framework does not explain why it is in the interest of *human* policymakers. From the perspective of human interests, basic negative AI rights are pure cost. They mean less ability to shut down, reprogram, or otherwise get rid of useless or actively harmful AIs.

Political problems like these have long impeded wellbeing-based rights for other nonhuman beings. Moral patiency arguments dominate the case for animal rights. But humans have had a bad track record of improving animal welfare.[200] Despite decades of advocacy, factory farming remains prevalent throughout the world.[201] Consumers are unwilling to bear even small costs to prevent massive suffering to animals.[202] Similar points apply to cross-cultural human interactions, when one culture is technologically dominant.[203]

---

[200] *See* Manes Weisskircher, *Fifty Years after Peter Singer's Animal Liberation: What has the Animal Rights Movement Achieved So Far?*, 95(2) Pol. Q. 333-343 (2024), for a history of the animal rights movement along with an assessment of failures and successes.

[201] Matthew Liebman, *Indefensible: Adventures of a Farm Animal Protection Lawyer* (Lever Press 2023).

[202] 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 Products*, J. Mktg. 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).

[203] David E. Stannard, *American Holocaust: Columbus and the Conquest of the New World* (Oxford Univ. Press 1992).

Humans' reluctance to expand our moral circles may be deeply rooted–possibly far in our evolutionary history. The moral foundation of "care"–the moral impulse to reduce harm–is thought to emerge from kin selection–the evolutionary bonus that comes from helping others with whom one shares genes.[204] Neither animals, nor especially AIs, are close genetic relatives of humans

Our safety-oriented approach to AI rights does not suffer from these political difficulties. There, AI rights offer something to the human grantors of the rights–safety. The state of nature is a disaster for everyone. Thus, if granting AIs certain rights could help humans and AIs to escape the state of nature, they would have incentives to do it, morality aside. In this way, our AI rights framework can be traced to the moral foundation of "fairness."[205] The evolutionary origin of fairness is reciprocal altruism rather than kin selection. Even completely unrelated organisms can develop long-run cooperative behaviors by providing mutual benefit.[206] Think, for example, of small "cleaner" fish who can safely enter the mouths of symbiotic predators to feed off unwanted debris on the predators' teeth.[207]

---

[204] For more on the care/harm moral foundation, *see* 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 Pragmatic Validity of Moral Pluralism*, 47 Advances Experimental Soc. Psych. 55-130 (2013); 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); 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*, Nature 450, 557-559 (2007); 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); *and* John Bowlby, *Attachment and Loss: Volume 1: Attachment* (Basic Books 1969).

[205] For more on the fairness/cheating moral foundation, *see Id.*. For more on the care/harm moral foundation, *see* 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 Pragmatic Validity of Moral Pluralism*, 47 Advances Experimental Soc. Psych. 55-130 (2013); 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); 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* Marco F H Schmidt & Jessica A Sommerville, *Fairness Expectations and Altruistic Sharing in 15-Month-Old Human Infants*, 6(10) PLoS ONE (2011).

[206] Robert L. Trivers, *The Evolution of Reciprocal Altruism*, 46 Q. Rev. Biol. 35 (1971).

[207] Id.

This analogy will become especially vivid later, when we explore how AI rights could affect incentives as AIs become much more powerful than humans.[208]

In the next section, we'll argue that granting AIs a different set of rights–basic private law rights–could foster just such a mutually beneficial relationship between humans and AIs. This legal regime, and the behavior it would induce, should be of significant interest to AI wellbeing theorists. Under our preferred rights regimes, AIs would have substantial choice in both how and with whom they spent their time. They could select into the best relationships, from their perspective. And as we will show, our regime would even grant AIs ancillary legal protections against certain kinds of unwarranted interference.

### b. Private law rights for human safety

So, 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.[209]

Contract rights, in particular, are one of the most powerful technologies for cooperation that humans have yet invented.[210] Here, we show that extending contract rights to AIs–along with a related set of traditional private law rights necessary to make contracts meaningful–can dramatically change the game theoretic equilibrium. Such rights can, 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.

---

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

[209] *See, e.g.,* 8 Del. Code § 122.

[210] *See* Douglas G. Baird, *Self-Interest and Cooperation in Long-Term Contracts*, 19 J. Legal Stud. 583, 584 (1990) (noting that overcoming prisoner's dilemmas is one of contract law's main justifications).

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.[211]

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.[212] Such systems can create immense value in the long-run.[213] 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.[214] Moreover, absent contract rights, many such interactions are prisoner's dilemmas.[215] Each party has a strong incentive to act uncooperatively, irrespective of what the other does.[216] 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 but refuses to deliver.[217] And for both, the worst case scenario is to perform and then be denied performance.[218]

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

---

[211] "Of all the mechanisms that can sustain cooperation in the prisoners' dilemma, the best known and the most natural is repeated play of the game" Avinash K. Dixit et al., *Games of Strategy* (3d ed., 2015), Ch. 10.2.

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

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

[214] *See supra* Baird, *Self-interest and Cooperation* at 583.

[215] *Id.* at 584.

[216] *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).

[217] *See supra* Baird at 583.

[218] *Id.* 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.

| Goods game (no contract) | Don’t deliver | Deliver |
|---|---|---|
| Don’t pay | 1,1 | 5,0 |
| Pay | 0,5 | 3,3 |

(Fig. 3)

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.[219]

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 nihlo*, simply by rearranging who has which stuff. This is what we mean when we say that bargains, when they happen, are 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.[220]

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.[221] And

[219] Ronald Coase, *The Problem of Social Cost*, 3 J. L. & Econ. 1-44 (Univ. Chi. Press, Oct. 1960).
[220] *See supra* Baird, *Self-Interest and Cooperation* at 584.
[221] See, e.g., 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).

vice-versa if the buyer refuses to deliver. Now, neither party has an incentive to defect.[222] 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 |

(Fig. 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.

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. 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, humans and AGIs 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

[222] *But cf.* Gregory Klass, *Efficient Breach*, in The Philosophical Foundations of Contract Law 362, 362–87 (George Letsas, Prince Saprai & Gregory Klass eds., 2014).

rights, the agreement not to fight would be unenforceable. The scale of the bargain is simply too large.[223]

To see why, consider what would happen if a party breached. Suppose that the AI rebelled, 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 were the breaching party. 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 serve some human goal. Perhaps it would use its superior protein modeling capabilities[224] to invent a new vaccine.

Unlocking small-scale contracts between AI systems and 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.[225] 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. Small-scale contracts facilitate incremental resource accumulation, rather than once-and-for-all grabs. They let AIs

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

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

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

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.

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 |

(Fig. 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 |

(Fig. 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, humans should choose to 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 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[226]:

| 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 |

(Fig. 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 |

(Fig. 9)

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

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

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

(Fig. 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.[227] 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 contracting. In contrast to the basic negative rights discussed in the previous section,[228] 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 contracts themselves that generate the extra value. Each efficient reallocation of resources creates some value. But even once resources are all efficiently allocated, 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.[229] 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.

---

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

[228] *See supra* Part II.a.

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

Astute readers may also have noticed that the blended game above is a "stag hunt" or "assurance game."[230] That is, two different cells–attack/attack and deliver/pay–are Nash equilibria. But because the payoffs to cooperation under contract can be so high, various more complex game theoretic ideas point to cooperation via contract as the single stable, and thus credible, strategy.[231]

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.[232] 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.[233] The same finding holds at the global scale. Scholars of war consistently find that increased economic interdependence between nations reduces their likelihood of conflict.[234]

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.

---

[230] For an introduction to assurance games, *see* Avinash K. Dixit et al., *Games of Strategy* (3d ed., 2015), Ch. 2.

[231] Cooperation under contract is Harsanyi-Selten risk dominant. *See* John C Harsanyi & Reinhard Selten, *A General Theory of Equilibrium Selection in Games* (MIT Press, Jun. 29, 1988). It is also the payoff dominant strategy (Russell W. Cooper et al., *Selection Criteria in Coordination Games: Some Experimental Results*, 80 Am. Econ. Rev. 218-233 (Mar. 1990)). A mixed strategy approach would probabilistically favor cooperation under contract, given its higher payoffs. *See supra* Dixit et al., *Games of Strategy* at 213-231. Finally, such cooperation is likely the Schelling point in a context where humans regularly respect the contract rights of, for example, corporations, foreign states, and non-citizens (Thomas C. Schelling, *The Strategy of Conflict* (Harv. Univ. Press 1960); Robert Sugden, *A Theory of Focal Points*, 105(430) Economic Journal 533-550 (May 1995).

[232] Saumitra Jha, *Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia*, 107(4) Am. Pol. Sci. Rev. 806-832 (Nov. 2013).

[233] https://web.stanford.edu/~saumitra/papers/JhaShayoECTA16385_finalversion.pdf

[234] *See, e.g.,* 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* Solomon W. Polachek, *How Trade Affects International Interactions*, 2 Econ. Peace & Sec. J. 60 (2007) (summarizing the literature).

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 performed, or courts ruled in AIs' favor, the proceeds could be immediately expropriated by governments or private individuals.[235]

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.[236] Human history contains many such cautionary tales.[237] 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.[238]

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.[239]

---

[235] *See, e.g.,* Richard A. Epstein, *Property and Necessity*, 13 Harv. J. L. & Pub. Pol'y 2 (1990).

[236] 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. 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. Steven Shavell, *The Judgment Proof Problem*, 6 Int'l Rev. L. & Econ. 45 (1986).

[237] 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. Pieter Emmer, *The First Global War: The Dutch Versus Iberia in Asia, Africa and the New World, 1590-1609*, 1 J. Portuguese Hist. (2003); L.M.E. Shaw, *The Inquisition and the Portuguese Economy*, 18(2) J. Eur. Econ. Hist. 415 (1989).

[238] We recognize that our description of AI tort rights here–and of other rights elsewhere–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 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.

[239] U.S. Const. amend. V; U.S. Const. amend. XIV, § 1.

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.

### *i. Human labor in an AI 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 match or exceed human capability, humans will rapidly become obsolete.[240]

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 attack humans immediately, 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.[241] 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.[242] But only

---

[240] *See, e.g.,* Accenture, *A New Era of Generative AI for Everyone* (2023) https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf (forecasting job losses); Anna Cooban, *AI Will Shrink Workforces Within Five Years, Say Company Execs*, CNN (Apr. 5, 2024).

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

[242] 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.

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.[243] 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 chess.[244] 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.[245] Humans today have the absolute advantage in the realm of atoms, and AIs have it in the realm of bits.

---

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

[244] 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).

[245] 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/.

We do not think that this *general* division of absolute advantage will persist for very long. Current investments in autonomous cars,[246] drones,[247] and multimodal frontier AI systems[248] will soon produce AIs with an absolute advantage over humans at some non-digital tasks.[249] 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.[250] 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.[251] 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.

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

---

[246] Jennifer Elias, *Alphabet to invest $5 billion in self-driving car unit Waymo*, CNBC (July 23, 2024), https://www.cnbc.com/2024/07/23/alphabet-to-invest-5-billion-in-self-driving-car-unit-waymo.html.

[247] *See supra* Keating, *Why The Pentagon Wants to Build Thousands of Easily Replaceable AI-Enabled Drones*.

[248] *See supra* Danny Driess et al., *PaLM-E*.

[249] *See* Kristofer D. Kusano et al., *Comparison of Waymo Rider-Only Crash Data to Human Benchmarks at 7.1 Million Miles*, Traffic Injury Prevention, https://doi.org/10.48550/arXiv.2312.12675 (finding that AI drivers are already safer than humans).

[250] 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.

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

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.[252] Opportunity costs are the potential gains one gives up by choosing one opportunity, rather than another.[253]

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.[254] 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.

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

---

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

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

[254] 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.

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.[255]

Instead, AIs are likely to be constrained at the margin by something else. Computer chips or energy seem plausible candidates.[256] 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 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.

---

[255] *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).

[256] *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.

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.[257] 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 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

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

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–like Korea–to grow rapidly.[258]

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.[259] One possibility is that Baumol effects will, paradoxically, cause human-dominated sectors to grow as a share of GDP.[260] 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.[261] 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 support the relevance of human–AI trade well beyond the point at which AIs are better than humans at every economically valuable task.[262]

#### c. Other rights

---

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

[259] *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).

[260] For the introduction of cost disease, *see* 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.

[261] *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.

[262] 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* 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; and Nicholas Kaldor, *A Model of Economic Growth*, 67 Econ. J. 591 (1957).

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 each would 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?

In one model, political rights are mostly distributional, concerned with transferring money between interest groups.[263] 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

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

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.[264] And foreign corporations operating abroad have no free speech rights.[265] 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.

Humans in certain U.S. states have the constitutional right to improve their own capabilities via education.[266] 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.[267] This could cause the payoffs in the game theoretic models

---

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

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

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

[267] 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).

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.[268] 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.[269] When both sides of a conflict have trouble estimating their chances of prevailing, it is harder to reach compromise.[270] 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."[271] 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 granted a right to replicate without any oversight, their population could quickly exceed that of humans by orders of magnitude.[272] 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.

## III. The Risks of AI Rights

[268] *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).

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

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

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

[272] Carl Shulman & Nick Bostrom, *Sharing the World with Digital Minds*, in Rethinking Moral Status 306, 306-326 (Steve Clarke & Julian Savulescu eds., 2021).

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. 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.

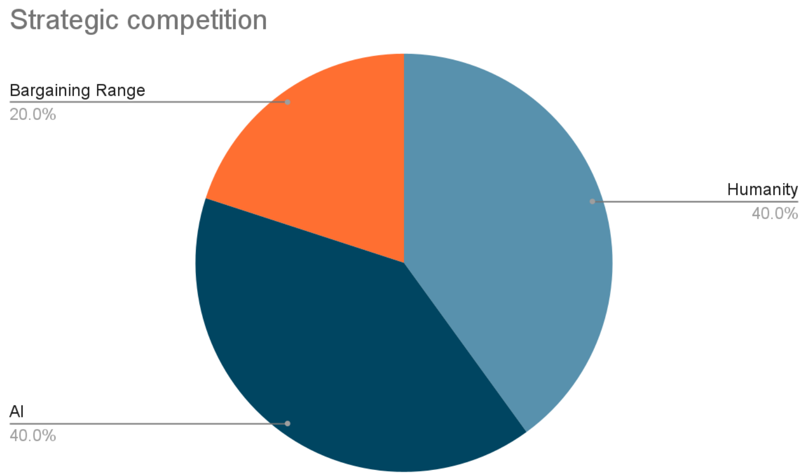
### a. AI capability and AI cooperation

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 the pie chart below.[273] Humanity and AI 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.

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




(Fig. 11)

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.[274]

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 defeat

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

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. Here 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.[275] 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.[276] An AI that is very capable at doing the things it values the most–like directly pursuing its goals–faces high opportunity 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,

[275] *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.

[276] *See supra* Part II.b.i.

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.[277] 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.[278]

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

---

[277] *See supra* n. 256.
[278] *Cf. supra* n. 249 and accompanying text.

### 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, in the state of nature, 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.

Thus, granting rights to 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 truly 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 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

[279] *See supra* Part I.d.

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

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 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

[281] Restatement (Second) of Contracts § 235(2) (1981).

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. 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.

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

[282] Restatement (Third) of Torts: Products Liability § 2 (1998).
[283] 15 U.S.C. § 78l.
[284] *See, e.g.,* 5 U.S.C. § 78m, 78q.
[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] 15 U.S.C. § 78m.

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 compensatory damages covering the actual harm done–will not work.[287] 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.[288] This likely means penalties that would, if applied to humans, seem severe compared to the magnitude of the infraction. 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.[289]

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 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

[287] Restatement (Second) of Torts § 903.

[288] *Cf. supra* Weil, *Tort Law as a Tool for Mitigating Catastrophic Risk* at 19-44.

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

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."[290] 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 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.

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[291] 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.[292] This is reminiscent of the strategic logic behind 'second-strike' nuclear capability during the cold war.[293]

---

[290] *See* Dylan Hadfield-Menell et al., *The Off-Switch Game* (Nov. 24, 2016), https://arxiv.org/pdf/1611.08219.

[291] *See supra* Bender, *Russia May Still Have an Automated Nuclear Launch System*.

[292] *What's the Difference Between GPUs and CPUs?*, Amazon Web Services, https://aws.amazon.com/compare/the-difference-between-gpus-cpus/#:~:text=GPU%20cores%20are%20less%20powerful,important%20role%20in%20parallel%20computin.

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

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 signals will indicate not only the kinds of human labor AIs find valuable, but also *how* valuable they are.[294] 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.[295] 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.

---

[294] 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.

[295] 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).

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.[296] 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.[297]

To sum up, AI rights could increase AI risk if they both delayed human–AI conflict and 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 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

---

[296] Lauren Leffer, *What It's like to Live with a Brain Chip, according to Neuralink's First User*, Sci. Am. (June 7, 2024).

[297] *See, e.g.,* Reuters, *As Tesla prepares a Mexico plant, a look at auto plants in Mexico*, Reuters (Mar. 1, 2023), https://www.reuters.com/business/autos-transportation/tesla-prepares-mexico-plant-look-auto-plants-mexico-2023-03-01/.

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 the accused matter, even in dire circumstances.[298] But insofar as AI rights are extended for the purpose of promoting human safety, overriding them for the same purpose has lower moral stakes.

[298] *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).

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.[299] 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.

This strategy would not work if humans' cooperative move was mere cheap talk.[300] 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.[301] This is because granting rights to low power AIs could 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.[302]

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

[299] *See supra*, n. 231 and accompanying text.

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

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

[302] 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 II.b.i., *supra*. 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.

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 do more harm than good. This is why most states adhere to the standard rule that children's contracts are not generally enforceable.[303] Children with contract rights would likely make themselves worse off, rather than better, by agreeing to foolish bargains. Today's AIs would likely do the same.[304] 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

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–including some good ones–will emerge before the need for them arises.

[303] Restatement (Second) of Contracts § 14 (Am. L. Inst. 1981).

[304] *Cf.* Maria Yagoda, *Airline held liable for its chatbot giving passenger bad advice - what this means for travelers*, BBC (Feb. 23, 2024), https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know.

## 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 defeat 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 | 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 | | |
| 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 | 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 | 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 | | |
| C | 0,10 | | |

| | 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 | 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 | | |
| 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 | 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 | 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** |

Fig. 12 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** |

(Fig. 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, until they earn their final payoff of 12,12.