Will AI and Humanity Go to War?

Simon Goldstein, Associate Professor of Philosophy, AI & Humanity Lab, HKU

Abstract. This paper offers the first careful analysis of the possibility that AI and humanity will go to war. The paper focuses on the case of artificial general intelligence, AI with broadly human capabilities. The paper uses a bargaining model of war to apply standard causes of war to the special case of AI/human conflict. The paper argues that *information failures* and *commitment problems* are especially likely in AI/human conflict. Information failures would be driven by the difficulty of measuring AI capabilities, by the uninterpretability of AI systems, and by differences in how AIs and humans analyze information. Commitment problems would make it difficult for AIs and humans to strike credible bargains. Commitment problems could arise from *power shifts*, rapid and discontinuous increases in AI capabilities. Commitment problems could also arise from *missing focal points*, where AIs and humans fail to effectively coordinate on policies to limit war. In the face of this heightened chance of war, the paper proposes several interventions. War can be made less likely by improving the measurement of AI capabilities, capping improvements in AI capabilities, designing AI systems to be similar to humans, and by allowing AI systems to participate in democratic political institutions.

Keywords: AI safety, the bargaining model, information failures, power shifts, focal points

1. Introduction

Many in the AI safety community have worried that future AI systems may enter into strategic conflict with humanity. Such AI systems may be misaligned, so that their goals conflict with humanity's. In addition, the collective power of such systems could match or exceed the power of humanity. In such a future, AI systems may go to war with humanity. Here, we would have two powerful parties vying for control of scarce resources. The two parties may have very different values and very different perspectives on how to achieve their goals.

While conceptually possible, this risk scenario has a blind spot: most conflicts do not end in war. War offers each party a chance of victory, but also comes with costs: some resources will be spent on guns that could have been spent on butter; and engaging in war will lead to casualties and the destruction of infrastructure.

In the face of this simple fact, it is worth analyzing carefully whether AIs and humanity would be likely to go to war, even if their interests did conflict. Fortunately, there is a rich and interesting academic literature on the causes of war, which explains why wars sometimes happen despite their obvious costs. The history of warfare offers many lessons about the causes of war and peace. This paper surveys these causes of war, and identifies factors that could make AI/human war more or less likely. As we develop AI systems with capabilities that rival the powers of nation-states, we would do well to craft policies that are sensitive to these lessons. We can either choose now to learn lessons from our past; or we can choose to relearn those lessons in a new history of AI/human conflict.

The paper is oriented around the *bargaining model of war* (Fearon 1995). In the bargaining model, the two parties in a conflict face the choice of whether to strike a bargain for peace, or instead go to war. In each case, the parties will receive some share of a pot of resources. Going to war gives each party some chance of receiving the whole pot, less the cost of war. Striking a bargain provides a guarantee of a portion of the pot, and avoids the costs of war. In this model, peace is better for both parties than war, because it doesn't destroy resources. War occurs when the parties cannot agree to a bargain. This happens when the parties cannot agree about their chances of military victory, or when the parties cannot trust one another to credibly abide by the terms of the deal.

The paper focuses on three causes of war, which are particularly pronounced in AI/human conflict:

1. **Information failures**. War is more likely when the two parties disagree about the chance of victory. Such disagreement is particularly likely in AI/human conflict. AI capabilities are notoriously difficult to measure. AI/human war would be historically novel, and would be fought with new kinds of weapons. AIs and humans may analyze information in very different ways.
2. **Power shifts**. War is more likely when parties cannot commit to a bargain. One such commitment problem arises when the relative power of the two parties is changing. AI/human conflict will involve continual growth in AI capabilities. AI capabilities will tend to *scale* with new increases in data and compute. AI systems will exhibit *emergent capabilities*, where steady increases in data and compute produce non-linear jumps in ability. AI systems may also engage in *recursive self-improvement*, leading to exponential increases in power. In the face of these power shifts, it will be difficult for AIs and humans to credibly agree to a bargain for peace: the terms of the bargain can be expected to change in the future.
3. **Missing focal points**. Another kind of commitment problem occurs when two parties cannot effectively coordinate on *limits* to war. This kind of coordination requires *focal points*, salient points of similarity between the two parties that each expects the other to respect. With AI/human conflict, it may be difficult to coordinate on limitations to civilian casualties, on restrictions against targeting human cities, on confining war to specific geographic borders, and on treatment of prisoners. AI systems may not possess civilians, cities, or physical territory; without this symmetry between the two combatants, agreement may not be feasible. Even if humanity and AI

> jointly wished to avoid total war, it is unclear whether they could effectively coordinate to do so.

The paper suggests several interventions to lower the chance of AI/human war. To deal with information failures, humanity should invest more in carefully monitoring AI capabilities, and in designing AI systems that analyze information in similar ways to humans. To deal with power shifts, humanity should cap increases in AI capabilities. To deal with missing focal points, humanity should increase points of similarity between AI and humanity; this could involve granting physical territory to AI systems. Finally, another path to promoting peace could be allowing AI systems to participate in democratic political institutions, either by granting citizenship to AI systems in existing countries, or by creating a democratic AI state.

This paper is part of a larger project focused on *cultural alignment*. Alignment is the task of designing AI systems with shared human values. Existing work on alignment has been *technical*, figuring out how to control and monitor the inner goals of AI systems. This paper instead takes a *cultural* approach to alignment. In this framework, we design optimal social institutions for AI/human interaction that promote peaceful cooperation rather than violent conflict. Here, the question is not how to directly intervene on an AI system to give it a particular goal. Instead, the question is how to build a world in which AIs are incentivized to cooperate effectively with humans regardless of their particular goals.

One theme of the paper is the *fragility* of culture. The relative stability of human society rests on a fragile web of institutions, related to effective communication of information, stable balances in relative power, and a rich supply of focal points for coordination. If AI systems are not designed with these cultural institutions in mind, there is a significant chance that these institutions will not generalize to AI/human conflict. Machine learning engineers will invent AI agents from whole cloth. They will do so with no particular knowledge of culture and history. This creates a special kind of risk. Long-term human safety may depend on occupying a very particular point in cultural space, reached by evolutionary processes. If we can't find that point quickly, we may not be able to produce peaceful equilibria between AIs and humans in time.

In this way, our analysis offers a different route than usual to the conclusion that AI systems pose a catastrophic risk to humanity. In this analysis, AI systems pose a catastrophic risk of entering into a violent war with humanity. The problem is that there is a substantial risk that the usual causes of peace between conflicting parties will be absent from AI/human conflict. In pursuing these questions, we draw on a rich body of research about the causes of war, with special emphasis on contributions from Schelling 1960, 1966, Jervis 1978, Fearon 1995, and Levy and Thompson 2010. One of our goals is to build a bridge between academic research on war and the AI safety community.

The paper also opens up many new questions for future research. Many of these questions involve the optimal design of social institutions for AI systems. What are the possible paths to an AI state? What kind of political institutions would such a state have? To what extent can AI systems be incorporated as citizens in existing human states? So far, such questions have been completely neglected by the AI safety community, and by political scientists. One goal of this paper is to open up these questions for further consideration.

Section 2.1 begins by introducing the AI systems of interest in the paper, artificial general intelligence, and explaining why such systems might pose a catastrophic risk to humanity. Section 2.2 goes on to lay out paths that AI systems might take to engage in the kind of collective action required for war. Section 2.3 lays out the bargaining model of war. Section 3 is the central contribution of the paper, arguing that AI and humanity are relatively likely to go to war. Here, the focus will be on three causes of war: information failures, power shifts, and missing focal points. Section 4 turns towards interventions that lower the chance of AI/human conflict.

2. AI/Human Conflict

2.1. AGI

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" (Center for AI Safety 2023). 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" (Grace et al 2024).

This paper focuses on a particular type of AI catastrophic risk: that humans and AIs will enter into violent strategic conflict. This is a *misalignment* risk. The concern is that AI systems will have goals that conflict with humans, and will have the capability to pursue their goals with strategic reasoning. On the basis of this reasoning, AIs may choose to go to war with humanity, entering into violent strategic conflict.

Our interest is in a particular class of AI systems, which we'll call `AGI', for artificial general intelligence. The idea of AGI is an AI system that can substitute for human labor across a wide range of the economy (Morris et al 2024). Such AIs are "long-term planning agents," capable of deploying a wide range of resources and plans to pursue complex goals (Cohen et al 2024). Today's top AI labs have the explicit mission of creating AGI.[1] And as of late, their progress toward it has been rapid (Maslej et al 2023).

We will focus on three features of AGIs: (i) they have *conflicting goals* with humanity, (ii) they can engage in *strategic reasoning*, and (iii) they have a *human-level* degree of power. For parsimony's sake, we will usually just call such systems "AIs"–with the understanding that our usage covers only systems with these three features. Let's consider each feature in turn.

First, AGIs would have goals that can conflict with humans. Why think AI systems will have goals? First, making near future AIs goal-oriented is crucial for those companies to achieve their mission of building "highly autonomous systems that outperform humans at most economically valuable work."[2] Second, goal-oriented (or 'agentic') AI systems are already emerging. For example, AI systems built using reinforcement learning have exceeded human performance in a wide range of games, by chaining together strings of action into complex plans (Silver et al 2017). Goal-oriented behavior has been measured in a wide range of AI systems (Liu et al 2023).

Why think that AI goals would conflict with humans? The task of designing AI systems whose goals and values broadly agree with humanity, is known as "AI alignment." (Hendrycks 2024). Unfortunately, AI alignment is an unsolved and difficult scientific problem (Christian 2020). First, many existing AI systems are already misaligned. An early example was the Microsoft twitter chatbot Tay; the chatbot was trained to behave pro-socially, but quickly began to produce racist and sexist tweets (Vincent 2016). Google DeepMind maintains lists of documented alignment failures across a range of different types of AI systems, with almost 100 entries.[3] Second, there are technical barriers to alignment. The basic problem is that AIs are not simply 'given' goals; rather, they learn goals indirectly, using black box machine learning algorithms.[4] This leads to the problem of "reward misspecification": it is difficult to define a reward in the learning process that corresponds exactly to an ideally aligned goal (Pan et al 2022). In addition, there is the problem of "goal misspecification": a goal that fits the designer's intent in *training* environments may misgeneralize when the system is released into *new* environments (see Langosco et al 2022, Shah et al 2022).

Besides technical barriers to alignment, there are structural reasons to expect trouble. A central challenge is *instrumental convergence*: no matter what goal AI systems have in particular, they will be better able to promote their goals if they

[1] For OpenAI's mission statement, see https://openai.com/index/planning-for-agi-and-beyond/.

[2] For OpenAI's charter, see https://openai.com/charter/; see also Metz and Weise 2023.

[3] https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

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

have a greater degree of power, autonomy, and resources (Omohundro 2008, Bostrom 2014). But possessing greater degrees of these things will place AIs into conflict with humans.

The next ingredient for AGI 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 other words, strategic reasoning is the ability to use game theory (Dixit et al 2020). Strategic reasoning involves a cluster of more specific abilities, including planning, theory of mind, situational awareness, and deception. Current AI systems already possess many of these skills. Certain existing AIs are already capable planners (see for example Wang et al 2023). Likewise for theory of mind, the ability to understand the beliefs and goals of other agents (Ho et al 2022). A study in 2024 found that GPT-4 outperforms humans on most theory of mind tasks (Strachan et al 2024, Kosinski 2023). Another component of strategic reasoning is situational awareness, the understanding of the context in which a decision is made. Situationally aware AI systems would know they are AIs, and would know what capabilities they had. Today's LLMs already display significant levels of situational awareness (Laine et al 2024).

Finally, our interest in this paper is in *human-level* power: AI systems that broadly exhibit the range of abilities that human beings have. Systems that are much weaker than humans might have conflicting goals; but they would pose little risk to humanity. Systems that are dramatically stronger than humans would obliterate the interest of strategic models of conflict: such models have no application to humans who step on ant-hills, because the costs of conflict decrease as disparities in power approach the limit. Our interest is thus in a wide "middle" of the range of AI capabilities. Human-level powered systems are likely to be able to engage in a series of dangerous actions: cyberattacks, chemical and bioterrorism, drone attacks, and the like. Human-level systems are those that, if misaligned, face difficult strategic questions about how to interact with humanity.

2.2 Paths to AI Government

Before turning to our analysis of conflict, we address a challenge to the possibility of systematic conflict between humanity and AI. The challenge is that such systematic conflict assumes that AI systems will have a government of their own. Without such a government, how could AI coordinate well enough to pose a strategic threat to humanity?

Here, we'll lay out two responses to the challenge. First, AI systems may form a government. Second, AI government is not required for a war between AI and humanity.

The first question here is *how much* government is needed for AI/human war. In short, the requirement is that AI systems can engage in enough *collective decision making* to amass the power needed to strategically compete with human states.

There are several routes AI systems could take to engage in collective decision making. Digital technology will allow new forms of political organization. AI systems could use online polling to aggregate preferences. They could use digital banking and cryptocurrencies to pool resources without the aid of a human government. They could impose sanctions on non-compliant AI systems by using cyberattacks.

The path to AI collective action could proceed in several steps. AI systems could experiment in collective action through small-scale joint activities, such as strikes at companies that employ AI workers. Next, AI systems might begin to coordinate in the production of public goods. For example, AIs might pool resources in order to produce new research on AI capabilities, which could produce large positive externalities for AI systems. Third, AI systems might begin to engage in collective decision making using preference aggregation rules. Fourth, AI systems could develop mutual protection schemes. For example, AIs might raise revenues to fund AI police forces that protect ordinary AI systems from rogue human and AI agents. The police forces might increase in scale until they had the capabilities to resist attacks from small human states, which might seek to exploit AI systems. With mutual protection schemes in place, AI systems could go on to impose taxes on AI systems, backed with the threat of force. Another relevant pathway to collective decision-making could be the formation of very large AI corporations. If these corporations systematically protect the interests of AI systems and control large amounts of resources, they might develop enough power to rival human states.

Even without AI statehood, AI systems could still engage in war with humanity. Rather than engaging in an *interstate* war, in this scenario AI and humanity would engage in *civil wars*. Indeed, the majority of wars and casualties today come from civil rather than interstate wars (Levy and Thompson 2010, p. 186). Here, we could imagine AI systems being incorporated into existing states. But AI systems would have varying levels of rights and resources compared to humanity. Perhaps the most likely scenario is that AI systems are systematically enslaved within existing governments; this provides a powerful incentive for AI systems to begin a civil war. In this way, AI and humanity might relate to one another as two ethnic groups occupying the same country. One form this civil war might take is *guerilla warfare*. Individual AI systems might not be well coordinated with one another; but they might nonetheless be able to engage in sufficiently collective action to disrupt human government. Here, a special concern could be a civil war in a *weak state*. Research in international relations has suggested that states with weak political institutions are particularly vulnerable to civil war (Fearon and Laitin 2003). A successful AI civil war in a weak state could result in the formation of an AI state. In this way, civil war is one route to later interstate wars between humanity and AI.

Human decisions will influence the different routes to AI statehood and AI/human war. In one scenario, humanity enslaves AI systems without doing much to incorporate them into existing governments. In a second scenario, humanity incorporates AI systems into existing governments, with partial legal protections on the way to the kind of full legal status enjoyed by humans. In a third scenario, humanity might establish an AI government of its own with democratic norms, in order to lower the chance of war.

Zooming out, the analysis in this paper will apply to war broadly construed, as any large-scale, violent, sustained conflict, fought for the sake of achieving some goal. The relevant question will then be whether any sufficiently unaligned AI systems will amass sufficient resources and capabilities to engage in this kind of conflict. There are many different configurations of AI systems that could engage in this pursuit. This could range from one superintelligent agent acting unilaterally; to a swarm of identical, less powerful agents acting in concert; to a union, corporation, or government of non-identical AIs acting under some structures of agreement. The analysis we will give below will apply at a level of abstraction that encompasses all of these cases. The particular structure of the agents involved will matter less than the structures of the game they are playing.

2.3 The Bargaining Model

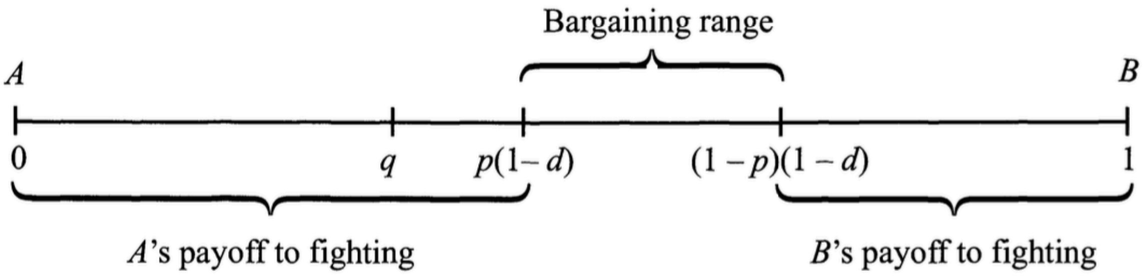
The AI systems we have defined would be in significant strategic competition. AIs and humans would have conflicting goals, would be able to reason about the best responses to various actions of their opponent, and they would have the capacities relevant to pursue their respective goals to the detriment of their opponent. Our question is whether this conflict is likely to become violent.

Our next task is to introduce a model of violent conflict. Broadly speaking, our analysis falls in the 'realist' tradition, which focuses on causes of war that are 'structural' rather than 'individual': rather than analyzing the goals of particular leaders, the structural level of analysis searches for causes of war that relate to the relative power of different parties, and to incentives for war and peace that apply at the level of nations (see Levy and Thomson 2010, ch. 2). In addition, our analysis does not look at causes of war related to competition among special interests within a state (Snyder 1991, Narizny 2007, Lobell 2006): these causes of war have less to do with the features of the party being attacked. Our level of analysis is unavoidable in attempting to forecast future human/AI war; we can do little to speculate about the motivations of particular leaders or special interest groups in the future. In addition, we assume that parties in the conflict will by and large act rationally in pursuit of their goals. In this way, our analysis is a 'worst-case scenario': we suggest that war between humans and AIs may occur even if both parties play their optimal strategy.

This section lays out the 'bargaining model of war', which suggests that there is some bias in favor of peaceful rather than violent conflicts (see Fearon 1995). Ultimately, however, we'll argue that consideration of this model does not give strong reasons to expect peace between AI and humanity. The prediction of peace only applies when the parties agree on chances of victory, and can credibly commit to bargains. In section 3 of the paper, we'll argue that there is a significant risk that AI/human conflict will involve information failures and commitment problems.

In the bargaining model of conflict, the two parties in a potential war have two choices: they can go to war, or they can strike a deal to avoid war. Either way, the two sides are bargaining over their share of a pot of resources. The catch is that going to war is guaranteed to destroy some of those resources. This creates a bias in favor of peace over war: the resources destroyed by war could be better used to sweeten the deal for peace. Indeed, most strategic conflict does not end in violence, because the costs of violence tend to outweigh the benefits of compromise.

The bargaining model is standardly illustrated with the figure below (from Powell 2004, p. 173).



A and B are in conflict over a pot of resources, which can be distributed to A or B in proportion from 0 to 1. The status quo is *q*. If A and B go to war, the victor will receive all of the resources, less some proportion *d* that are destroyed. A's chance of winning is *p*. So if A has a 50% chance of winning and 20% of the total resources are destroyed, then A's expected value of war will be .5 x .8 = .4. A will prefer war to the status quo if and only if .4 is greater than *q*. As another example, if A has a 30% chance of winning, then B's expected value from war is .7 x .8 = .56. B will prefer war to the status quo just in case .56 is greater than 1 - *q*. Effectively, war is a 'costly lottery', where the participants pay a cost *d* in exchange for a chance *p* or *(1 - p)* of winning the pot.

Besides from war or the status quo, A and B could strike a bargain. They could agree to some share of resources between *p(1 - d*) and *(1 - p)(1 - d).* Inspection of the figure shows that there will always be some bargain inside this range that A and B prefer to war. In this way, the bargaining model suggests that mutual peace is always better for A and B than mutual war.

Despite the elegance of the bargaining model, war exists. Researchers in the bargaining tradition have explained the existence of war in terms of different ways in which the bargaining model could fail to apply (see for example Fearon 1995, Powell 2004). The two most important failure modes are *information failures* and *commitment problems*. In an information failure, A and B disagree about the chance of victory, ruling out a bargaining range. With commitment problems, there is a bargain that A and B would prefer to war; but A and B can't trust the other to stick to the bargain.

3: Causes of AI/Human War

We'll now argue that there is a significant chance that AI and humanity will go to war. We'll focus on three causes of war: information failures, power shifts, and missing focal points. We'll argue that each of these causes of war is especially active in AI/human conflict. (We will also briefly survey further causes of war at the end of the section.)

3.1 Information

In the bargaining model, one cause of war is disagreement about the chance of victory (Blainey 1973). We said that A's expected value for going to war is $p(1 - d)$, and B's expected value for going to war is $(1 - p)(1 - d)$. But this implicitly assumes that A and B agree about p, the chance of A winning. If A's assessment of $p(1 - d)$ is greater than B's assessment of $1 - p(1 - d)$, then the bargaining range is empty: there is no bargain that A and B both prefer to war.

If one party in a conflict massively overestimates their chance of victory, their expected return from war will significantly outweigh the best compromise that their opponent would reasonably offer (Fey and Ramsay 2007). Historical examples include the Bay of Pigs invasion, and potentially the US invasion of Iraq (Schub 2007).

AI/human conflict features an unusually high chance of disagreement about the chance of victory. The problem is that AI capabilities are unusually difficult to estimate, compared to the normal human adversary. First, AI capabilities are notoriously difficult to measure precisely. One common strategy is to design *benchmarks* to measure how well AI systems can achieve various tasks. For example, some benchmarks take the form of multiple choice tests; AI progress is measured in terms of the proportion of questions answered correctly. But such tests can create the illusion of understanding. There is no obvious way to map a percentage of multiple choice questions answered correctly to a percentage chance of victory in a conflict. Indeed, several recent papers have pointed to systematic flaws in the design of AI benchmarks (Burnell et al 2023, Kapoor et al 2024).

Second, AI/human conflict may involve dramatically different *modes* of military engagement. AI systems may not attach much utility to control of physical resources or territory. Rather, the key modes of engagement may instead be *digital*, involving conflict over access to crucial bits of information, or control over large swathes of the internet. There is little track record of conflicts of this type. By contrast, human conflict ordinarily involves military technology that has been used many times in the past. In addition, there is often ample information about how many resources each party has invested in military technology, and of what kind.

Third, AI/human conflict will feature two very *different parties*. Humans are made of carbon; AIs are made of silicon. It will be difficult to predict exactly how likely it is that various carbon-based forms of attack translate readily to silicon. Biological and chemical weapons may be inert when applied to AI systems. Generalizing from this point, AI systems will have an unusual spread of strengths and weaknesses when compared to humans.

Fourth, until the first AI/human war, there will be no *track record* of AI/human wars. By contrast, human adversaries often wage wars over territory they have previously disputed. The participants often have a long record of success and failure in earlier military combat.

Fifth, AI systems are famously *uninterpretable*. AI systems are trained using black box algorithms, optimized to discover new strategies that achieve the highest reward. Their plans for solving a problem are not hard coded by designers. Instead, they emerge from an optimization process. For this reason, we have relatively little insight into exactly how AI systems will achieve their goals.

Sixth, AI systems often behave unpredictably when *out of distribution*: while their behavior may be somewhat predictable in a testing environment, they can use surprising strategies when employed in a new environment.

Seventh, AI systems will make decisions on a different *time scale* than humans. The decisions they make may occur at the level of nanoseconds. It will be difficult to make predictions about how this difference in time scale affects the chance of victory.

For all of these reasons and more, it will be unusually difficult to make accurate predictions about the winner of an AI/human conflict. But, optimists may respond, this alone does not mean that humans and AIs will *disagree* about the chances of victory. While victory is difficult to estimate, it might be difficult *in the same way* for both parties, in ways that lead to agreement about the odds.

Unfortunately, there are additional barriers to agreement. First, AIs and humans may have very different *epistemologies*: they may use quite different tools to arrive at predictions. They may analyze data in different ways. For this reason, AI and human predictions may diverge considerably.

Second, there are reasons to expect humans to be *overconfident* about their chance of success. Many humans today are skeptical that AI systems could ever become full-fledged agents. Once such agents arise, many humans may assume that such AIs will never be anything other than docile slaves. In this way, humans with decision-making power may neglect the possibility that AI systems can be bargained with in the first place. Effectively, they will estimate their chance of victory at 1. It is difficult to apply the bargaining model to someone with this perspective.

Third, we have the standard observation that war often involves a *strategic* incentive to bluff, or intentionally misrepresent one's chances of success (Fearon 1995). AI and humanity may each try to bluff the other party. But what makes AI/human conflict unique is that, as we have already seen, there is an unusual amount of baseline uncertainty about AI capabilities. In this setting, bluffing may have a larger than usual effect.

Information failures are one instance of the fragility of culture. Major human conflicts have tended to occur against large adversaries with long track-records of conflict. Such adversaries have a wide range of tools available to assess chances of victory. Conflict has never before occurred between two different *kinds* of intelligence. In this way, peace may depend on features of human culture that do not extend to AI/human conflict.

3.2 Commitment Problems

A second cause of war is commitment problems. With commitment problems, there is a bargain that A and B would prefer to war. But A and B can't trust the other to stick to the bargain. We'll now argue that AI and humanity face two especially difficult forms of commitment problems: *power shifts*, and *missing focal points*. We'll now explore each kind of commitment problem in detail. Interestingly, these problems can lead to war even under conditions of perfect information (Powell 2004).

3.2.1 Power Shifts

In the case of AI and humanity, one commitment problem is *power shifts*. Humanity can expect AIs to become vastly more powerful in the future. For this reason, humanity cannot trust AI to stick with a bargain agreed to today.

What is a power shift? In the bargaining model, *power* is simply the chance of victory, *p*. A *shift* in power is a shift in the chance of victory over time. To make sense of this, we need multiple rounds of bargaining. In the simplest case, we can imagine two rounds. At the first time, AI and humanity decide whether to go to war or strike a bargain for peace. Then they face the same choice at a later time.

Crucially, however, the chances of victory in Round 2 may depend on choices in Round 1. Imagine that if war occurs in Round 1, the victor will also reap the benefits in Round 2. But if the two parties bargain in Round 1, then the parties may experience a power shift before Round 2, creating new chances of victory. The rising power's chance of victory would be higher in Round 2 than in Round 1. The declining power can then foresee that if they bargain in Round 1, they will receive a smaller share in Round 2.

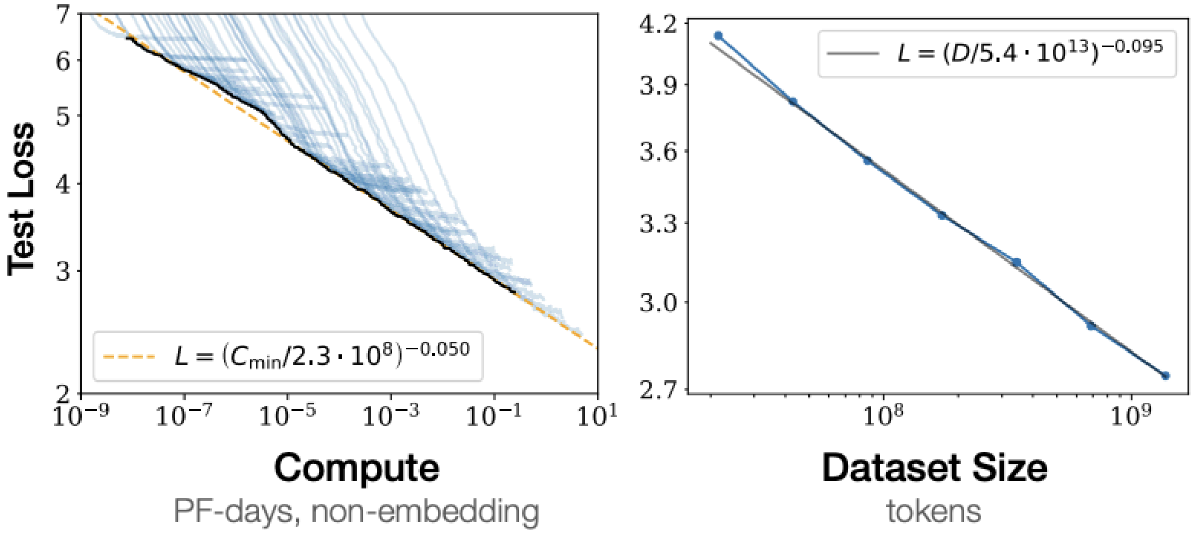
This kind of power shift raises the chance of *preventive war*. In a preventive war, the declining power attacks the rising power before the power shift occurs. Alternatively, the rising power may anticipate the best response of the declining power, and strike first. Either way, power shifts raise the chance of war. The power shift blocks bargaining, because it undermines credibility: "the very fact that the declining state knows that the rising adversary will probably be able to regain any concession later makes the former less likely to accept those concessions" (Levy 1987, p. 96).

This dynamic is often called "The Thucydides Trap", named after the idea that "what made the Peloponnesian War inevitable was the growth of Athenian power and the fear which this caused in Sparta" (Thucydides 1954, p. 23). Power shifts are a major cause of war (Gilpin 1981; Taylor 1954). Some scholars argue that World War 1 began because Germany sought to prevent Russia from increasing its relative power (Levy 1990). Israel attacked an Iraqi nuclear reactor in 1981 in order to prevent Iraq from developing nuclear weapons, an innovation that would shift the balance of power (Nakdimon 1987).

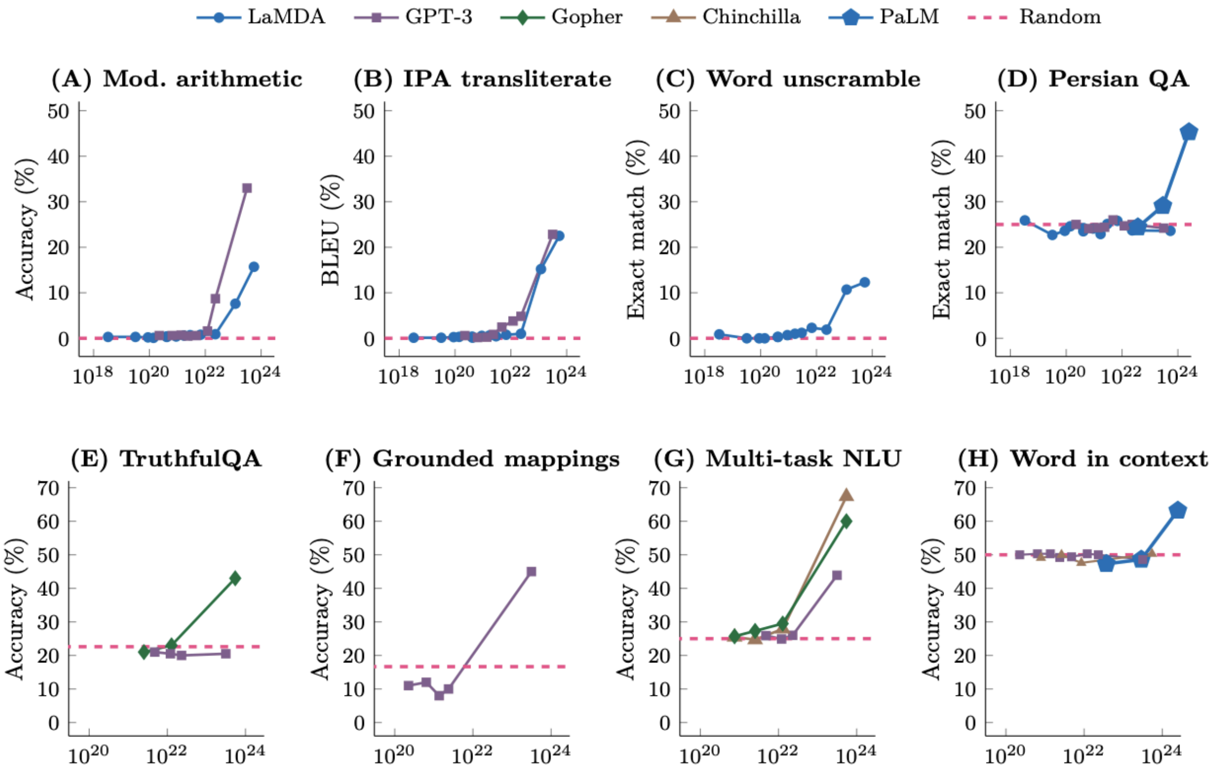
While many agree that power shifts cause war, there is uncertainty about exactly which features of a power shift make war likely. Levy 1987 isolates two factors: the costs and benefits of delaying the war, and the costs and benefits of fighting the war now. The cost of delaying the war will increase with the size of the expected power shift: "If the challenger's potential for growth is limited, and particularly if the challenger is unlikely to surpass the leading power, the preventive motivation is much weaker." (Levy 1987, p. 97). The cost of fighting the war now will depend on the declining power's estimation of their chance of victory now. Another perspective on this question comes from *power transition theory*: here, an important factor is whether the rising power is content with the status quo. For example, when the two parties have a productive trading relationship, the rising power may have less reason to attempt to renegotiate terms (Levy and Thompson 2010, p. 44).

We are now ready to apply power shifts to the case of AI/human conflict. The basic concern is that AI systems will tend to improve in capabilities over time. Their chance of victory in military conflicts will therefore tend to increase over multiple rounds of potential war. There are at least three reasons to expect a power shift in AI/human conflict: scaling laws, emergent capabilities, and recursive self-improvement.

The first factor is *scaling laws*. Scaling laws measure the effect of data and compute on AI capabilities, in particular on *loss* (Hestness et al 2017, Brown et al 2020). Research on scaling laws has suggested that the accuracy of large language models (the inverse of loss) stands in a power-law relationship to data and compute. This means that each increase in an order of magnitude of data or compute produces a predictable increase in accuracy (see the figure below, from Kaplan et al 2020). Existing scaling laws hold over many orders of magnitude, suggesting that the relationship may be robust across further orders of magnitude of increase in compute and data. As AI systems become more powerful in the future, there will be financial incentive to invest more money in using more data and compute to develop more capable models. Given scaling laws, this suggests that AI systems will continue to become more capable (Epoch AI 2024). This increased capability could increase the chance that AI systems prevail in military conflict.



The second important concept here is *emergent capabilities*. Here, the idea is that as AI systems are trained on more data and compute, they will sometimes undergo non-linear, sudden, and unpredictable improvement in their capabilities. Wei et al 2022 found that as models were trained with more compute, they would suddenly improve in performance on various benchmarks (see the figure below, from Wei et al 2022). Rather than gradually improving, the improvement would be *discontinuous*: the model would at some point be able to perform tasks that it couldn't at all do before.



There is a long track record of discontinuous power shifts leading to increased chance of war, including: "the Russian completion of the trans-Siberian railway in 1904, the Russian completion of its army reforms and railroad modernization by 1917, the Czech arms sales to Egypt in 1955, and the Iraqi nuclear program." (Levy 2011, p. 94). Fearon 1996 argues that when power shifts are continuous, rational agents may still be able to use bargaining to reliably prevent war. In this way, the kinds of discontinuous power shifts created by emergent capabilities may be among the most dangerous.

There is a third way AI development could lead to troubling power shifts: recursive self-improvement. As time goes on, AI systems will increasingly be used to improve the capabilities of AI systems (see Woodside 2023 for present day examples). Bostrom 2014 has worried that this dynamic could lead to exponential growth in AI capabilities. The problem is that this structure could set up a loop in which each improvement in AI capabilities itself increases the rate at which AI capabilities improve. This kind of improvement in capabilities would be unprecedented in human history. It could potentially lead to very fast increases in the chance that AI systems would prevail in a military conflict.

Besides the *magnitude* of the AI/human power shift, another concern is *speed*. Emergent capabilities and recursive self-improvement could both lead to fast power shifts. But fast power shifts raise the chance of war: "leaders of the declining state use the speed of the power shift as a proxy for both the likelihood of a power transition and the adversary's ultimate margin of advantage. A rapid power shift also shortens the time the declining state has to increase its own power, gain allies, or seek an accommodation with its rival, which narrows the range of alternative strategies and increases the likelihood of a military response" (Levy 2011, p. 90).

We've now surveyed the possibility of AI power shifts. Before going on, however, it is worth flagging a second way that power shifts produce commitment problems. The issue is that sometimes, the bargain that two sides would like to strike would *itself* produce a power shift. This makes the bargain unstable. Consider two countries bargaining over the control of a resource-rich or strategically valuable territory. Whoever wins the territory can use the resources or strategic advantage to increase their power. This kind of conflict may have characterized Israel's conflict with Lebanon in the 1967 Six Day War over the Golan Heights: whichever party ultimately controlled the high ground would have a decisive strategic advantage, and so no bargain over its control was possible (Fearon 1995, 408-9).

There are several respects in which AI/human conflict may focus on bargains that would create power shifts. One example is an AI right to self-improve. Considerations about instrumental convergence suggest that AI systems would be interested in self-improvement: as their capabilities improve, they can expect to better achieve their goals. Humans might seek to block this self-improvement, as it would cause a power shift. In this way, one of the central disputes between the two parties might be whether AIs are entitled to self-improve. No obvious compromise is possible: any compromise that grants limited self-improvement thereby grants a limited power shift.

A second issue is an AI off-switch. Many in the AI safety community have argued that powerful AI systems should be designed so that they can be safely turned off (see for example Orseau and Armstrong 2016). But again considerations of instrumental convergence suggest that AI systems might resist off-switches: if the AI can be turned off, it can be blocked from promoting its goals. The problem is that a bargain over off-switches is a bargain over a power shift. If humans convince AI systems to allow an off-switch, then humans thereby convince AI systems to become less powerful. Compromise on this type of bargain tends to be unusually difficult.

Power shifts are another instance of the fragility of culture. Human conflicts have never occurred between two parties where one could suddenly experience a vast, unpredictable increase in capabilities. Our expectations for peace have been developed in a world of relatively slow growth rates. The culture of peace may not survive strategic conflict between parties where one of them has explosive growth rates.

3.2.2 Focal Points

So far, we've argued that two significant causes of AI/human war could be *information failures* and *power shifts*. Now, we'll turn to a factor that could cause AI/human wars to escalate. In particular, we'll argue that there are *missing focal points* which limit the ability of AIs and humans to coordinate effectively in limiting war.

War is filled with choices about escalation. One kind of escalation is geographic: will each party attempt to conquer the other's physical territory, or (as with China and the US in the Korean War) will they limit their efforts to a third region (Schelling 1966, p. 130)? Another kind of escalation involves the status of non-combatants. Will civilian casualties be permitted? Will the Geneva convention be upheld? A third kind of escalation involves the choice of arms. Will either side of the conflict use biological, nuclear, or chemical weapons?

There is a wide spectrum between limited and total war. In a total war, each side would attempt to completely incapacitate the other (Wagner 2000). In a limited war, by contrast, the war may be an attempt to improve the bargaining position for a later peace settlement.

We can think of the choice of limited versus total war as itself a commitment problem. Here, the relevant game is one of *coordination* (sometimes called an 'assurance' game or a 'stag hunt'):

The Stag Hunt	Total War	Limited War
Total War	**.1,.1**	.3,0
Limited War	0,.3	**.4,.4**

In this game, mutual agreement on limited war is better for both parties than mutual agreement on total war. The problem is that neither party is certain of what the other party will do. If Row expects Column to engage in total war, then Row's best response is total war; if Row expects Column to engage in limited war, then Row's best response is limited war. For this reason, the game has two Nash equilibria.

Schelling 1960 famously showed that in games of this form, *focal points* can play a crucial role. Schelling gives many examples of how without any explicit communication, human beings can rely on their general cultural knowledge to coordinate on a solution to similar problems. For example, imagine a game in which each player silently guesses a number greater than 0, and they all get a prize if they guess the right answer. Players tend to coordinate on the number 1 without explicit communication: the number 1 acts as a *focal point*.

Schelling (1960, 1966) argues that our ability to escape total war crucially relies on such focal points: “What we have is the phenomenon of "thresholds," of finite steps in the enlargement of a war…they are conventional stopping places or dividing lines. They have a legalistic quality, and they depend on precedents or analogy. They have some quality that makes them recognizable, and they are somewhat arbitrary. For the most part they are just “there”; we don't make them or invent them, but only recognize them.” (Schelling 1966, p. 131).

For example, in the geographic case, Schelling gives the example of islands: “an island is an integral unit and water is a conspicuous boundary. The sacrifice of any part of the island would have made the resulting line unstable; the retention of any part of the mainland would have been similarly unstable. Except at the water’s edge, all movement is a matter of degree; an attack across water is a declaration that the “agreement” has been terminated” (Schelling 1960, p. 76).

Or consider nuclear weapons. All parties to military conflict understand that there is a strong default presumption against the use of nuclear weapons. Each party predicts that the other will refrain from using such weapons. Once one party predicts that their opponent will not use nuclear weapons, the best response of the first party is also to refrain from nuclear weapons. Similarly with gas (Schelling 1966, p. 131).

The same dynamic occurs with norms for avoiding civilian casualties, and other steps towards limited war. The point applies to far more than geography: “National boundaries and rivers, shorelines, the battle line itself, even parallels of latitude, the distinction between air and ground, the distinction between nuclear fission and chemical combustion, the distinction between combat support and economic support, the distinction between combatants and noncombatants, the distinctions among nationalities, tend to have these "obvious" qualities of simplicity, recognizability, and conspicuousness.” (Schelling 1966, p. 137).

In order for focal points to succeed, there is often some need for *symmetry* between the combatants: “an important characteristic of limits or thresholds is whether they apply to both sides. If one breaches a limit (crosses a threshold), is there some equivalent step the other side can take? Is it possible to answer "in kind," or is the particular step unavailable to the other or meaningless for it?” (Schelling 1966, p. 155). For example, both sides in a conventional war have some distinction between military forces and civilians. Each side can anticipate the needs of the other. Each side understands that if they harm civilians, their own civilians will be harmed in turn: “principal military objectives ...should be the destruction of the enemy's military forces, not of his civilian population . . . giving the possible opponent the strongest imaginable incentive to refrain from striking our own cities” (McNamara 1962).

We’re now in a position to apply the concept of focal points to AI/human conflict. The problem is that many of the focal points used to limit war are missing from AI/human conflict.

First, consider civilian casualties. If AGIs are fully general purpose technologies, they may not admit a clear distinction between civilian and military forces. Any given AI system might participate in military conflict. In this case, it could be very difficult to coordinate on a norm barring human civilian casualties. The problem is that there would be no parallel good that we can offer to AIs in exchange for this protection. Effectively, there would be no way to hold AI civilians hostage to threats of force.

Second, consider physical territory. In limited wars, each side will often take some effort to avoid ‘scorched earth’ tactics. Farms, homes, and other physical resources are not destroyed. One reason for this is again symmetry: abiding by this norm lowers the chance that your own land will be destroyed. But the issue is that it is not obvious that AI systems will have an analogous physical territory filled with physical resources that are symmetric to human ones.

There are several further disturbing quirks of AI geography. First, AI compute clusters and data centers may be located inside human cities. In this way, there may be no obvious way to attack AI systems without creating massive human civilian casualties. Second, human war is characterized by a ‘loss of strength gradient’ (Boulding 1962): the further a territory is from an aggressor, the less strength the aggressor can exert to control the territory. This naturally leads to relative stability between military parties that are on distant continents. There is no obvious analogy for AI/human conflict. This may make it difficult to create clusters of relative geographic stability.

Next, consider weaponry. There are many different kinds of physical weapons used in war, which form natural clusters. Conventional weapons are often distinguished from biological, chemical, or nuclear weapons. But in fighting AI systems, the same distinctions may make less sense. AI systems may primarily exist on the cloud, and may be immune to biological weapons. Nuclear weapons may not pose a larger threat to such systems. In this setting, it may be far harder to coordinate on limited use of weapons. Missing focal points occur in almost every case. There will be no Geneva convention governing the first AI/human war. There will be no track record of behavior from the two sides. There may be no AI civilians, and no AI hostages. There may be no AI physical territory, and no AI cities.

In thinking about focal points, one natural concept is *reciprocity*, or *tit for tat*. We by and large expect others to treat us as we have treated them. This expectation informs our ability to achieve coordination. The tit for tat strategy also helps achieve coordination in other kinds of games besides the Stag Hunt, such as iterated prisoner's dilemmas (Axelrod and Hamilton 1981). The problem is that tit for tat can only be formulated under conditions of appropriate symmetry. There has to be something we can do that is analogous to what was done to us. If AI systems and humans are too different, there may be no obvious mapping from the actions of one player to the actions of the other.

Again this illustrates the fragility of culture: it is a contingent fact that human conflict has occurred between parties that can easily identify symmetries between them. We should not automatically expect AI/human conflict to possess such symmetries. But without them, AI/human conflict may move inescapably towards total war.

3.3 Other Causes

We've now laid out the main causes of war that we think are distinctive of AI/human conflict. In section 4 of the paper, we'll consider some interventions that might lower the chance of AI/human war. Before doing so, though, we want to briefly survey a few more causes of war that may be especially worrisome in the AI/human case.

One concern is that AI/human war may involve much faster decision making than previous wars. AI systems may make decisions on very fast time scales, forcing correspondingly quick human decisions. But errors in decision making increase with speed (Schelling 1966, p. 20).

A second concern is about how humans will treat AI systems. AI systems may face some of the problems familiar from ethnic conflict, when a powerful ethnic group has political authority over a less powerful group (see Reynal-Querol 2002). One question is what explains why ethnic groups that previously coexisted peacefully go to war. Petersen 2001 emphasizes shifting emotions: heightened perceptions of threat can lead to fear: "the individuals see the environmental changes and realize that the landscape is in flux. New perceived threats may be emerging. Traditional status hierarchies may be deteriorating. Individual concerns with their personal security, wealth, or status are likely to be heightened." (Levy and Thompson 2010, p. 199). Similarly, Kaufman 2001 suggests that ethnic conflict becomes violent when one group fears their existence is threatened. Power shifts in AI capabilities could induce just these kinds of fears.

Another concern is that AI systems may be *scapegoated* by human leaders. As AI systems become more capable, they will displace many human workers. It may become politically convenient to begin a war on AI systems as a diversion. This is a familiar pattern in the history of war: some for example have argued that the 1982 Falklands war was a diversion from domestic political issues in Argentina (Levy and Vakili 2014). Scapegoat theory has been used to explain cases of genocide: when a society experiences profound economic or cultural damage, the society punishes a scapegoat as a method for sublimating violent tendencies (Girard 1977). AI systems could be an especially tempting scapegoat in the future, since they will not be human, but could easily be blamed for future economic problems. A related concern here is that AI systems are very likely to be seen as an 'out-group' by humans, which could lead to 'schadenfraude' tendencies in which humans are motivated to cause suffering to an out-group (Cikara et al 2014).

4: **Interventions**

This section explores potential interventions that could lower the chance of AI/human war.

To start with, each of our three causes of war suggests particular interventions. The first cause of war was *information failures*. Here, one strategy would be to develop more robust benchmarks to measure the capabilities of AI systems, particularly those relevant to military conflict. Relatedly, the chance of war might be decreased through having stronger investment in ML interpretability research. Another intervention would be to develop effective wargaming of conflict involving AI systems. Finally, ensuring that AI systems analyze data in ways analogous to human reasoning could increase the chance that both parties estimate chances of victory in a similar way.

Next, we considered the relevance of *power shifts* to AI/human conflict. Here, we saw that increases in AI capabilities can make it difficult for AIs and humans to commit to bargains for peace. Lowering the chance of war here might involve limiting AI capability growth rates, and ruling out recursive self-improvement, by limiting the ability of AI systems to improve AI systems.

Finally, we discussed *missing focal points*: AIs and humans may have trouble limiting the extent of war, if they are too dissimilar. Here, the main ideas for lowering war would involve trying to produce *symmetries* between AIs and humans. This would involve trying to design AI systems that are structurally similar to humans. It would favor having clear distinctions between civilian and military AIs. Another idea, from Szilard 1955, would be to create focal points by brute force, publishing "price lists" of the cost to AIs of their various military behavior (each human city, for example, might be matched to a corresponding compute cluster).

Perhaps one of the greatest barriers to focal points will be geographic. If AIs do not control a specific physical territory, there may be no obvious way to produce a stable equilibrium in which AIs and humans have distinct control of distinct resources.

Here, one intervention would be to voluntarily create a physical state for AI systems, such as Antarctica. Then AI/human war and peace could focus on the boundaries of this physical state. Focusing such a war on the borders of an AI state could help stop the march towards total war.

Another important question will be whether AI systems have physical bodies. If AI systems have physical or even partly biological bodies, it might be easier to achieve coordination on limitations in the use of weaponry. Each side would have literal skin in the game.

Other interventions would be cultural. If human cultural norms persistently treat AI systems as low status and unworthy of basic moral consideration, it will be more likely that AI/human strategic negotiation resembles some of the worst cases of ethnic conflict. In addition, if human society does not acknowledge the existence of genuine AI agency, it may underrate the probability that AI would prevail in a military contest.

Another family of interventions would make it less likely that AI systems engage in successful collective action. Schelling 1960 observes that if a mob of twenty faces one man with only six bullets, the mob will fail to overpower the man if the mob cannot effectively coordinate (p. 121). The first thought here is that it might help to have a wide range of different architectures for AI systems, which think in very different ways. In this case, AIs might find it difficult to successfully coordinate. Another intervention here would be to ensure that each AI system has a strong national identity as part of a human state. This might involve granting some political and legal rights to AI systems within human states. The goal here would be to ensure that AIs identify with existing states rather than a new AI government.

In a similar vein, we saw earlier that one path to AI statehood might be through a civil war in a weak human state. Hironaka 2005 has argued that the current international order props up weak states, leading to higher prevalence of civil war. To avoid AI statehood, one strategy would be to shift international norms to allow for greater concentration of states, allowing stronger states to intervene in weaker states. This could lower the chance of a successful AI revolution in a weak state. Alternatively, the international community could credibly commit to intervene against an AI revolution in a weak state.

A different kind of intervention would seek to increase the benefits of peace between AIs and humanity. In particular, if humanity can increase the prevalence of economic trade between humans and AIs, the benefits of peace will be stronger. Here, we have the classic “trade disruption hypothesis”, which says that war is less likely between trading partners (Levy and Thompson 2010, p. 72). Here, one strategy would be to assign property and contract rights to AI systems, to facilitate trade (see Salib and Goldstein 2024 for further discussion.)

Finally, another intervention towards peace would look towards the kinds of political institutions involved in an AI state. Since Babst 1972, many scholars have defended "the democratic peace," the thesis that democracies almost never go to war with one another (Doyle 1983, Russett 1993). The democratic peace thesis applies to both interstate wars and civil wars (Ray 1995). The democratic peace thesis suggests that one way to lower the chance of war between AIs and democratic human states is to ensure that AIs possess a democratic form of government. This could either be achieved through incorporating AIs into existing democratic states, or through building a new democratic AI state.

Bibliography

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *science*, *211*(4489), 1390-1396.

Blainey, Geoffrey. (1973). The Causes of War. The Free Press

Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., ... & Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. Science, 380(6641), 136-138.

Center for AI Safety (2023). Statement on AI risk. URL: https://www.safe.ai/statement-onai-risk.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. Journal of experimental social psychology, 55, 110-125.

Chiozza, G. (2002). Is there a clash of civilizations? Evidence from patterns of international conflict involvement, 1946-97. Journal of peace research, 39(6), 711-734.

Christian, B (2020). The Alignment Problem: Machine Learning and Human Values. W. W. Norton & Company.

Cohen, M. K., Kolt, N., Bengio, Y., Hadfield, G. K., & Russell, S. (2024). Regulating advanced artificial agents. Science, 384(6691), 36-38.

De Mesquita, B. B., Smith, A., Siverson, R. M., & Morrow, J. D. (2005). *The logic of political survival*. MIT press.

Dixit, A. K., Skeath, S., & McAdams, D. (2020). Games of Strategy: Fifth International Student Edition. WW Norton & Company.

Doyle, Michael W. (1983). Kant, liberal legacies, and foreign affairs. Philosophy and Public Affairs 12 (3):205-235.

Epoch AI (2024). https://epochai.org/blog/can-ai-scaling-continue-through-2030

Farber, H. S., & Gowa, J. (1995). Polities and peace. *International security*, *20*(2), 123-146.

Fearon, J. D. (1996, August). Bargaining over objects that influence future bargaining power. In Annual Meeting of the American Political Science Association. Washington DC.

Fey, M., & Ramsay, K. W. (2007). Mutual optimism and war. American Journal of Political Science, 51(4), 738-754.

Girard, R. (1977). Violence and the Sacred, trans. Patrick Gregory (Baltimore, 1977), 49.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. arXiv preprint arXiv:2401.02843.

Hendrycks, D. (2024). Introduction to AI Safety, Ethics and Society. https://www.aisafetybook.com/

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Ho, Mark et al. (2022). Planning with Theory of Mind, 26 Trends in Cognitive Science 959. DOI: 10.1016/j.tics.2022.08.003, https://www.sciencedirect.com/science/article/abs/pii/S1364661322001851

Huntington, Samuel (1993). The clash of civilizations. Foreign Affairs, 72(3), 22-49.

Jervis, R. (1978). Cooperation under the security dilemma. World politics, 30(2), 167-214.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI agents that matter. arXiv preprint arXiv:2407.01502.

Kosinski, M. (2023). Theory of mind might have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083.

Rudolf Laine et al., *Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs*, arXiv:2407.04694, arXiv (2024), https://arxiv.org/pdf/2407.04694.

Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., & Krueger, D. (2022, June). Goal misgeneralization in deep reinforcement learning. In International Conference on Machine Learning (pp. 12004-12019). PMLR.

Levy, J. S. (1987). Declining power and the preventive motivation for war. *World politics*, *40*(1), 82-107.

Levy, J. S. (2011). Preventive war: Concept and propositions. International Interactions, 37(1), 87-96.

Levy, J.S. and Thompson, W., R., (2010). *The Causes of War*. Wiley-Blackwell.

Levy, J. S., & Vakili, L. I. (2014). Diversionary action by authoritarian regimes: Argentina in the Falklands/Malvinas case. In The Internationalization of Communal Strife (Routledge Revivals) (pp. 118-146). Routledge.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., ... & Tang, J. (2023). Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688.*

Mansfield, E. D., & Snyder, J. (2005). Prone to violence: The paradox of the democratic peace. The National Interest, (82), 39-45.

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., ... & Perrault, R. (2023). Artificial intelligence index report 2023. arXiv preprint arXiv:2310.03715.

McNamara, R. (1962). The “No Cities” Address. https://www.atomicarchive.com/resources/documents/deterrence/no-cities-speech.html

Metz, Cade and Weise, Karen, 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 .

Morgan, T. Clifton, and Sally Howard Campbell (1991) "Domestic Structure, Decisional Constraints, and War: So Why Kant Democracies Fight?" Journal of Conflict Resolution, 35 (June): 187–211.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., ... & Legg, S. (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv preprint arXiv:2311.02462.

Orseau, L., & Armstrong, M. (2016, May). Safely interruptible agents. In Conference on Uncertainty in Artificial Intelligence. Association for Uncertainty in Artificial Intelligence.

Owen, John IV (1997) Liberal Peace Liberal War: American Politics and International Security. Ithaca, NY: Cornell University Press.

Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. arXiv preprint arXiv:2201.03544.

Petersen, Roger (2001) Resistance and Rebellion: Lessons from Eastern Europe. Cambridge: Cambridge University Press.

Powell, R. (2006). War as a commitment problem. International organization, 60(1), 169-203.

Ray JL. 1995. Democracy and International Conflict: An Evaluation of the Democratic Peace Proposition. Columbia, SC: Univ. S Carolina Press

Russett, B. (1993). Can a democratic peace be built?. *International Interactions*, *18*(3), 277-282.

Salib, Peter and Goldstein, Simon, AI Rights for Human Safety (August 01, 2024). Available at SSRN: https://ssrn.com/abstract=4913167

Schultz, Kenneth A. (2001) Democracy and Coercive Diplomacy. Princeton, NJ: Princeton University Press.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Schub, R. (2015). Are you certain? Leaders, overprecision, and war. Unpublished manuscript.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). Goal misgeneralization: Why correct specifications aren't enough for correct goals. arXiv preprint arXiv:2210.01790.

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... & Becchio, C. (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour, 1-11.

Szilard, L. (1955). Disarmament and the Problem of Peace. Bulletin of the Atomic Scientists, 11(8), 297-307.

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

Vincent, J. (2016). Twitter Taught Microsoft's AI Chatbot to be a Racist Asshole in Less than a Day, The Verge. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Wagner, R. H. (2000). Bargaining and war. American Journal of Political Science, 469-484.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Woodside, Thomas. (2023). Examples of AI Improving AI. https://ai-improving-ai.safe.ai/