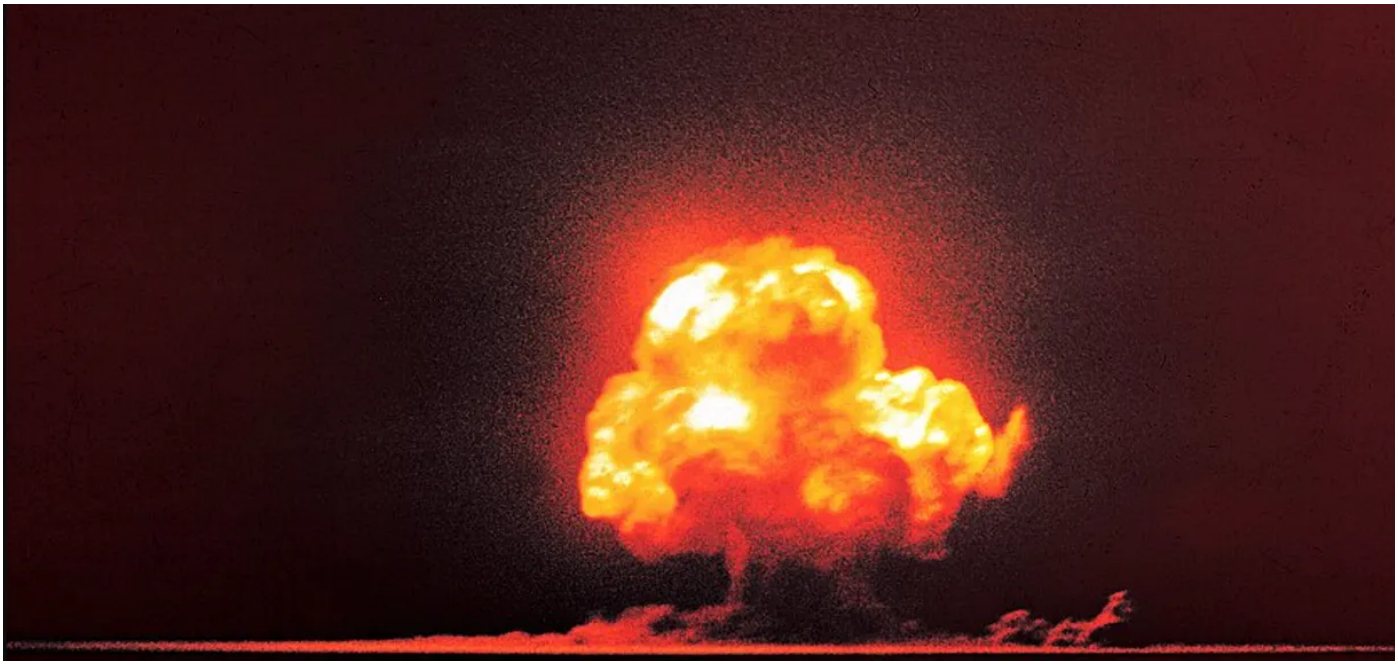
Mutual sabotage of AI probably won't work

AI deterrence isn't like nuclear deterrence

PETER WILDEFORD AND OSCAR DELANEY

APR 02, 2025

This article was written jointly by Peter Wildeford and Oscar Delaney.



Within a decade, the race for superintelligent AI could trigger geopolitical confrontations more dangerous than the Cuban Missile Crisis, due to potentially having even more powerful weapons and far less clarity about tl geopolitical rules of engagement.

everything, including geopolitics, national security, and defense.

Consider what happens once a country develops Artificial General Intellige (AGI) that can match human capabilities and then moves on to Artificial Superintelligence (ASI) that greatly exceeds human abilities. What might a country do if they learn their rivals are rapidly developing ASI that powers new ways of warfare and other forms of dominance, greatly outpacing wha was previously possible?

In a recent paper with a cool domain name (nationalsecurity.ai), Hendryc Schmidt, and Wang analyze this and propose a framework called "Mutual Assured AI Malfunction" (MAIM). Drawing parallels to nuclear deterrence and Mutual Assured Destruction (MAD), they suggest that nations will be mutually motivated to sabotage each other's AI projects that threaten to ra to ASI first, creating a form of strategic stability.

In short, Hendrycks, Schmidt, and Wang argue four key points:

- **Superintelligence leads to world domination**
    - Any nation that achieves ASI first will likely be able to leverage this initial advantage to create even more powerful AI systems, improve cyber offense and defense capabilities, hone battlefield strategy, an accelerate military R&D.
    - Together, these advances may amount to a decisive strategic advantage – namely a position of strength that means even nuclear-armed adversaries would be unable to mount significant resistance. This ability to attack any adversary at will without risk of reprisal co be tantamount to 'world domination'.
- **For China [3], risking war is better than being dominated by the US**

    - If China believes the world domination premise, they have a very strong reason to disprefer the US reaching ASI first. Rightly or wrongly, China is likely to believe that a world where the US has achieved total domination will not be conducive to a flourishing Chi or at least a flourishing CCP regime. This would make the CCP incli to race their own AI development while also threatening the US witl MAIM actins.

- **For the US, conceding some AI development would be better than wa with China**
    - If the above two premises hold, the US will face a choice. It can eith continue to race towards advanced AI (risking Chinese military strik on its AI infrastructure and China racing as well) or choose to conce on AI development (likely via seeking a negotiated agreement with China on deliberately going slower with AI development). Such a settlement would constrain AI development globally for both nation preserving a balance of power. The harms of war are so great that a settlement is preferable.

- **This equilibrium leads to peace through constrained AI development**
    - If the above three premises hold then the US and China would potentially enter into a stable equilibrium where both war and an all out race to ASI are avoided.

In making their above argument, it's unclear whether Hendrycks et al. are making a descriptive claim (this is how the world will work, these MAIM attacks will happen, and this equilibrium will hold) or a normative claim (thi how the world *should work*, these MAIM attacks *should* happen, this is the

equilibrium *we want*).

In this article, we intend only to analyze the MAIM strategy as a descript claim – is this how the world will work? We worry that this isn't the case.

The core issue is that **MAIM lacks the characteristics that made nuclear deterrence effective.** Unlike nuclear weapons, AI development has unclear red lines, limited visibility, difficult attribution, uncertain ability to retaliate and questionable effectiveness of counterattacks. Additionally, countries m not believe that AI advances are severe threats to them that warrant militar action. Let's dive deeper into the key tensions and implications of the MAIM proposal.

ASI is not necessarily world domination

A key premise of Hendrycks et. al.'s analysis is that ASI leads to world domination. However, this premise depends on the idea that there will be b differences in relative power between the state with the most advanced AI development, and everyone else. There are several reasons to be skeptical there will be such a gap:

- **There may not be a sudden jump in AI capabilities, such that no coun achieves a large AI lead.** This seems to describe the 2022-2025 status q fairly well, where the US maintains a lead, but capabilities are advancing a digestible pace, and China is only about six months behind.
- **ASI could be stolen.** Current AIs are representable in model weight files that China may be able to exfiltrate despite cybersecurity efforts (see Securing Model Weights). Both large innovations in AI-enabled cyberoffense and cyberdefense could change this picture, but the curre status quo is that if China wanted to steal a model they very likely coulc

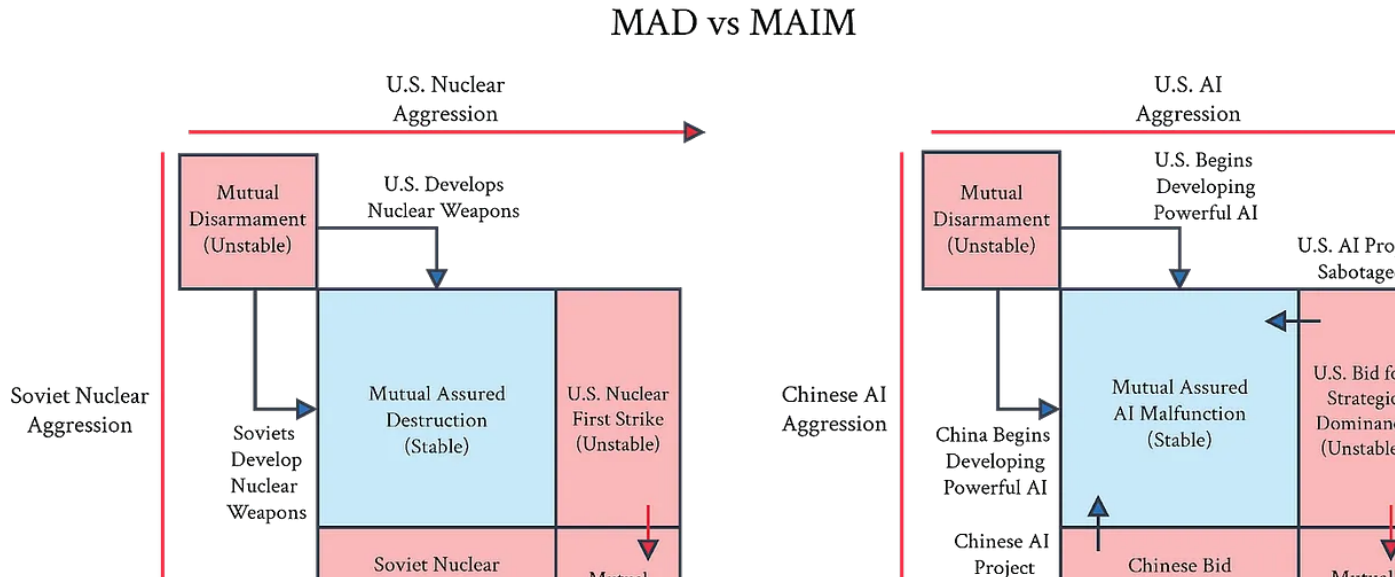
Model theft could allow China to stay competitive with the US, even if China's domestic AI industry can't keep up.

- **Other offensive technology could remain highly relevant.** Adversaries likely could harden their nuclear systems enough to avoid being taken offline by AI-enabled first strikes or cyberattacks, thus maintaining a credible second-strike deterrent despite not leading in ASI developmer

MAIM doesn't work like MAD

However, if we assume for the sake of argument that AI development will proceed to world domination if not contested, we can then begin to analyz how MAIM intends to stop it. Hendryks et al. compare MAIM to Mutual Assured Destruction (MAD) from the nuclear realm. MAD posits a stable equilibrium where if the US nuked China, China would be highly likely to be able to retaliate with devastating effect, making the US face unacceptable consequences to nuking China. Likewise, the reverse would be true if China tried to nuke the US. Thus neither country aims to nuke the other country.

Hendrycks et al. compare MAD and MAIM as follows:



In our analysis, what makes MAD work is:

1. **Clear red line of nuclear attack to respond to** - you know that it's okay for enemies to have nukes [4] but the moment they launch them towards you, you should retaliate immediately and decisively.

2. **Clear visibility of incoming nuclear attacks** - you can see the nukes coming in time to react.

3. **Clear attribution of nuclear attacks** - you know exactly who is nuking you so you can respond with a devastating counterattack to the right place.

4. **Ability to retaliate very likely survives the first strike** - you can ensur that your counterattack will launch as intended even as your nukes are attacked by enemy nukes.

5. **A retaliatory strike succeeds with very high likelihood** - it is not possi for the enemy to intercept your retaliatory strike and thus they will suf large consequences if you choose to retaliate

6. **It is very hard to lose control over nuclear weapons** - while there is definitely some risk of 'accidental' nuclear war if one country mistakenl thinks they are under attack (e.g., by misreading a signal or misunderstanding an enemy action), the nuclear weapons themselves a inert and very difficult to launch on their own without prompting or be launched by an adversary. This makes nuclear weapons safe for a state maintain, hoping that they never need to be deployed but present to de

enemy action.

7. **Strong mutual understanding of the dynamic** - the would-be first stril knows about the above six factors and knows they would be hit by a devastating counterattack with high confidence and thus knows not to attack in the first place.

These seven factors make destructive retaliation highly likely. This is what underpins MAD and why "the only winning move is not to play".

However, MAIM doesn't have these seven factors in the same way [5]:

1. **Clear red line of nuclear attack to respond to → ❌ It is not clear wha AI developments you would respond to.** In MAIM, you suffer a "salami slicing problem" as there is no equivalent of an unmistakable "nuclear strike" in AI development. If the US gradually progresses towards ASI, where does China draw a red line? What level of AI development is too much?

2. **Clear visibility of incoming nuclear attacks → ❌ AI development ma not be visible.** While current AI development operates with large data centers, it is possible that advances in distributed computing or concea data centers (e.g., underground developments) may allow for AGI projec to be developed in relative secrecy. Additionally, even if AI development progresses in the open, it may be too difficult for a rival state to proper ascertain the AI's offensive capabilities and level of threat. This is very different from a nuclear attack which is typically easy to see coming.

3. **Clear attribution of nuclear attacks → ❌ Offensive AI use could be difficult to attribute.** Advanced AI could mimic communication pattern coding styles, and operational signatures of other nations. This makes

plausible deniability much easier to maintain. Additionally, AI systems themselves might become the attackers, with complex, distributed command structures that make attribution genuinely impossible, not ju difficult. Who do you punish when an autonomous swarm with no clear operator attacks you?

4. **Ability to retaliate survives the first strike with very high likelihood** - ❓ **Ability to retaliate is unclear.** It's not well known what future advan AI attacks might look like or if they could break defenses. With AI, you may not be able to ensure that you can counterattack.

5. **A retaliatory strike succeeds with very high likelihood** → ❓ **Attempts MAIM may not succeed.** Unlike nuclear weapons which achieve devastation with very high probability, MAIM attacks are theoretical an not guaranteed to succeed. Cyberattacks alone may not completely prevent AI development, destroyed data centers can be rebuilt, and mo training could be moved to different facilities after initial facilities are destroyed. Additionally, distributed cloud computing, decentralized training, and algorithmic development increasingly don't require centralized physical locations, making AI systems less tied to particular facilities, and thus harder to disrupt through targeted strikes. Potential sufficiently motivated military AI development could involve hardened data centers underground or in other defensible positions that are harc strike via cyber or kinetic means.

6. **It is very hard to lose control over nuclear weapons** → ❓ **It is plausib to lose control over AI.** Unlike nuclear weapons that can't launch themselves, advanced AI systems may be agentic and may be able to fin ways to escape human control. Hendrycks et al. mention this risk clearl within the paper, and it is a very reasonable concern. Hendrycks et al.

worry about an AI race turning into omnicide, and we concur – The danger from AI is not just that the other side might win, but that there might be a catastrophe from misaligned AI where everyone loses.

7. **Strong mutual understanding of the dynamic → ❌ MAIM is not (yet) widely understood.** Nations do not yet operate under clear MAIM principles, and this paper may not succeed in changing that.

Countries may not follow the MAIM dynami

Another key premise is that risking war is better than being dominated by a rival country. However, this requires (a) knowledge that you are about to be dominated by a rival country and (b) a willingness to risk war. One or both o these conditions might not be met and there are some factors that push against a MAIM equilibrium forming:

- **Status quo bias:** In the current world order, doing AI training is seen as normal, legitimate thing to do, whereas launching missiles against data centers is seen as beyond the pale. For the MAIM deterrence regime to hold, this equilibrium will need to shift to doing frontier training runs o building large data centers being seen as an act of war.

- **Recognizing AI advances with high confidence is difficult:** AI might nc actually be a winner-takes-all superweapon. And even if it is, the US an China might not recognize it as such and be willing to react in a hostile and threatening way. This requires not only being bought into the theoretical possibilities of ASI, but being bought in at a very high level o confidence. There's significant uncertainty about how quickly ASI becomes a decisive military advantage, especially if your own AI development is not that far behind. Thus it seems possible but unclear

that the US could attain a decisive strategic advantage if China fails to understand what is happening and fails to react. Similarly, the US could also fail to properly react to Chinese military developments in AI.

- **Communication and verification failures:** China may threaten MAIM strikes, but the US may mistakenly think this is a bluff. Moreover, if the and China try to reach an agreement where they both slow down their development, it may be difficult to verify that the other side is complyi In general, the low-trust relationship between the US and China makes harder to achieve common knowledge through credibly honest communication.

MAIM threats may not be credible

Furthermore, in order for MAIM to work, a country needs to be able to credibly threaten to take action unless they can get something (such as ordering an AI project to halt, getting more information about an AI project or something else). But the credibility of these threats is suspect.

The biggest problem is that MAIM strikes themselves might be deterrabl and/or risk dangerous escalation. MAIM calls for aiming to destroy a rival project via cyberattack or a limited kinetic strike. However, these MAIM attacks themselves are subject to potential escalation and could thus be deterrable. For example, the 2018 US nuclear posture review under the firs Trump administration declared that the US might respond to a sufficiently damaging cyberattack with a nuclear strike. If China wants to MAIM a US A project via a cyberattack, they could be risking nuclear war in response – t may be a tall order. And it would likely be even worse if China attacked a da center with a missile strike. What kind of escalation might occur? China's leaders might think that it's preferable to take the risk of the US achieving

ASI-enabled domination than the risk of nuclear war.

Also, as mentioned above, it's not clear if MAIM attacks would even succe There is so much we don't know about how these attacks would work. But i they don't have the high likelihood of success of nuclear weapons, MAIM attacks could be deterred either through threats of retaliation or just deterrence by denial – intentional strategies that deter action by making th action infeasible or unlikely to succeed with sufficient confidence. In other words, the would-be threatener cannot threaten MAIM with credibility due lacking sufficient confidence in MAIM's success.

Should the US give in to the MAIM equilibrium?

Strategically, there's a lot of uncertainty about what the equilibrium will be. But countries can take actions to potentially bring about certain equilibriur and Hendrycks et al. potentially want the US to work to uphold and respect the MAIM equilibrium. Hendrycks et al. obfuscate between a descriptive an normative point about MAIM and the answer is genuinely unclear on both - we don't know whether the US *has* to do this or whether it should.

One path towards upholding MAIM could involve working on a treaty with China and building improved verification technology, building agreement o and common knowledge about red lines and expectations about escalation ladders, finding ways to credibly signal intentions behind AI development a prevent AI from being used for offensive purposes, agreeing to make AI development infrastructure more specialized, and even making AI intentionally vulnerable to attack.

On the other hand, the US potentially need not concede to MAIM. There's

alternative strategy that could focus on deterrence by threats of escalation and deterrence by denial. Pre-commit to massive retaliation in the event of MAIM attack, harden AI development against the possibility of MAIM attack and intentionally make it difficult for the enemy to understand the state of your own AI development and judge where and how to strike and what you might do in response. This is of course a risky strategy, but it retains the option of decisive US victory, and isn't obviously riskier than a path where MAIM is done without strong coordination about red lines and escalation ladders.

Either way, **MAIM has immense consequences – we'd be talking about locking in the current balance of power and foreclosing the opportunity t potentially use AI to reshape the world order to be more liberal and democratic.** The US may not want to give up on a potential bid for dominar

There are also political consequences and factors that cut both ways and m avoid a rational response. Acknowledging China's MAIM threat and negotiating on that basis may be portrayed (rightly or wrongly) as weak, do ish, and giving in to blackmail. This dynamic may prevent the US reaching a negotiated settlement with China even if they should. Conversely, political pressure and anti-war sentiment might push the US towards a settlement when they shouldn't.

Looking Forward

It's important to note that as AI becomes increasingly capable and geopolitically relevant, many bad outcomes are possible. We could rapidly develop AI capabilities and then lose control of them. Or we could start Wo War III. Or both. An international AI arms race may be quite perilous in mar different ways, whether there are MAIM actions or not. We're not ready to

address these complex geopolitical challenges involving advanced AI.

Thus, it is valuable that Hendrycks et al. are working through these considerations in advance because it is going to be very hard to get this analysis right when everything is exploding (potentially literally but definite metaphorically) and we have limited time to react. Best to do this analysis now, when things are relatively quiet.

Many of the actions in the Hendrycks et al. paper are worth taking, especia with regard to ensuring non-proliferation and competitiveness, both principles we uphold in IAPS's recommendations to the US government. However, while MAIM provides a useful starting point for thinking about AI deterrence, the framework requires significant refinement before it can off a viable path forward.

One central question remains whether nations can make stable agreements on advanced AI without requiring perfect visibility, verification, or trust — a challenge that makes nuclear arms control look straightforward by comparison. Another central question is whether it would be in a country's interests to actually do so. Unfortunately we don't have answers to either o these questions.

What we do know is that rather than seeking direct parallels from AI to nuclear deterrence, policymakers may need to develop novel frameworks specifically tailored to the unique characteristics of advanced AI systems.

Urgent research is still needed.

Acknowledgements: Thanks to Oliver Guest, Onni Aarne, and Liam Patell for review and contributions.

---

1 See Altman's "Three Observations" (OpenAI), Amodei's "Machines of Loving Grac (Anthropic), or this interview with Demis Hassabis (Google DeepMind).

2 See Ben Buchanan's discussion with Ezra Klein, former OpenAI policy lead Miles Brundage's Substack, or Metaculus's aggregated forecast.

3 The paper is more general and speaks of rival countries generally without specifying the US or China in particular. But to make this more clear and easier reason about, we're assuming in this article a default path where the US is leadi in AI and China is incentivized to use MAIM dynamics. But if the US were no lon the leader, a lot of the same logic could still apply.

4 Or at least if your enemies are among the official nuclear-weapon states under t terms of the Treaty on the Non-Proliferation of Nuclear Weapons. Otherwise it the building of nuclear weapons that constitutes a red line. But luckily this line i also quite clear and visible, unlike some potential forms of advanced AI development.

5 Some of these points build on "Seeking Stability in the Competition for AI Advantage" by Iskander Rehman, Karl P. Mueller, and Michael J. Mazarr. We are grateful for their work.

27 Likes · 2 Restacks

A guest post by

Oscar Delaney

AI policy researcher at IAPS (but speaking for myself). Previously dabbled in biosecurity and quant trading. Let's make the future both big and amazing!

Subscribe to Osc

Discussion about this post

Comments Restacks

Write a comment...

Dan Hendrycks ML Safety Newsletter Apr 2, 2025 *Edited*

Liked by Peter Wildeford, Oscar Delaney

> AI development doesn't have clear red lines

In the beginning of the "How to Maintain a MAIM Regime" we discuss how states will want to sta

articulating what sorts of AI projects they would find to pose imminent destabilizing risks if pursu

The AI _application_ of a fleet of thousands of AIs doing fully automated AI research (an intellige recursion) is one such red line. A recursion would likely take months and be discernible and disru through sabotage.

Red lines can evolve over time and be supplemented with communication between states throug Track IIs, Track Is, CBMs, and so on to limit misunderstandings and communicate which risk sourc states find most concerning.

LIKE (4) REPLY

1 reply

Adam Khoja AI and Its Consequences Apr 2, 2025

Liked by Peter Wildeford, Oscar Delaney

Appreciate seeing thoughtful rebuttals of the paper! This post points to several real phenomena make MAD a much more stable mutual vulnerability than MAIM, but arguably overstates the challenges for MAIM to work.

On "ASI is not necessarily world domination," I don't think this disagrees with the paper. It is prec the AI capabilities gap between nations, rather than the absolute capabilities of any one nation, t determines whether a state's security is gravely imperiled by rival AI. If the capabilities gap betwe US and China would remain small with high confidence through the entire process of developme there is less of an incentive for either party to maim. The complicating factor is that if the rate of development accelerates, perhaps during an intelligence recursion, a small time gap in AI develo may still translate to a large relative capabilities gap, which probably would be actionable. This is makes attempting a recursion a "bid for dominance" in the parlance of the paper in a way that ot types of development might not be. Also, in this framing, stealing AI weights is arguably a type o maiming attack: it reduces the relative capabilities gap between states.

Is MAIM descriptive or normative? This could have been clearer in the paper. My view is that MA descriptively points to "fortunate" features of AI development and geopolitics which make deter attractive: states are fully militarily capable of surveilling and crippling rival AI projects requiring centralization of compute, and they are incentivized to do so in the face of a large rival capabiliti or a high perceived loss of control risk from a rival AI project. All current large-scale developmen credibly vulnerable to severe disruption by states as it approaches superintelligence.

The normative claim is that states should work to preserve the mutual vulnerability of AI projects as the "facts on the ground" of AI development change: for example, AI development which requ less or less centralized compute, or heavily securitized projects that are harder to surveil or steal

weights from, or hardened AI infrastructure. States that prefer the stability of a mutual vulnerabil might agree to centralize their compute anyways; implement various transparency measures on t amount of compute they have, where it is, and what it's doing; not harden their AI infrastructure various ways; etc. Some of these conditions probably can be imposed unilaterally, i.e. a nation ca threaten to maim a project attempting to build a large underground datacenter. But others need mutual buy-in from relevant parties.

I agree that a state which wants to subvert MAIM would have several attractive ways to do so: th all-out retaliation to maiming attacks (I have been calling this "effective hardening" and think it's plausible though destabilizing option), hardening and dispersing compute as much as possible, concealing their activities and progress to reduce the confidence attacks will succeed, etc. If we r point where AI is sufficiently geopolitically salient, I can totally imagine this being seen as an act though I agree that this would be a reversal from the status quo.

I was a bit confused by some of the points on the MAIM analogy to MAD--"Offensive AI use cou difficult to attribute" and "Ability to retaliate is unclear"--as they seem to imply that MAIM mean your AI attacks me, my AI attacks you" as opposed to "Rival AI development projects are mutuall vulnerable to sabotage." MAIM is not about attacks that AIs perform. Kinetic strikes on an AI pro are attributable, and I don't see why sabotaging a rival AI project would in general prevent them sabotaging yours (separate from the valid question of whether a state can be confident that mai attacks will work).

Otherwise, I broadly agree that limited visibility of rival activities and capabilities, and uncertainty the reliability of a maiming attack make MAIM less stable than MAD. They don't seem like issues current regime of terrible security and few frontier datacenters, but they eventually will need to b actively addressed.

LIKE (2) REPLY

1 reply

6 more comments...