

# International Cooperation vs. AI Arms Race

BRIAN TOMASIK

Foundational Research Institute

brian.tomasik@foundational-research.org

Dec. 2013\*

## Abstract

There's a decent chance that governments will be the first to build artificial general intelligence (AI). International hostility, especially an [AI arms race](#), could exacerbate risk-taking, hostile motivations, and errors of judgment when creating AI. If so, then international cooperation could be an important factor to consider when evaluating the [flow-through effects](#) of charities. That said, we may not want to popularize the arms-race consideration too openly lest we accelerate the race.

## Contents

1	Will governments build AI first?	2
2	AI arms races	2
3	Ways to avoid an arms race	3
4	Are these efforts cost-effective?	4
5	Should we publicize AI arms races?	4
6	How do our prospects look?	5
7	Robot arms races	6
8	Nanotech arms races	7
9	Feedback	8
	References	8

---

\*First written: 5 Dec. 2013; last update: 29 Feb. 2016

## 1 Will governments build AI first?

AI poses a national-security threat, and unless the militaries of powerful countries are very naive, it seems to me unlikely they'd allow AI research to proceed in private indefinitely. At some point the US military would confiscate the project from Google or Facebook, if the US military isn't already ahead of them in secret by that point.

While the US government as a whole is fairly slow and incompetent when it comes to computer technology, specific branches of the government are on the cutting edge, including the NSA and DARPA (which already funds a lot of public AI research). When we consider historical examples as well, like the Manhattan Project, the Space Race, and ARPANET, it seems that the US government has a strong track record of making technical breakthroughs when it really tries.

Sam Altman [agrees](#) that in the long run governments will probably dominate AI development: "when governments gets serious about [superhuman machine intelligence] SMI they are likely to out-resource any private company".

There are *some* scenarios in which private AI research wouldn't be nationalized:

- An unexpected AI foom before anyone realizes what was coming.
- The private developers stay underground for long enough not to be caught. This becomes less likely the more government surveillance improves (see "[Arms Control and Intelligence Explosions](#)" by Shulman

and Armstrong (2009)).

- AI developers move to a "safe haven" country where they can't be taken over. (It seems like the international community might prevent this, however, in the same way it now seeks to suppress terrorism in other countries.)

Each of these scenarios could happen, but it seems reasonably likely to me that governments would ultimately control AI development, or at least partner closely with Google.

## 2 AI arms races

Government AI development could go wrong in several ways. Plausibly governments would both the process by not realizing the risks at hand. It's also possible that governments would use the AI and robots for totalitarian purposes.

It seems that both of these bad scenarios would be exacerbated by international conflict. Greater hostility means countries are more inclined to use AI as a weapon. Indeed, whoever builds the first AI can take over the world, which makes building AI the ultimate arms race. A [USA-China race](#) is one reasonable possibility.

Arms races encourage risk-taking – being willing to skimp on safety measures to improve your odds of winning ("[Racing to the Precipice](#)" by Armstrong, Bostrom, and Shulman (2016)). In addition, the weaponization of AI could lead to worse expected outcomes in general. [CEV](#) (Yudkowsky, 2004) seems to have less hope of success in a Cold War sce-

nario. ("What? You want to include the evil *Chinese* in your CEV?") With a pure CEV, presumably it would eventually count Chinese values even if it started with just Americans, because people would become more enlightened during the process. However, when we imagine more crude democratic decision outcomes, this becomes less likely.

In *Superintelligence: Paths, Dangers, Strategies* (Ch. 14), Nick Bostrom (2014) proposes that another reason AI arms races would crimp AI safety is that competing teams wouldn't be able to share insights about AI control. What Bostrom doesn't mention is that competing teams also wouldn't share insights about AI *capability*. So even if less inter-team information sharing reduces safety, it also reduces speed, and the net effect isn't clear to me.

Of course, there are situations where arms-race dynamics can be desirable. In the original prisoner's dilemma, the *police* benefit if the prisoners defect. Defection on a tragedy of the commons by companies is the heart of [perfect competition](#)'s efficiency. It also underlies competition among countries to improve quality of life for citizens. Arms races generally speed up innovation, which can be good if the innovation being produced is both salutary and not risky. This is not the case for general AI. Nor is it the case for other "races to the bottom".

### 3 Ways to avoid an arms race

Averting an AI arms race seems to be an important topic for research. It could be partly informed by the Cold War and other nuclear

arms races,

as well as by [other efforts](#) at nonproliferation of chemical and biological weapons. Forthcoming robotic and [nanotech weapons](#) might be even better analogues of AI arms races than nuclear weapons because these newer technologies can be built more secretly and used in a more targeted fashion.

Apart from more robust arms control, other factors might help:

- Improved international institutions like the UN, allowing for better enforcement against defection by one state.
- In the long run, a scenario of [global governance](#) would likely be ideal for strengthening international cooperation, just like nation states [reduce intra-state violence](#).
- Better construction and enforcement of nonproliferation treaties.
- Improved game theory and international-relations scholarship on the causes of arms races and how to avert them. (For instance, arms races have sometimes been modeled as iterated prisoner's dilemmas with imperfect information.)
- How to improve verification, which has historically been a weak point for nuclear arms control. (The concern is that if you haven't verified well enough, the other side might be arming while you're not.)
- Moral tolerance and multicultural perspective, aiming to reduce people's sense of nationalism. (In the limit where neither Americans nor Chinese care which government wins the race, there would be no point in having the race.)

- Improved trade, democracy, and other forces that historically have reduced the likelihood of war.

#### 4 Are these efforts cost-effective?

World peace is hardly a goal unique to effective altruists (EAs), so we shouldn't necessarily expect low-hanging fruit. On the other hand, projects like nuclear nonproliferation seem relatively underfunded even compared with anti-poverty charities.

I suspect more direct [MIRI](#) -type research has higher expected value, but among EAs who don't want to fund MIRI specifically, encouraging donations toward international cooperation could be valuable, since it's certainly a more mainstream cause. I wonder if GiveWell would consider studying global cooperation specifically beyond its [indirect relationship](#) with catastrophic risks.

#### 5 Should we publicize AI arms races?

When I mentioned this topic to a friend, he pointed out that we might not want the idea of AI arms races too widely known, because then governments might take the concern more seriously and therefore start the race earlier – giving us less time to prepare and less time to work on FAI in the meanwhile. From David Chalmers (2010), "[The Singularity: A Philosophical Analysis](#)" (footnote 14):

When I discussed these issues with cadets and staff at the West Point Military Academy, the question arose as to whether the US military or other branches of the

government might attempt to prevent the creation of AI or AI+, due to the risks of an intelligence explosion. The consensus was that they would not, as such prevention would only increase the chances that AI or AI+ would first be created by a foreign power. One might even expect an AI arms race at some point, once the potential consequences of an intelligence explosion are registered. According to this reasoning, although AI+ would have risks from the standpoint of the US government, the risks of Chinese AI+ (say) would be far greater.

We should take this information-hazard concern seriously and remember the [unilateralist's curse](#) (Bostrom, Douglas, & Sandberg, 2016). If it proves to be fatal for explicitly discussing AI arms races, we might instead encourage international cooperation without explaining *why*. Fortunately, it wouldn't be hard to encourage international cooperation on grounds other than AI arms races if we wanted to do so.

Also note that a government-level arms race could easily be *preferable* to a Wild West race among a dozen private AI developers where coordination and compromise would be not just difficult but potentially impossible. Of course, if we did decide it was best for governments to take AI arms races seriously, this would also encourage private developers to step on the gas pedal. That said, once governments do recognize the problem, they may be able to impose moratoria on private development.

How concerned should we be about accidentally accelerating arms races by talking about them? My gut feeling is it's not too risky, because

- It's hard to contain the basic idea. Super-powerful AI is already well known not just by governments but even in popular movies.
- Developing verification measures, technology restrictions, and so on require governments knowing what technology they're dealing with.
- If governments can think about these issues ahead of time (decades before strong AI becomes feasible), they're more likely to go for cooperation and less likely to panic and build up their own defenses, because they see that there's time for negotiations to potentially work before losing that much ground. Right now most AI research appears to be done in public, so there's not a huge cost for a given country in delaying at this point.
- Most risk analysts don't express concerns like these too much when talking about military arms races. Of course, there's selection bias; maybe most of the military does think it's dangerous to talk about these issues in public, and we only hear from the minority that defects from this view. But I've never heard criticism against people who talk too much about arms races in public, except this one comment from my friend.
- Talking about arms-race scenarios specifically makes it much more clear *why* we

need global governance and improved cooperation. It's more persuasive than just saying, "Wouldn't it be great if the world could sing Kumbaya?"

That said, I remain open to being persuaded otherwise, and it seems important to think more carefully about how careful to be here. The good news is that the information hazards are unlikely to be disastrous, because all of this material is already publicly available somewhere. In other words, the upsides and downsides of making a bad judgment seem roughly on the same order of magnitude.

## 6 How do our prospects look?

In *Technological change and nuclear arms control* (1986), Ted Greenwood suggests that arms control has historically had little counterfactual impact:

In no case has an agreement inhibited technological change that the United States both actually wanted to pursue at the time of agreement and was capable of pursuing during the intended duration of the agreement. Only in one area of technological innovation (i.e., SALT II constraints on the number of multiple independently-targetable reentry vehicles, or MIRVs, on existing missiles) is it possible that such agreements actually inhibited Soviet programs, although in another (test of new light ICBMs [intercontinental ballistic missiles]) their program is claimed by the United States to violate the SALT II Treaty that the Soviets have stated they

will not undercut.

In "Why Military Technology Is Difficult to Restrain", Greenwood (1990) adds that the [INF Treaty](#) was arguably more significant, but it still didn't stop technological development, just a particular application of known technology.

John O. McGinnis (2010) [argues against](#) the feasibility of achieving global cooperation on AI:

the only realistic alternative to unilateral relinquishment would be a global agreement for relinquishment or regulation of AI-driven weaponry. But such an agreement would face the same insuperable obstacles nuclear disarmament has faced. [...] Not only are these weapons a source of geopolitical strength and prestige for such nations, but verifying any prohibition on the preparation and production of these weapons is a task beyond the capability of international institutions.

In other domains we also see competition prevail over cooperation, such as in most markets, where usually there are at least several companies vying for customers. Of course, this is partly by social design, because we have anti-trust laws. Competition in business makes companies worse off while making consumers better off. Likewise, competition to build a quick, hacky AI makes human nations worse off while perhaps making the unsafe AIs better off. If we care some about the unsafe AIs for their own sakes as intelligent [preference-satisfying agents](#), then this is less of a loss than

it at first appears, but it still seems like there's room to expand the pie, and reduce suffering, if everyone takes things more slowly.

Maybe the best hope comes from the possibility of global unification. There is just one US government, with a monopoly on military development. If instead we had just one world government with a similar monopoly, arms races would not be necessary. Nationalism has been a potent force for gluing countries together and if channeled into internationalism, perhaps it could help to bind together a unified globe. Of course, we shouldn't place all our hopes on a world government and need to prepare for arms-control mechanisms that can also work with the present-day nation-state paradigm.

## 7 Robot arms races

Robots require AI that contains clear goal systems and an ability to act effectively in the world. Thus, they seem like a reasonable candidate for where artificial general intelligence will first emerge. Facebook's image-classification algorithms and Google's search algorithms don't need *general* intelligence, with many human-like cognitive faculties, as much as a smart robot does.

Military robotics seems like one of the most likely reasons that a robot arms race might develop. Indeed, to some degree there's already an arms race to build drones and autonomous weapons systems. [Mark Gubrud](#):

Killer robots are not the only element of the global technological arms race,

but they are currently the most salient, rapidly-advancing and fateful. If we continue to allow global security policies to be driven by advancing technology, then the arms race will continue, and it may even reheat to Cold War levels, with multiple players this time. Robotic armed forces controlled by AI systems too complex for anyone to understand will be set in confrontation with each other, and sooner or later, our luck will run out.

## 8 Nanotech arms races

Nanotechnology admits the [prospect of severe arms races](#) as well. "[Can an MM Arms Race Avoid Disaster?](#)" lists many reasons why a nanotech race should be less stable than the nuclear race was. In "[War, Interdependence, and Nanotechnology](#)," Mike Treder suggests that because nanotech would allow countries to produce their own good and energy with less international trade, there would be less incentive to refrain from preemptive aggression. Personally, I suspect that countries would still be very desperate to trade *knowledge* about nanotech itself to avoid falling behind in the race, but perhaps if a country was the world's leader in nanoweapons, it would have incentive to attack everyone else before the tables turned.

Mark Gubrud's (1997) "[Nanotechnology and International Security](#)" presents an excellent overview of issues with both AI and nanotech races. He suggests:

Nations must learn to trust one another

enough to live without massive arsenals, by surrendering some of the prerogatives of sovereignty so as to permit intrusive verification of arms control agreements, and by engaging in cooperative military arrangements. Ultimately, the only way to avoid nanotech confrontation and the next world war is by evolving an integrated international security system, in effect a single global regime. World government that could become a global tyranny may be undesirable, but nations can evolve a system of international laws and norms by mutual agreement, while retaining the right to determine their own local laws and customs within their territorial jurisdictions.

According to Jürgen Altmann's talk, "[Military Uses of Nanotechnology and Nanoethics](#)," 1/4 to 1/3 of US federal funding in the National NT Initiative is for defense – \$460 million out of \$1.554 billion in 2008 (video time: 18:00). The US currently spends 4-10 times the rest of the world in military nanotech R&D, compared with "only" 2 times the rest of the world in overall military R&D (video time: 22:28). Some claim the US should press ahead with this trend in order to maintain a monopoly and prevent conflicts from breaking out, but it's dubious that nanotech can be contained in this way, and Altmann instead proposes active arms-control arrangements with anytime, anywhere inspections and in the long run, progress toward global governance to allay security dilemmas. We have seen many successful bans on classes of tech-

nology (bioweapons, chemical weapons, blinding lasers, etc.), so nano agreements are not out of the question, though they will take effort because many of the applications are so inherently dual-use. Sometimes commentators scoff at enforcement of norms against use of chemical weapons when just as many people can be killed by conventional forces, but these agreements are actually really important, as precedents for setting examples that can extend to more and more domains.

Like AI, nanotech may involve the prospect of the technology leader taking over the world. It's not clear which technology will arrive first. Nanotech contributes to the continuation of Moore's law and therefore makes brute-force evolved AI easier to build. Meanwhile, AI would vastly accelerate nanotech. Speeding up either leaves less time to prepare for both.

## 9 Feedback

To read comments on this piece, see the [original LessWrong discussion](#).

## References

- Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & SOCIETY*, 31(2), 201–206.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: OUP Oxford.
- Bostrom, N., Douglas, T., & Sandberg, A. (2016). The unilateralist's curse and the case for a principle of conformity. *Social Epistemology*, 30(4), 350–371.
- Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10), 7–65.
- Greenwood, T. (1990). Why military technology is difficult to restrain. *Science, Technology & Human Values*, 15(4), 412–429.
- Gubrud, M. A. (1997). Nanotechnology and international security. In *Fifth foresight conference on molecular nanotechnology* (Vol. 1).
- McGinnis, J. O. (2010). Accelerating AI. *Nw-UL Rev.*, 104, 1253.
- Shulman, C., & Armstrong, S. (2009). Arms control and intelligence explosions. In *7th european conference on computing and philosophy (ecap), bellaterra, spain, july* (pp. 2–4).
- Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*.