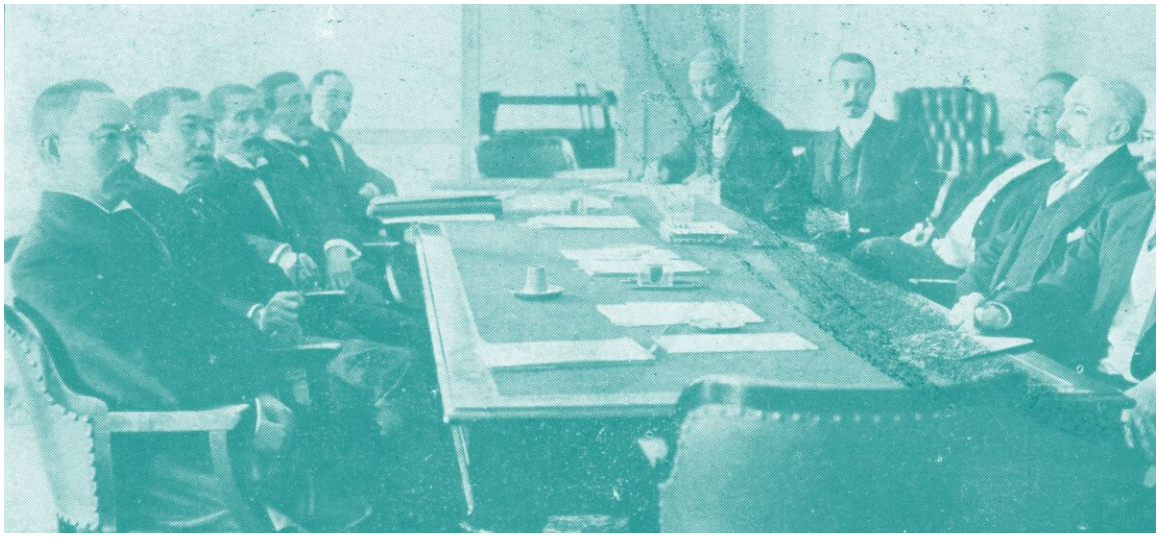
15

PART III · THE SILICON SHIELD

Supernegotiation

Govind Pimpale



Russian and Japanese peace envoys meet at Portsmouth Navy Yard, New Hampshire, 1905. A settlement depended on each side understanding the other's demands and the costs of further fighting.
Library of Congress.

SUMMARY

Rapid AI progress may increase the incentives for great powers to engage in wars. This is because AI-accelerated military R&D means that each side knows far less about what the other's military capabilities are, meaning that they are more likely to misjudge how costly fighting a war would be. Wars between great powers have historically been devastating. To avert this, we need to apply our progress in AI to negotiating deals, in addition to fighting wars.

AI negotiators allow us to leverage private information about military capabilities in negotiation: they can be shown secrets, bargain on it, and then verifiably forget it. We can also enhance negotiation by using mechanism design to make bluffs costly. We should start testing new forms of negotiation in diplomacy or commercial settings.

THE BULK OF TRANSFORMATIVE AI'S IMPACT ON A NATION'S WARFIGHTING ABILITY is likely to come from AI-assisted R&D instead of direct AI integration. Much of aerospace engineering—the kind needed to build cruise missiles, fighter jets,

SATELLITE CONSTELLATIONS, AND OTHER HARDWARE CRITICAL TO MILITARY DOMI-nance—is white-collar labor susceptible to automation by transformative AI. Advanced robotics for manufacturing will only accelerate this phenomenon, by enabling rapid prototyping and buildout. Today, the average major U.S. weapons program takes 12 years to deliver an initial capability.[1] In the future we are considering here, that time might be far shorter: a year or less. Additionally, automated intellectual labor will accelerate cyber vulnerability discovery, strategy work, and other non-tangible advantages.

These factors, together, make the military capabilities of any given country far more opaque than they used to be, even given the parallel effect of transformative AI on intelligence agencies. There will simply be fewer humans to extract information from and less time to gather facts.

This increased opacity is one factor why transformative AI is likely to increase the incentives for war. But first, we need to understand why nations fight wars in the first place. In a foundational 1995 paper,[2] scholar James Fearon notes that wars are *costly*. Nations that go to war will experience much loss of human life and destruction of infrastructure. Furthermore, wars are also *uncertain*—no one knows who will prevail *ex ante*. These two factors imply that there should always exist a peaceful negotiated settlement that both actors rationally prefer over going to war. Yet, wars still exist. Why?

One major reason, for Fearon, is private information. Nations have information that they cannot credibly communicate to counterparties. This information comes in two types: information about their resolve—that is, under what conditions they would go to war, and under what conditions they would sue for peace—and information about their military capabilities. But, even weak nations will bluff that their resolve is unbreakable, and that their military capabilities are unbeatable, to get the other party to offer them more favorable terms. Since bluffs cannot be distinguished from the truth, each side is likely to misjudge the other. If both a nation and its adversary think that they are the favorite in any hypothetical war, then it makes war much likelier, since both sides may see the actual fair settlement as worse than what they could get by going to war.

To avert this, a nation can prove that it has some military capability via a demonstration: performing nuclear tests, launching missiles, or conducting fleet exercises. This only works if the demonstration is not too costly. However, other military capabilities are destroyed if they are ever publicly disclosed. For example, if it turned out that the U.S. had managed to secretly transport a nuclear bomb into a Beijing warehouse, telling this to the Chinese negotiating team would immediately wipe out the military advantage. Unfortunately, most of the non-hardware advantages conferred by transformative AI will fall into the latter category. Cyber, intelligence, and strategic advantages all tend to be easily counterable by an enemy who becomes aware of the information.

1 U.S. Government Accountability Office, *Weapon Systems Annual Assessment: Requiring Mature Technologies Could Enable Shift to Rapid Delivery*, GAO-26-108457 (2026).

2 James D. Fearon, "Rationalist Explanations for War," *International Organization* 49, no. 3 (1995): 379–414.

This means that a state with such advantages cannot benefit from them during negotiation, only during war.

The real problem is not disclosure of the sensitive military information *per se*, but rather that human negotiators cannot verifiably forget information after they are told it. Thus, any disclosure to human negotiating teams inevitably results in that information being propagated to leadership, which then takes actions to negate the advantage.

But AI, unlike humans, can credibly commit to forgetting information. Returning to the example of the nuke in Beijing, if the Chinese and U.S. negotiating teams were AIs, then they could discuss this fact amongst themselves and write a treaty that is more favorable to the U.S. than it otherwise would have been (reflecting the advantage the bomb gives them). Then, after they settle on the treaty, the AIs would have their memories verifiably wiped. To prevent lying about false secret capabilities, the negotiating teams would need to go into the room with difficult-to-forge proof that such a capability did exist.

In order to realize this ability, though, we need both sides to develop AI models that are capable of wargaming out scenarios and understanding the probability of victory fairly. Both sides need to trust their negotiators in order for the process to work. Luckily, this is a capability that each nation's military has every incentive to develop: as AI models grow smarter and faster, an AI general commanding autonomous weapons will likely be the *de facto* mode of operation. Realistic evaluation environments will therefore be plentiful. Making this evaluation infrastructure mutually auditable is a difficult but tractable engineering challenge.

More difficult than the issue of private information relating to military capability is that of private information relating to resolve. As mentioned earlier, states have an incentive to bluff about their willingness to go to war, or at what cost they would be willing to fight. For example, would the U.S. go to war in defense of Taiwan? The U.S. wants China to believe that it would, but in actuality, it might not. If China knew that America might actually not fight, it would be heavily incentivized to invade, knowing that repercussions would be minimal. However, unlike with the private information relating to military capabilities, oftentimes the true answer of what the U.S. would do is set by fickle politics as opposed to cold military logic. Furthermore, while it might be feasible to bring hard-to-fake proof of secret military capabilities, it is difficult to bring proof that a nation is actually resolved to take action.

There are a few potential solutions. One is mechanism design. Right now, talk is cheap. The AI negotiators could bluff that the U.S. is willing to accept the destruction of a few cities in order to totally destroy China. We could design the negotiation process such that the settlement actually involves giving money (or some other good like compute) in proportion to how much damage the negotiators accepted during the wargame. Thus, bluffs of accepting damage are now costly.

Second, and far more speculatively, we might be able to build verifiably accurate behavioral models of the relevant political leaders which could be queried for what they would do in the relevant situations. We already have scanned a fruit fly brain,[3] and it is plausible that in the future, human brains might be scanned as well. Alternatively, if neurotechnology is not quite as advanced, one could create a "beta upload,"[4] an AI model trained on the actions of a person, and designed to emulate his or her externally observable behavior. This would allow the negotiators to call the other side's bluff by simulating what the leaders might actually do.

Making negotiation better is a tractable problem, one that deserves far greater investment than it currently receives. But because government officials will be reluctant to try novel techniques on such a high-stakes problem, we should begin using AI-enabled diplomacy on lower-stakes problems to iron out the issues. There are many such lower-stakes cases: trade deals with smaller countries, negotiations with corporations, and track II diplomacy (informal diplomacy between non-state actors).

The place to start is probably the private sector, however. There are already strong financial incentives to make the highest value deals, and the stakes are far lower. AI-enabled negotiation may become a *de facto* standard there before anywhere else if it provides value.

ABOUT THE AUTHOR

Govind Pimpale is a research fellow at the Foundation for American Innovation, where he focuses on AI policy. Before joining FAI, he worked at a startup developing reinforcement learning environments, and prior to that, as an AI evaluations researcher. He holds a BS in computer science and engineering from UCLA.



From *Futurecraft: The Day After AGI*, Foundation for American Innovation, 2026.
Read or listen online at futurecraft.thefai.org/essays/supernegotiation.

© 2026 by the Foundation for American Innovation. All rights reserved.

3 Robert Sanders, "Researchers Simulate an Entire Fly Brain on a Laptop. Is a Human Brain Next?" *Berkeley News*, October 2, 2024.

4 Vishal Prasad, "Beta Uploading: A Survey," *vishalblog*, March 11, 2026.