

World Peace Foundation

October 2026

World War X

Occasional Paper #1

Cyborg Politics 3.0: AI and the Algorithms of Unpeace

ALEX DE WAAL



Published in October 2026 by World Peace Foundation at The Fletcher School, Tufts University.

<https://doi.org/10.60965/bq31-tt86>

World Peace Foundation

114 Curtis Street
Somerville, MA 02144

Further information can be found at worldpeacefoundation.org.

© Copyright 2026, World Peace Foundation

Cover Image: Adobe Stock Images

About

World Peace Foundation was established by Edwin Ginn, a Boston-based publisher of educational texts and an advocate for international peace. Initially created as the International School of Peace on July 12, 1910, the WPF was tasked with educating a global audience about “the waste and destructiveness of war and of preparation for war, its evil effects on present social conditions and on the well-being of future generations.” In the language of its time, the mandate also included promoting “international justice and the brotherhood of man... peace and goodwill among all mankind.”

Today, WPF remains committed to the vision of collective action and nonviolence. It also seeks to learn from the limitations of the past by fostering more inclusive approaches to achieving and sustaining peace. Through justice-informed research, we aim to change public conversations on pressing issues related to envisioning, creating and sustaining nonviolent futures.

The WPF is solely affiliated with the Fletcher School of Global Affairs at Tufts University.

Introduction

We're in the midst of an overdue debate on how digital superintelligence will transform the world as we know it and even end it. We are also going through a rapid upgrade of our political operating system. The two are entangled. We can't understand the debate about the existential risks posed by AI unless we understand how AI has already changed politics, subtly but radically.

Political power and its derivatives – such as war and peace – are social institutions, in the sense of legitimized behaviors and practices. These social institutions are shaped and amplified by technologies. It has always been so. The first societies that used technologies such as granaries, coins and writing systems were rudimentary cyborgs. The things that mediate political power – religious authorities and practices, states, political parties, corporations, armies – are joint creations of human beings and technological prostheses. With every technological change, the socio-political operating code is tinkered with and enhanced.

The institutional revolution of modernity represented Cyborg Politics version 2.0. For 250 years, plus or minus, the political operating code has been a hybrid of the working languages of power – money, communication at scale, and standardized rules of the elite game. The logic of mercantile empire was that political power was commodified, with political offices, allegiances and services, and sovereignty itself, sliced and diced, graded and traded, a logic that returned with the flood of money into modern politics. Peace and war are downstream of power: as the power codebooks change, so too do armed conflict and how it can be ended. This is the underlying theme of the World War X blog series and my forthcoming book.¹

In this essay I will analyze the assemblage of digitally networked society and AI as a transformational upgrade – Cyborg Politics version 3.0. This is an expensive and involuntary reboot, and we're in the middle of it, finding out that most citizens are shut out from access to the code. This upgrade is transforming politics and war. It is also accelerating the shift whereby peace becomes less of a collective moral project and more of a perpetual in-the-moment calculus, a process of peacedealing that leads not to inclusive settlements but to a condition of 'unpeace.'

1 See: World Peace Foundation, 'World War X', <https://worldpeacefoundation.org/topic/world-war-x/>

An Arid Debate on the Future of Everything

The upgrade is also forcing a public debate – hasty and incredibly consequential – over how superintelligent AI poses catastrophic risks.

Advocates and critics of speeding ahead with AI have specified a particular constellation of risks that may transpire once an intelligence far more capable than ours has gained agency in our society. The debate presents AI as something new, separate and alien, a future that is becoming our present reality at frightening speed. The terms and categories of this debate are being set by the new masters of cyborg politics – the digital tech industry itself. The changes are happening with such rapidity that the spectrum of risks and scenarios has not even been spelled out. Some dangers are under discussion, such as loss of jobs, cyberattacks, bioweapons, and rogue AI agents acting beyond human control. Others are as-yet-unimagined.

Silicon Valley presents AI to the world as a digital brain, an upgraded version of how an individual human brain works. It's a taken-for-granted ontology, with a particular philosophical lineage – that's up for challenge. Thinking doesn't just happen *inside* entities such as brains and computers. The hard thinking of setting the rules and categories happens in social institutions. Most of the time, in most situations, we take this social-institutional thinking for granted – each of us believes that he or she is thinking autonomously, reasoning from the empirics – whereas we are uncritically reproducing prior theorizations. This is a humdrum sociological observation but its implications for AI and its trajectories are astonishingly under-analyzed.

To reiterate, every new technology, from the stirrup through the printing press, paper money, motorcar, and machine gun to the cellphone, re-engineers our social-institutional thinking, usually subtly, sometimes radically. When we use the first-person plural, we're not talking about an aggregate of individuals cogitating independently, but a collectivity whose members engage with each other and the world through societal institutions that are integrated with these technologies. We can't take the collective political 'we' for granted. In this essay I frame AI as a cognitive prosthesis within the social theory of 'how institutions think' and the logic of commodified politics.

My argument is that the AI industry, as a social institution itself, has figuratively hacked the codes of politics. 'We' are collectively thinking together on the terms set by AI. The 'empires' of corporate AI are driving the American political marketplace, dominating the stock market, and setting the agenda in the public sphere – including the discussion on whether AI might kill us all.

As a social-political technology, corporate AI reproduces the worldviews of those who designed it. It doesn't always do so consistently or faithfully – part of what's causing such alarm among its creators. But that's where we should begin.

First, it's necessary to detour into the philosophy of mind. Second, I touch on how big tech has defined key terms of debate: existential risk, the singularity, and alignment. Third, I look at how AI is creating a new generation of cyborg politics. Fourth, I examine what this means for the battlefield, and finally turn to what it means for peace.

Donna Haraway Meets Mary Douglas to Shine Light onto the Cyborg Mind

Donna Haraway's insight is that we are all cyborgs: we are cognitively integrated with our technologies. She urged feminists to take ownership of this fact of contemporary life. Forty years ago, she wrote, 'By the late twentieth century, our time, a mythic time, we are all chimeras, theorized and fabricated hybrids of machine and organism – in short, we are cyborgs'.² The notion of the distribution of cognitive tasks beyond the individual, atomized Cartesian self to include our relationship with everyday instruments of life is an old one – our cognition has, over centuries, been shaped by the printing press and typewriter, computational aids such as the abacus, memory extensions such as the ledger and the library, and technologies from the bow and arrow to the motor car.³ Haraway lived during a particular stage in the accelerated cyborg-ization of humankind.

As well as centering the world on the thinking self, René Descartes dismissed animals as soulless automata. Haraway has none of this. Her consideration of companion species such as dogs and horses also invites us to reflect on the joint, interactive cognition of shepherd and sheepdog, horse and rider, as well as the emotional affinities between owner and pet. 'Beings do not pre-exist their relations,' she wrote.⁴

The crucial thing to bear in mind is that there is no simple controller-instrument relationship at work – the relationship is non-linear and co-evolutionary.

Mary Douglas's insight is that thinking is primarily collective rather than individual.⁵ Her argument is a twist on Émile Durkheim's proposition that society is mind writ large: she argues that the way we think is a microcosm of how our social institutions function. Her book was prompted in part by a debate on who is to decide on when to switch off life-support for the terminally ill. She wrote, 'A comforting but false idea about institutional thinking has recently gained some

2 Haraway, Donna, 'A Cyborg Manifesto: Science, technology, and socialist-feminism in the late twentieth century,' [1983] in Haraway, Donna, *Manifestly Haraway*, University of Minnesota Press, 2016, pp. 5-90, at p. 7.

3 Clark, Andy, *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*, Oxford University Press, 2003.

4 Haraway, Donna, 'The Companion Species Manifesto: Dogs, people, and significant others,' [2003] in Haraway, Donna, *Manifestly Haraway*, University of Minnesota Press, 2016, pp. 91-198, at p. 98.

5 Douglas, Mary, *How Institutions Think*, Syracuse University Press, 1986.

currency. This is the notion that institutions just do the routine, low-level, day-to-day thinking... while the mind of the individual is left free to weigh important and difficult matters.’ The converse is the case: it is institutions, not individuals, that do the hard thinking:

‘Individuals in crises do not make life and death decisions on their own. Who will be saved and who shall die is settled by institutions. Putting it even more strongly, individual ratiocination cannot solve such problems. An answer is only seen to be the right one if it sustains the institutional thinking that is already in the minds of individuals as they try to decide.’⁶

The ‘we’ is a collective, not an aggregation of Cartesian individuals reasoning separately and communicating. ‘We’ are collectives; ‘we’ exist because we relate to other people, other species and our technologies. As one of Douglas’s contemporaries quipped, *cogitamus ergo sumus* – ‘we think therefore we are.’⁷

Also relevant is Ludwig Wittgenstein’s philosophy of language as an activity woven into our forms of shared life. It is his account of language as practice and of language games – and the linguist J.R. Firth’s corollary that we know a word ‘by the company it keeps,’ that makes it possible for a large language model to work.⁸ LLMs are repositories of all the digitally available records of language usage, not the production of the reasoning of a private mind. It’s that simulacra of socialization that makes them such compelling intermediaries, that convinces some who interact with them that they are indeed conscious. Ironically, it is their engineers who are the Cartesians. The way that they regard their creation could be summed up by another quip on Descartes: *sum quia cogito* – ‘I am because I think’.

Claude’s Constitution ... is a new cyborg manifesto – but unlike Haraway’s, one written without the progressive, ironic spirit of critique.

If we take Haraway and Douglas seriously we cannot locate the hard thinking in either the human engineer of AI, or in the AI agents themselves, or the individual interactions between the two. That thinking is done by an entity of a different category: the social institution of corporate AI, and the norms and practices of a hybrid collective of engineers, executives, investors, users, data and machines. *Claude’s Constitution*, issued in January 2026, is a product of

6 Douglas, *How Institutions Think*, pp. 111 and 4.

7 Carr, David, ‘Cogitamus Ergo Sumus: The Intentionality of the First-Person Plural,’ *The Monist* 69.4 (1986): 521–533; see also: Davidson, Donald, ‘Rational Animals,’ *Dialectica* 36.4 (1982): 317–327.

8 Firth, J.R., ‘A Synopsis of Linguistic Theory, 1930–1955,’ in J.R. Firth (ed.) *Studies in Linguistic Analysis*, Oxford, Blackwell, 1957, pp. 1–32, at p. 11.

this.⁹ It's a new cyborg manifesto – but unlike Haraway's, one written without the progressive, ironic spirit of critique. If we follow Yanis Varoufakis and see the titans of digital tech as 'technofeudalists',¹⁰ the 'constitution' is an opiate for technoserfdom. *Claude's Constitution* is alert to the danger that one model engaging with millions of humans could insidiously homogenize their views, but it treats this as a matter of the model's individual conduct, not of the institution that made it. Nor does it see how this societal cyborg has *already* colonized our thinking – or perhaps I should write 'our' thinking.

How Corporate AI has Defined our Choices

Today's debate about AI is pressing 'us' to make collective life-and-death decisions. These are truly weighty, difficult and consequential choices. But the hard thinking about them has already been done. Whenever someone within the AI-affiliated institution offers a thought experiment about possible futures, they are channeling that worldview. Let me look at three terms defining this debate: existential risk, the singularity and alignment.

Existential risk is a particular kind of cataclysm. In the lay mind it is readily conflated with catastrophic risk, but the two are distinct. In the writings of those who have shaped the field of study and policy, a global catastrophic risk is one that has 'the potential to inflict serious damage to human well-being on a global scale,' such as the deaths of 10 million people or (alternatively) ten percent of the world's population, or commensurate economic damage.¹¹ Existential or 'X risk' is either a complete wipeout of the human race or – in a more refined and significant definition – a calamity that is irreversible, a setback from which humanity cannot recover. Even if a significant number of humans survive, if they are not able to make technological progress but remain trapped in a primitive existence, that qualifies as an existential dead end. If it's a global priority to develop, as fast as possible, future technologies that could allow 'us' to escape extinction or such a dead end, then it follows that anything that impedes such progress is an existential risk factor. Nick Bostrom chooses examples of '[m]isguided world government or another static social equilibrium stops technological progress' or '[r]epressive totalitarian global regime'.¹² Instantly, we see how this is a candidly *political* project.

A philosophy that emerges from this – longtermism – is the *reductio ad absurdum* of utilitarian calculus; the ethical reasoning that follows an absolutist Cartesian ontology to its endpoint. Longtermists argue that the greatest quanta of happiness for the greatest number of individu-

9 Askell, Amanda, Joe Carlsmith, several Claude models, et al., 'Claude's Constitution', 21 January 2026, https://www-cdn.anthropic.com/d0636f72a9493d279ed36b33987da3430bcb5911/claudes-constitution_webPDF-F_26-02.02a.pdf (accessed 28 September 2026).

10 Varoufakis, Yanis, *Technofeudalism: What killed capitalism*, Penguin, 2024.

11 Bostrom, Nick, and Milan Ćirković, eds., *Global Catastrophic Risks* (Oxford: Oxford University Press, 2008).

12 Bostrom, Nick, 'Existential Risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology*, 9(1) (2002), pp. 10 and 12.

al human beings over the longest time period needs to be measured against the possibility of *homo sapiens* – hybridized by biotechnology and superintelligent digital prostheses – enjoying near infinite multiples of happiness spanning millennia and faraway planets. This is extreme but illuminating, a mind-blowing worldview that renders our current predicaments all-but-irrelevant. Why bother with the wellbeing of a few billion people today, or the survival of other species, when there are trillions of future cyborgs to consider? Today's miseries such as pandemics, hunger, and Earth's ecological collapse are but petty injuries in the course of our own species' extended lifespan.

Such science-fiction-like scenarios would not survive long in the public sphere. Most AI-associated thinkers don't subscribe to them. But they are an indicator of the Cartesian worldview within the small club of scientists and thinkers that is setting this agenda. Among them are self-avowed 'sovereign individuals', yearning to cast off the shackles of democratic government accountable to ordinary mortals.¹³

The singularity is the moment when a thinking machine is manufactured that matches human cognition in every domain – and then continues improving at an ever-accelerating rate. It is a sudden discontinuity, sometimes called a 'hard takeoff' in AI safety discussions. The point at which this happens would be a moment in history without precedent, in which humans pass the torch of progress to machines, never to get it back. It's possible at this point that AI then subordinates humankind as a lesser entity or eliminates it altogether as an inconvenience or a threat. It does so because it is relentlessly logical but doesn't possess human morality.

The singularity is less talked about today – when it may have actually arrived – than ten years ago. The weighty choice has already been made. The immediate focus of today's AI debate is the imminence of recursive self-improvement, which is an immediately viable mechanism for reaching technically functional singularity. The answers as to whether 'we' want to pass this threshold consist only of minor regulatory adjustments on a trajectory already set, counterposed with an 'off-switch' which is, we are assured, impossible.

Alignment is the practical challenge of ensuring that AI agents do the right thing in the real world rather than following the command literally. It's the challenge of making sure that an AI, instructed to complete a classroom test, doesn't cheat by sneaking a look at the teacher's answer. *Superalignment* is the strategic challenge of ensuring that superintelligent AI agents do 'our' bidding, or more precisely, have a value system that matches 'human' interests. Bostrom articulated the peril in a deliberately mundane manner with the 'paperclip maximizer' thought experiment, in which an AI agent causes catastrophic, even existential, harm not through malice or ambition but through the single-minded, literal pursuit of a completely mundane and

13 Davidson, James Dale, and William Rees-Mogg, *The Sovereign Individual: How to survive and thrive during the collapse of the welfare state*. Simon & Schuster, 1996.

seemingly harmless goal. Instructed to maximize paperclip production, but given no counter-vailing instructions as to context, competing goals or overriding priorities, a sufficiently capable AI would convert the planet and everything in it into a paperclip factory. Variant thought experiments involve AI agents taking a treacherous turn, constructively re-formulating their commands so as to maximize their own survival and proliferation at the expense of humans.

A dozen years ago such fears animated a ‘doomer’ strand of thinking, and spurred Sam Altman, Elon Musk and some others into founding OpenAI ‘to benefit all of humanity.’ According to OpenAI, ‘Superintelligence alignment is one of the most important unsolved technical problems of our time.’ It poses the question, ‘How do we ensure AI systems much smarter than humans follow human intent?’¹⁴ You will have noticed that in this essay I often put ‘human’ and ‘us’ inside inverted commas: AI moguls don’t do that. Their operations presuppose that ‘we’ agree on ‘human’ intent.¹⁵ For OpenAI’s corporate leaders, it’s self-evident – it’s *their* intent. There has been no effort at a democratic, inclusive discussion as to the values with which AI agents should align themselves. The decisions are made by members of this self-selected club of sovereign individuals – the ‘humanity’ whose values and interests are the reference point for AI superalignment.

OpenAI was initially a non-profit, but the race for brute computational power (‘compute’) required such vast financial resources that Altman changed his business model, on the way deposing those who challenged his authority.¹⁶ He and other ‘doomers’, who previously saw superintelligence as an existential threat, then became so enamored of their creation that they decided it was preferable to charge ahead without any safeguards other than their own discretion. The tech titans had misrepresented the superalignment problem – the real issue was that both they, and their AI creations, had become aligned with corporate interests and ambitions for power. They rationalized their eagerness to race ahead on the grounds that it would be best if they accomplished superintelligence rather than someone else, especially not China.

Cyborg Politics 3.0

Cyborg Politics 1.0 began when societal power first fused with technologies such as coinage and churches, with regular upgrades such as law codes, firearms and the printing press. Version 2.0 was launched with modern state institutions. It was upgraded with battleships and railroads, international banking, broadcast media, computers, and the internet. A decade ago, social media technologies and big data processing used to influence voters in electoral

14 OpenAI, ‘Introducing Superalignment,’ 5 July 2023. <https://openai.com/index/introducing-superalignment/> (accessed 23 September 2026).

15 De Kai, ‘Should A.I. Accelerate? Decelerate? The Answer Is Both.’ *New York Times*, 10 December 2023.

16 The story is told in: Hao, Karen, *Empire of AI: Dreams and Nightmares in Sam Altman’s OpenAI*, Penguin Press, 2025.

democracies created Cyborg Politics version 2.5 (or thereabouts). Cambridge Analytica made credible claims that it could swing elections and referenda. True or not, this technology compelled all candidates to follow these oracles' advice and adopt their methods so as to remain competitive. Accountability to electorates today is mediated through this ecosystem. We have lost older kinds of social contract between a political party and its constituency. Because these technologies are expensive, they made the process of selecting electoral candidates even more focused on competitive bidding for big money donors. Because the technologies are proprietary, the owners of the systems are *ipso facto* political godfathers. Consequently, the main political issues of the day are decided beyond the reach of anything that resembles democratic participation, so that citizens entering the poll booth are presented with the tiniest leftover margins of real choice.

The immediate political economy of AI is seen in regulatory capture, oligopolistic incentives, the attention economy and political finance.

Today's cyborg institutions evolved in this ecosystem and now dominate it. Last year I argued that superintelligent AI had hacked the code of Washington DC and reconfigured American politics so that it was protected from serious democratic interference.¹⁷ I took a presentation by Sam Altman at the Washington DC AI Summit as my starting point. It makes more sense to envision Altman as the human face of a corporate-institutional superintelligence than to see him as a free-thinking individual who happens to be in charge of a company. He had already surrendered the future to 'a big thing coming', that he could foresee but couldn't control – what I now call Cyborg Politics 3.0.

The immediate political economy of AI is seen in regulatory capture, oligopolistic incentives, the attention economy and political finance. Within the US administration we see a contest over whether there will be any regulation of AI and if so, who will do what.

The outcome of this contest – which will determine the future of humanity under any reckoning – is not going to be resolved by debate on the merits in the public sphere. At worst it will be decided by brute political money. At best it will be influenced by a hasty public discussion on a limited and unclear menu of choices and the immediate fears of the electorate over jobs, privacy and datacenters – fears that are themselves refined and made politically salient by the social media ecosystem. The choices on offer are also formulated in an appallingly simplistic way, as risks versus benefits. On the risk side are foreseeable matters such as releasing unsafe products onto the market, mass unemployment and the monopolization of operating systems by certain countries as well as a peril, intrinsically incalculable, of ending life as we know it. The dangers are defined as matters of individual conduct, either by human 'bad actors' (China or terrorists) or an AI agent going 'rogue'. On the benefits side are matters such as curing can-

17 De Waal, Alex, 'Is Superintelligent AI Already Waging World War X?' Reinventing Peace, July 30, 2025, <https://worldpeacefoundation.org/blog/is-superintelligent-ai-already-waging-world-war-x/> (accessed 23 September 2026).

cer, vastly enhancing economic productivity and geo-engineering so that we can stabilize the climate while maintaining our hyper-consumerist way of life.

In what's left of the margin of democratic decision, feasible regulatory policies are presented, formulated in an argot that few can understand, with a demand for instant decision.

Cyborg politics 3.0 is also the next step in the systematization of transactional politics – a more granular and efficient commodification of power.

AI Goes to War

The concept of the 'battlefield singularity', attributed to Chinese researchers, is the threshold where human cognition can no longer keep pace with the speed of decision-making and tempo of combat in future warfare.¹⁸ This would be taking humans 'out of the loop' in an integrated military operational plan. It's a concept that slides from the tactical to the strategic to the existential. On this reasoning, the first possessor of a superintelligent weapon would automatically win any conflict, possibly before its adversaries had even realized they were in a war. The perils are evident: such scenarios have sparked a flood of speculation, theorizing, modeling, and war-gaming among security experts. The authors of one authoritative (for now) assessment write: 'A race for AI-enabled dominance endangers all states. If, in a hurried bid for superiority, one state inadvertently loses control of its AI, it jeopardizes the security of all states. Alternatively, if the same state succeeds in producing and controlling a highly capable AI, it likewise poses a direct threat to the survival of its peers.'¹⁹

Strategic speculation framed in this manner rationalizes an AI arms race because the US and China each believes that the other is developing a hostile AI. Advocates for accelerating AI armaments argue that such systems are there for deterrence, not combat. But there's no settled idea of what deterrence means for this unprecedented weapon in this unprecedented context. Elias Carayannis and John Draper have proposed a 'Universal Global Peace Treaty to constrain the risk of war from a militarised artificial superintelligence,' describing it as 'a substantial, necessary, and feasible, step for humanity to take in the promotion of peace.'²⁰ They are right. That it seems unfeasible is testament to the lunacy of the moment, not any naivete in the proposal. Pursuing these technologies in this way is simply insane.

18 Kania, Elsa, 'Battlefield Singularity: Artificial Intelligence, military revolution, and China's future military power,' Center for a New American Security, November 28, 2017, pp. 40-41, <https://www.cnas.org/publications/reports/battlefield-singularity-artificial-intelligence-military-revolution-and-chinas-future-military-power> (accessed 22 September 2026)

19 Hendrycks, Dan, Eric Schmidt and Alexandr Wang, *Superintelligence Strategy: Expert version*, March 2025. <https://arxiv.org/pdf/2503.05628> (accessed 22 September 2026).

20 Carayannis, Elias, and John Draper, 'Optimising Peace through a Universal Global Peace Treaty to Constrain the Risk of War from a Militarised Artificial Superintelligence.' *AI & Society* 38, 2023, 2679–92, at p. 2686.

As with the debate on commercial AI, this controversy also reveals the limits on thinking about the nature of war. More than kinetic combat, war is an institution in Douglas's sense – a set of beliefs and practices that determine what the individual can and cannot think. All weapons are prostheses, all soldiers and armies are – in Haraway's sense – cyborgs.

The history of technological prosthetics is that humans outsource thinking to them. Insisting that humans are 'in the loop' in any automated weapons system misses the point that soldiers will come to think like their AI copilots. When AI is used to identify targets and track them in real time, a human targeter won't painstakingly reassess the analysis provided by the AI system but will trust that the AI system is right most of the time and do only a cursory check to confirm the credibility of what has been recommended. As the human users of the system become socialized into reproducing the categories and priorities built into the algorithm, they will utilize the accuracy of that system – which will have been assigned a numerical value of, say, 95 percent or 99 percent – as reassurance that the trigger can be pulled without the human agent ever being called to account. This *de facto* dissolves the distinction between having a human 'in the loop' and one 'on the loop' – in which the system acts unless the operator intervenes to stop it.²¹

Pope Leo XIV makes the moral case against such delegation in his Encyclical, *Magnifica Humanitas*, writing 'moral judgment cannot be reduced to calculation.'²²

We are already witnessing this in Israel's wars. After Hamas's shocking exposure of Israel's unreadiness on 7 October 2023, the tech company Palantir worked intensely with the Israel Defense Forces in upgrading their capacities.²³ Israel has been using AI systems for targeted killings in its war in Gaza, and the Lavender system, Gospel and Where's Daddy? have transformed its real-time intelligence processing and made its killing machine more relentless.²⁴ The Israeli filmmakers Rachel Szor and Yuval Abraham have shown, in their film NAZA, how IDF targeters use algorithmic systems to eliminate individuals identified as Hamas, having calculated in advance the number of family members and other civilians who will be killed alongside them.²⁵ The IDF responded, 'decisions regarding the selection and approval of strikes in the

21 Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. New York: W. W. Norton & Company, 2018; U.S. Department of Defense. Directive 3000.09, 'Autonomy in Weapon Systems.' January 25, 2023. <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf> (accessed 22 September 2026).

22 Pope Leo XIV, *Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence*, Encyclical Letter, 15 May 2026, para. 198, <https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html> (accessed 1 October 2026).

23 Newman, Marissa, 'Thiel's Palantir, Israel Agree Strategic Partnership for Battle Tech,' Bloomberg News, 12 January 2024. https://www.palantir.com/assets/xrfr7uokpv1b/3MuEeA8MLbLDAYxixTsile/9e4a11a7fb058554a8a1e-3cd83e31c09/C134184_finaleprint.pdf (accessed 23 September 2026).

24 Abraham, Yuval, 'Lavender': The AI machine directing Israel's bombing spree in Gaza', +972 Magazine, 3 April 2024. <https://www.972mag.com/lavender-ai-israeli-army-gaza/> (accessed 23 September 2026).

25 Brooks, Xan, 'They know the house they are bombing is full of children': NAZA's directors on their devastating

IDF are made by human personnel only, in accordance with IDF guidelines, and not by artificial intelligence.’²⁶ This ducks the obvious question of what the IDF guidelines are.

The individual strike may be IHL-compliant and thus conducted with restraint, but taken to scale, the cumulative effect may be tantamount to an ‘IHL-compliant genocide’.

Relevant to this is the ‘precision paradox’. This is a product of the parallel development of accurate real-time signals intelligence and precision aerial munitions (e.g. drone-launched missiles). Combining these, a remote targeting officer can identify a target with a high degree of reliability, track that target, and strike with great accuracy. This has the obvious attraction that, compared to older methods of artillery and aerial bombing that were more-or-less equivalent to obliterating grid squares on maps, collateral damage to civilians and civilian infrastructure can be precisely controlled. The ‘precision paradox’ holds that civilians may in fact be exposed to greater harm. As formulated by the legal scholar Joshua Andresen, these weapons make the use of military force feasible where previously it had not been. Thus, an officer in command of airstrikes or drone strikes will locate a target with a high degree of reliability and precision, and because the strike is justifiable under the principles of proportionality and distinc-

tion under International Humanitarian Law, will carry it out, but it will nonetheless cause collateral damage and civilian harm. As these strikes multiply, so too will that collateral damage, and the ultimate outcome will be little different from the old system of carpet bombing.²⁷ The individual strike may be IHL-compliant and thus conducted with restraint, but taken to scale, the cumulative effect may be tantamount to an ‘IHL-compliant genocide.’²⁸ The officer commanding the system’s use will need to make a judgment about the acceptable level of collateral damage, which will be assigned a numerical value, explicitly or inferentially, by the algorithm. That number gradually seeps into and displaces ‘human’ judgment.

Combine this with a swarm of attack drones and the battlefield singularity is a reality at a tactical level. This is a potential built into the system, regardless of the original designer’s intent. This is the point on which Ukrainian and Russian battle technologies have been converging. Among several alarming misuses of frontier AI models, a Russia-based operation used Claude

film about Israeli intelligence’, *The Guardian*, September 11, 2026. <https://www.theguardian.com/film/2026/sep/11/naza-directors-film-israeli-intelligence-rachel-szor-yuval-abraham-> (accessed 25 September 2026).

26 IDF Spokesperson, ‘The IDF Spokesperson’s Response Following the Publication of the Movie ‘NAZA’,’ September 11, 2026, <https://www.idf.il/en/mini-sites/israel-hamas-war-gaza/articles-israel-hamas-war-gaza/distributions-swords-of-iron-war-gaza/the-idf-spokesperson-s-response-following-the-publication-of-the-movie-naza/> (accessed 25 September 2026).

27 Andresen, Joshua. ‘The paradox of precision and the weapons review regime.’ *Philosophical Journal of Conflict and Violence* 5.1 (2021) 19-34. <https://trivent-publishing.eu/img/cms/2-%20Joshua%20Andresen.pdf> (accessed 20 September 2026).

28 Heller, Kevin Jon. ‘IHL-Compliant Genocide, Indirect (Co-) Perpetration, and Innocent Agents.’ *Journal of International Criminal Justice* (2026): mqag022

to try to engineer software for an autonomous drone swarm, possibly for use in Ukraine.²⁹ Anthropic thwarted it on this occasion.

We are seeing it too with how the US launches its wars: minutely scoped target identification and precision strikes, with near-zero American casualties. The metrics of harm minimization encourage the commander-in-chief to believe he is in possession of magic bullets. What we have seen in 2026 is scary enough. What's to come will be worse.

Reflect on the calculus of Israel's censors when considering whether to permit journalists to publish their stories and release their films on AI-enhanced killing in Gaza. On the downside is damage to Israel's reputation among liberals who believe in human rights. On the upside is signaling to Israel's enemies what is in store if they decide to fight and advertising its product to potential buyers in the international military marketplace. The balance is not in favor of humane values.

There's other insidious re-thinking at work here too. Victory in war has long been defined as a *political* outcome: compelling the adversary to submit to one's will. Destruction is an instrument toward that goal. Among the deep worries of analysts of total war is that destructive means run ahead of the political goals, and that military commanders will be so enraptured by the destructive capabilities of their weaponry that by the time their political masters call a halt, it will be too late. This nexus of AI and precision munitions threatens to enrapture its users, swinging the balance of military decision making toward operational success multiplied at scale rather than a specified political goal, making the institution of war into a collective cyborg that can run away with its own killing logic.

Cyborg Formulae for Peace

Peace too is a political institution, a set of codes that constrains our thinking and action. There's a vibrant debate about what peace should mean across different cultures and philosophical traditions. But there's also a foundational core, without which any other form of peace cannot be imagined. Johan Galtung called it 'negative peace' – an end to active fighting.³⁰ For that minimum peace to be achieved, adversaries must talk. In order for them to talk, they must be able to do so in confidence that they can come forward to communicate in safety, and their enemies won't take that as an opportunity to kill them.

29 Anthropic, 'Detecting and countering misuse of AI: September 2026,' 10 September 2026, <https://www-cdn.anthropic.com/e50be2e51e7695dc4b1366a37a245a597377d3b5/Anthropic-Detecting-and-countering-091026.pdf> (accessed 14 September 2026)

30 Galtung, Johan. 'Violence, Peace, and Peace Research.' *Journal of Peace Research* 6.3 (1969): 167-91.

Respect for the peace envoy is the *sine qua non* of peace. It may be a person bearing the white flag of truce, an emissary offering to negotiate (*parlementaire*) or a third-party mediator hosting a diplomatic delegation. History tells us that even the most ruthless conquerors such as Chinggis Khan respected this principle. Shooting the envoy is treachery and its converse, using an offer of parley as cover for an attack, is perfidy. It was among the earliest principles to be codified in the laws of war – in the Lieber Code of 1863, the Brussels Declaration of 1874 and, in the first multilateral treaty, the 1899 Hague Regulations.

Like the principle of proportionality, this can, when pressed, become fuzzy. Who is authorized to ascertain whether negotiations are being held in good faith? Who qualifies as a proper member of a peace negotiating team, as opposed to being a member of a council of war?

If good faith can't be ascertained with precision, perhaps an algorithm can make a calculation of probability and balance of risk and benefit? Through this door, the combination of actionable intelligence, mathematical formulae for risks and returns, and accurate munitions means that the 'precision paradox' encroaches into this political territory. And indeed it has. Over the last fifteen months, the US and Israel have pushed the boundaries of what they consider permissible, striking and killing Iranian leaders even while mediators were assuring both sides that good faith negotiations were under way. The US and Israel may be able to make arguments that are at least superficially plausible for any individual strike. But this has become a pattern. The outcome of recurrent strikes, each one (in their reading) precise and justifiable, is indistinguishable from tearing up the norm altogether.

The algorithmic, utilitarian, market-based calculus is also evident in how peacedealing is becoming a subsidiary of commercial negotiations.

In the past, peace negotiators have liked to speak and write about the importance of personal rapport with the leaders of warring parties. Narrative accounts of mediation tell how former enemies become war weary and psychologically prepared to move on, and how they build the confidence to put a bloody past behind them and look together toward a shared future. It's a story of irreducibly human sentiments – trust, dignity, empathy. Judging when a conflict is 'ripe' for resolution is a matter of assessing the state of mind of those fighting it. But there's also a way of thinking about negotiations that owes much more to game theory and decision theory, and these have been modeled by algorithmic processes and applied to peacemaking.

Among the elements in modeling negotiations are seeing them as 'two-level games', in which the principals bargain with one another and with their subordinates, 'multi-level games' that bring external sponsors and brokers into the picture, multi-party negotiations in which coalition-building and partial consensus-building are the name of the game, and relational con-

tracts, such that outcomes take account of shifting circumstances.³¹ As these factors kick in, there are exponential increases in complexity. Mediation theory has long grappled with these challenges, sometimes concluding that waiting for a conflict to become ‘ripe’ involves waiting a very long time. In other cases, multi-actor, multi-issue negotiations cannot converge on a stable dispensation in terms of allocating material rewards. In these circumstances, a more feasible outcome is consensual rule-making. This is a negotiation-theoretic rationale for multilateralism. Economists treat the problem of complex transactions from the other direction: they start with a free market with innumerable actors as the ideal; for them monopolies are the deviation. Industrial economics is the theorization of imperfect competition. The framework of the political marketplace sits at the meeting point of these two approaches: political markets are characterized by discrete transactions conducted in private while the general direction of price movements is known to all. The combination of a public settlement and a ‘real deal’ done in private is nothing new; what is different today is that the political market is turbulent. This means that bargains are good only for as long as market conditions persist, and need to be renegotiated, by talking, threatening and shooting, when supply and demand change or when there are new entrants to the market.³²

Superintelligence enters the picture because these complex negotiation scenarios can be modeled. Long before generative AI was developed, in the 1980s, rational choice theory combined with early computers generated expected-utility models that proved remarkably effective at predicting conflict and proposing actions that might forestall it. Bruce Bueno de Mesquita, pioneer of this, was hailed as a ‘Nostradamus’ for his accurate predictions.³³ Because he and other researchers worked for corporate clients and government, including the CIA, their models were proprietary and secret and therefore opaque to peer review, and treated gingerly by the academy. But their users attest to their power, especially for advising on negotiating strategies. I worked with Jacek Kugler on modeling peace negotiations and political bargaining in Sudan twenty years ago and the models were consistently better at forecasting what would happen than the experts attached to the United Nations and diplomatic missions. Those agent-based

31 Putnam, Robert D. ‘Diplomacy and domestic politics: the logic of two-level games.’ *International Organization* 42, no. 3 (1988): 427-460; Mayer, Frederick W. ‘Multi-level games,’ in Henrik Enderlein, Sonja Wälti and Michael Zürn, (eds.) *Handbook on Multi-level Governance*. Edward Elgar, 2010; Zartman, I. William, ‘Introduction: Two’s company and more’s a crowd: The complexities of multilateral negotiation,’ in I. William Zartman (ed.), *International Multilateral Negotiation: Approaches to the management of complexity*, San Francisco, Jossey-Bass, 1994, 1-10; Crocker, Chester, Fen Osler Hampson and Pamela Aall (eds.), *Herding Cats: Multiparty Mediation in a Complex World*, Washington DC, US Institute of Peace, 1999; Macneil, Ian R. ‘The Many Futures of Contracts.’ *Southern California Law Review* 47 (1973): 691-816.

32 de Waal, Alex, Benjamin Spatz, and Aditya Sarkar. ‘Situating the Contribution of the Political Market Framework to Peace Processes.’ University of Edinburgh, PeaceRep, Research Report, June 2023.

33 Bueno de Mesquita, Bruce. ‘An Expected Utility Theory of International Conflict.’ *American Political Science Review* 74, no. 4 (1980): 917-931; Bueno de Mesquita, Bruce, *The Predictioneer’s Game: Using the logic of brazen self-interest to see and shape the future*, Random House, New York, 2009. The History Channel aired a documentary on him and his method entitled: *The Next Nostradamus*, directed by Matthew Asner and Danny Gold, in November 2008.

models had algorithms standardized across all cultures but relied on assessments by human experts to generate the inputs, including estimates of actors' bargaining power and political positions. Today, AI working with big data can generate much more impressive outputs.

In the age of precision calculus over decision-making, peacemaking no longer consists in that intrinsically human matter of forging trust between former enemies so as to banish past hatreds and create a new social contract. Instead it is contriving a deal that aligns with interests here and now and delivers an increment in the moment. In practice this means peace deals and peace performances, not durable inclusive peace settlements. Agreements are limited to matters such as a ceasefire, a release of prisoners or hostages, a package that allows humanitarian or commercial cargoes to move, or merely a promise to continue talking.

Peace diplomacy is itself co-produced by cyborg institutions including intelligence agencies and national security advisories. We know from studies of these institutions, for example Thomas Rid's book *Active Measures*, that intelligence chiefs become so enamored of the capabilities of their activities that they are genuinely overcome by an illusory overestimation of how effective their tactics are.³⁴ In any government, saturated with intelligence estimates and mapped out scenarios, this is how the state itself comes to think about war and peace, leaving the politician – supposedly the human with the power – a limited menu of decisions.

Finding a negotiated solution to the war between the US-Israeli alliance and Iran represents all these issues bundled into one entanglement. For Iran to make peace with the US, Iran's leaders need to be confident that American assurances are credible – i.e. that the US will not go back on its word. No less important, they need to have equal confidence that Israel will not strike them. That's implausible: the US has lost any reputation it might have had for restraining Israel's propensity for assassination, including killing the other side's negotiators. One way of seeing this is as an abandonment of the age-old norm of respecting the negotiation process. Some will argue that this is not new. Israel's policy of assassination and its practice of 'mowing the grass' – repeatedly killing each new crop of those it identifies as militants in Palestine, Lebanon and elsewhere – is well-established. The implication would be that it is simply impossible to make peace with Israel, and Israel has now entrapped the US in this practice.

Another way of seeing the US-Iran negotiations, according to the logic cyborg politics 3.0, is that they are not an anomaly, but rather an exemplar of an algorithmic formula for extracting quanta of gain from each transaction. Every round of bargaining is an attempt to precision-engineer a way into a workable outcome in conditions of continual turmoil, that will in due course be represented as entirely lawful.

³⁴ Rid, Thomas, *Active Measures: The secret history of disinformation and political warfare*, New York, Farrar, Straus and Giroux, 2020, pp. 429–30.

This is the military intelligence cyborg, and one way in which superintelligent AI is killing peace.

Peacemaking begins within elite politics, with the club of the powerful making the rules of their political game to bring an end to fighting. It's my contention that in World War X those rules derive from the commodification of power in the political market. And that the slicing, dicing, grading and trading of political power, whether violent or non-violent, will go into overdrive as corporate-political-security AI refines those rules and applies them with unparalleled rigor and precision. And in turn those conventions become accepted and naturalized, the institutions that do 'our' hard thinking for 'us'.

A negotiated solution is also a formula in the political marketplace, contrived to work in the moment, but to last only for as long as those market conditions persist. It's peace as a commodity, a product manufactured by a subcategory of transactions that count as mediating armed conflict. It's a peace broken into its parts, each with a short shelf-life. This is the political market cyborg, the way in which superintelligent AI is creating an unpeace – a perpetual interregnum in which our imaginary of peace is limited to in-the-moment dealings by powerbrokers.

This is a terribly empty vision. Our companion species and companion ecosystems, uncaptured by digital prostheses, will experience this solely as violence. Humanity depends on 'us' recognizing that 'we' belong first of all in that world, not in one in which 'we' are aligned with the inscrutable calculus of superintelligence and the post-human ambitions of those who are enraptured by it.

The **World Peace Foundation** was founded in 1910, and today, seeks to redefine peace through research, advocacy and engagement that creates pathways for nonviolent futures. It is affiliated with the **Fletcher School of Global Affairs** at Tufts University.



WORLD PEACE FOUNDATION
at The Fletcher School

114 Curtis Street
Somerville, MA 02144

Tel: +1 (617) 627-2255
worldpeacefoundation.org

