# CP- AI

## 1NC

### 1NC- AI CP- Long

#### Counterplan Text: Public colleges and universities ought not restrict constitutionally protected speech with the exception of artificial intelligence source code, research, and development.

#### AI speech is freedom of expression

**Weaver 14** [John Frank Weaver, May 15, 2014, "Robots Deserve First Amendment Protection," Slate Magazine, <http://www.slate.com/blogs/future_tense/2014/05/15/robots_ai_deserve_first_amendment_protection>] .html] NB

C-3PO's First Amendment claim is different. First, he would never run into an accident scene willingly—if he's there, it's only because R2-D2 pushed him there, in which case, the police should arrest that droid. Second, C-3PO isn't a tool someone uses to engage in self-expression—he creates and expresses ideas all his own. Does the First Amendment apply to robots? A literal reading of the text of the First Amendment suggests that it does: It simply states that the government "shall make no law ... abridging the freedom of speech, or of the press." Nothing there specifically suggests freedom of speech is limited to people. (In contrast, U.S. copyright and patent laws [clearly indicate that only human beings qualify as authors and inventors](http://www.slate.com/blogs/future_tense/2013/12/10/ai_intellectual_property_rights_how_artificial_intelligence_might_monetize.html).)

#### AI speech would be protected because it respects democracy and self-governance rights of rational beings

**Massaro & Norton 15** [Toni M. Massaro & Helen Norton “SIRI-OUSLY? FREE SPEECH RIGHTS AND ARTIFICIAL INTELLIGENCE” Northwestern University Law Review] NB

Arguments rooted in democracy and self-governance link freedom of speech to the political cornerstones of liberal democracy and to notions of public discourse. For example, Alexander Meiklejohn famously noted that what matters for freedom of speech is not that all speak, but that “everything worth saying shall be said.”17 Taken literally, speaker identity should be irrelevant to Meiklejohn’s inquiry, and strong AI speech should be protected no less than human speech provided that its speech contributes to the democratic process—i.e., that it is “worth saying.” More recently, Robert Post draws his theory of freedom of expression from principles of self-government under which there must be a “chain of communication . . . ‘sufficiently strong and discernible’ to sustain the popular conviction that representatives spoke for the people whom they purported to represent.”18 For Post, the First Amendment is “designed to protect the processes of democratic legitimation.”19 In his view, because corporations do not themselves “experience the value of democratic legitimation,”20 they do not themselves hold free speech rights equivalent to individuals but instead hold derivative First Amendment rights to speak in ways that “may be useful to natural persons who seek to participate in public discourse.”21 In other words, corporations “do not possess original First Amendment rights to participate in public discourse as speakers,” but they can be rights holders in ways that differ from natural persons.22 The logical extension of Post’s theory to strong AI speakers is that such speakers should also be protected if and when they produce information useful to natural persons who seek to participate in public discourse.23 That a computer, not a human, produces the useful information should not matter. To be sure, under this view, limits can and should be imposed where the speech does not serve this audience-sensitive value, and strong AIs as derivative rights holders may hold rights that differ from those held by natural persons. Other democratic speech theorists, such as Jack Balkin, argue that emerging communicative technologies require a refocus of free speech theory to protect democratic culture.24 Balkin defines democratic culture as “a culture in which individuals have a fair opportunity to participate in the forms of meaning making that constitute them as individuals.”25 That is, he goes beyond representative democracy justifications for free speech. His primary anxiety is that technologies promise wider participation but also carry the means of controlling democratic participation in new ways, and he argues for attention to the latter in theorizing about First Amendment constraints on regulation of digital networks.26 Balkin’s account focuses directly on humanness when he notes that: “Human beings are made out of culture. A democratic culture is valuable because it gives ordinary people a fair opportunity to participate in the creation and evolution of the processes of meaning-making that shape them and become part of them.”27 But he then adds that the “processes of meaning-making include both the ability to distribute those meanings and the ability to receive them.”28 Human creativity is sparked by an endless array of cultural stimuli, and AI speech can contribute to receivers’ meaning-making too. Balkin’s democratic culture perspective thus would not rule out cases in which strong AI speakers contribute to the democratic disco. Indeed, Balkin’s more explicitly ecumenical account of how humans make meaning—from a wide variety of idiosyncratically relevant sources—renders such computer speech more obviously important than do more traditional, public discourse models.

#### AI satisfies the test for legal personhood- first amendment rights have been granted to nonhuman beings in the past and affective computing is increasing

**Massaro & Norton 15** [Toni M. Massaro & Helen Norton “SIRI-OUSLY? FREE SPEECH RIGHTS AND ARTIFICIAL INTELLIGENCE” Northwestern University Law Review] NB

Legal persons thus already include not only individuals, but also corporations, unions, municipalities, and even ships, though the law makes adjustments based on their material differences from humans.38 “Legal persons” often hold a variety of legal (including constitutional) rights and duties even though they may be very different from “moral” or “natural” or “human” persons. They can sue and be sued, for example. Stating that some class of nonhuman speakers may be rights holders in certain contexts simply means that that they are legal persons in those contexts—and, to date, human status is not a necessary condition for legal personhood.39 To be sure, not all rights are, or should be, necessarily available to all legal persons. For example, that a legal person has the right to sue and be sued— or to speak—does not necessarily mean that it has, or should have, the right to vote or a right to privacy.40 In this Article we take care to focus only on the possibility of free speech rights for strong AIs, and not on any other set of constitutional rights. Second, technology is changing in ways that may at some point enable some computers to satisfy certain criteria for legal personhood. For example, one difference between computers and humans used to be human- like corporality. That difference is rapidly disappearing, as some computers are now being inserted into sophisticated and human-like physical shapes. As Ryan Calo recently observed, “robots, more so than any technology in history, feel to us like social actors.”41 Although embodiment surely will affect many important legal and policy issues,42 nothing in having a physical body need determine (though it may enhance) the “selfhood” principles of freedom of expression identified here. Computers’ inability to experience emotions offers another potential source of distinction. Computer-generated speech—whether robotic or detached from a human-like form—does not entail a speaker in possession of human emotions, with emotions’ speech-curbing as well as speech- generative potential. Nor does a computer have the human need or desire, one assumes, to communicate noninteractively with itself in the way a person might write poetry or a diary with no intention of sharing this with others.43 But here too things are changing with emerging developments in affective computing.44 Although computers today are more accurately described as capable of expressing emotions rather than as having them,45 some computer scientists do not rule out a computer one day having emotions, according to their definitions of emotions (which too are evolving as scientists learn more about the workings of human cognition and emotion).46 Most threatening to strong AI speaker claims to First Amendment coverage are theories that limit such coverage to humans precisely because they are human—i.e., simply because blood flows through their veins47— rather than because of criteria such as corporality, affect, or intentionality that are associated with humans but may (or may not) be associated with strong AI speakers at some point in the future.48 Humanness, according to this view, is both necessary and sufficient. Lawrence Solum’s response to this argument remains powerful: But if someone says that the deepest and most fundamental reason we protect natural persons is simply because they are human (like us), I do not know how to answer. Given that we have never encountered any serious nonhuman candidates for personhood, there does not seem to be any way to continue the conversation.49 In other words, speaker autonomy arguments face increasing pressure not only to identify intrinsic qualities of moral personhood that are unique to humans, but to explain why those qualities should matter for purposes of conferring free speech rights (other than that they are uniquely human). We thus agree with Solum that even speaker-driven autonomy theories do not necessarily rule out First Amendment rights for strong AI speakers.50

#### Universities are crucial to partnerships for AI research and development

**Gershgorn 1-27** [Dave Gershgorn, 1-27-2017, "A massive AI partnership is tapping civil rights and economic experts to keep AI safe," Quartz, <https://qz.com/896501/a-massive-ai-partnership-is-adding-civil-rights-organizations-to-keep-computers-from-repeating-human-mistakes/>] NB

This week, the Partnership brought on new members that include representatives from the American Civil Liberties Union, the MacArthur Foundation, OpenAI, the Association for the Advancement of Artificial Intelligence, Arizona State University, and the University of California, Berkeley. The organizations themselves are not officially affiliated yet—that process is still underway—but the Partnership’s board selected these candidates based on their expertise in civil rights, economics, and open research, according to interim co-chair Eric Horvitz, who is also a director at Microsoft Research. The Partnership also added Apple as a “founding member,” putting the tech giant in good company: Amazon, Microsoft, IBM, Google, and Facebook are already on board. The Partnership is now the most high-profile, comprehensive, and mainstream organization considering how AI will shape our future. It not only has representatives from nearly every major tech company that’s heavily invested in machine-learning research, but also has backing from organizations that routinely study the impacts of technology and bias on modern society. To succeed in its mission would mean organizing a nascent field, uncertain and fragmented in its view of how AI should be implemented, while establishing guidelines that match its members’ propitious rhetoric. The Partnership’s most recent additions suggest it is also uniquely concerned with understanding AI’s ability to create (or mimic) disparity. “In its most ideal form, [the Partnership] puts on the agenda the idea of human rights and civil liberties in the science and data science community,” says Carol Rose, the executive director of the ACLU of Massachusetts who is joining the Partnership’s board. “[This is] so the people who are developing machine intelligence are aware, mindful, and cognizant of the impact of their choices, because they’re not neutral choices. “There’s a decision on, are you going to use artificial intelligence to perpetuate biases that exist in our human society? Is artificial intelligence going to be developed in a way that serves the 1% but not the 99%? Or instead, can artificial intelligence and [machine learning] be used to address deep issues of global climate change or poverty? Those are fundamental ethical and moral issues it’s important that the scientific community engage in. What they do isn’t apolitical, it’s deeply political.” The ACLU has been doing similar work with universities like MIT and Harvard as a part of its [Technology for Liberty](https://aclum.org/our-work/projects/technology-for-liberty/) initiative. Rose says that if the Partnership turns out to be toothless, she’ll back out and stop contributing.

#### Multiple net benefits:

#### 1. AI could go rogue and hurts the economy

**Dredge 15** [Stuart Dredge, 2-18-2015, "Artificial intelligence and nanotechnology 'threaten civilisation'," Guardian, <https://www.theguardian.com/technology/2015/feb/18/artificial-intelligence-nanotechnology-risks-human-civilisation>] NB

Artificial intelligence and nanotechnology have been named alongside nuclear war, ecological catastrophe and super-volcano eruptions as “risks that threaten human civilisation” in a [report by the Global Challenges Foundation](http://globalchallenges.org/publications/globalrisks/about-the-project/). In the case of AI, the report suggests that future machines and software with “human-level intelligence” could create new, dangerous challenges for humanity – although they could also help to combat many of the other risks cited in the report. “Such extreme intelligences could not easily be controlled (either by the groups creating them, or by some international regulatory regime), and would probably act to boost their own intelligence and acquire maximal resources for almost all initial AI motivations,” suggest authors Dennis Pamlin and Stuart Armstrong. “And if these motivations do not detail the survival and value of humanity, the intelligence will be driven to construct a world without humans. This makes extremely intelligent AIs a unique risk, in that extinction is more likely than lesser impacts.” The report also warns of the risk that “economic collapse may follow from mass unemployment as humans are replaced by copyable human capital”, and expresses concern at the prospect of AI being used for warfare: “An AI arms race could result in AIs being constructed with pernicious goals or lack of safety precautions.”

#### 2. AI will take a treacherous turn- it doesn’t align with human values

**Bostrom 14** [Bostrom, Nick. “Superintelligence: Paths, dangers, Strategies,” Oxford University Press. (Director, Futrure of humanity Instiutue, University of Oxford) 2014 ] NB

An existential risk is one that threatens to cause the extinction of Earth- originating intelligent life or to otherwise permanently and drastically destroy its potential for future desirable development. Proceeding from the idea of first- mover advantage, the orthogonality thesis, and the instrumental convergence thesis, we can now begin to see the outlines of an argument for fearing that a plausible default outcome of the creation of machine superintelligence is existen- tial catastrophe. First, we discussed how the initial superintelligence might obtain a decisive strategic advantage. This superintelligence would then be in a position to form a singleton and to shape the future of Earth-originating intelligent life. What hap- pens from that point onward would depend on the superintelligence’s motivations. Second, the orthogonality thesis suggests that we cannot blithely assume that a superintelligence will necessarily share any of the final values stereotypically associated with wisdom and intellectual development in humans—scientific curi- osity, benevolent concern for others, spiritual enlightenment and contemplation, renunciation of material acquisitiveness, a taste for refined culture or for the sim- ple pleasures in life, humility and selflessness, and so forth. We will consider later whether it might be possible through deliberate effort to construct a superintel- ligence that values such things, or to build one that values human welfare, moral goodness, or any other complex purpose its designers might want it to serve. But it is no less possible—and in fact technically a lot easier—to build a superintel- ligence that places final value on nothing but calculating the decimal expansion of pi. This suggests that—absent a special effort—the first superintelligence may have some such random or reductionistic final goal. Third, the instrumental convergence thesis entails that we cannot blithely assume that a superintelligence with the final goal of calculating the decimals of pi (or making paperclips, or counting grains of sand) would limit its activities in such a way as not to infringe on human interests. An agent with such a final goal would have a convergent instrumental reason, in many situations, to acquire an unlimited amount of physical resources and, if possible, to eliminate poten- tial threats to itself and its goal system. Human beings might constitute potential threats; they certainly constitute physical resources. Taken together, these three points thus indicate that the first superintel- ligence may shape the future of Earth-originating life, could easily have non- anthropomorphic final goals, and would likely have instrumental reasons to pursue open-ended resource acquisition. If we now reflect that human beings consist of useful resources (such as conveniently located atoms) and that we depend for our survival and flourishing on many more local resources, we can see that the outcome could easily be one in which humanity quickly becomes extinct.1 There are some loose ends in this reasoning, and we shall be in a better posi- tion to evaluate it after we have cleared up several more surrounding issues. In particular, we need to examine more closely whether and how a project develop- ing a superintelligence might either prevent it from obtaining a decisive strategic advantage or shape its final values in such a way that their realization would also involve the realization of a satisfactory range of human values. It might seem incredible that a project would build or release an AI into the world without having strong grounds for trusting that the system will not cause an existential catastrophe. It might also seem incredible, even if one project were so reckless, that wider society would not shut it down before it (or the AI it was building) attains a decisive strategic advantage. But as we shall see, this is a road with many hazards. Let us look at one example right away. The treacherous turn With the help of the concept of convergent instrumental value, we can see the flaw in one idea for how to ensure superintelligence safety. The idea is that we validate the safety of a superintelligent AI empirically by observing its behavior while it is in a controlled, limited environment (a “sandbox”) and that we only let the AI out of the box if we see it behaving in a friendly, cooperative, responsible manner. The flaw in this idea is that behaving nicely while in the box is a convergent instrumental goal for friendly and unfriendly AIs alike. An unfriendly AI of suf- ficient intelligence realizes that its unfriendly final goals will be best realized if it behaves in a friendly manner initially, so that it will be let out of the box. It will only start behaving in a way that reveals its unfriendly nature when it no longer matters whether we find out; that is, when the AI is strong enough that human opposition is ineffectual. Consider also a related set of approaches that rely on regulating the rate of intel- ligence gain in a seed AI by subjecting it to various kinds of intelligence tests or by having the AI report to its programmers on its rate of progress. At some point, an unfriendly AI may become smart enough to realize that it is better off concealing some of its capability gains. It may underreport on its progress and deliberately flunk some of the harder tests, in order to avoid causing alarm before it has grown strong enough to attain a decisive strategic advantage. The programmers may try to guard against this possibility by secretly monitoring the AI’s source code and the internal workings of its mind; but a smart-enough AI would realize that it might be under surveillance and adjust its thinking accordingly.2 The AI might find subtle ways of concealing its true capabilities and its incriminating intent.3 (Devising clever escape plans might, incidentally, also be a convergent strategy for many types of friendly AI, especially as they mature and gain confidence in their own judgments and capabilities. A system motivated to promote our interests might be making a mistake if it allowed us to shut it down or to construct another, potentially unfriendly AI.) We can thus perceive a general failure mode, wherein the good behavioral track record of a system in its juvenile stages fails utterly to predict its behavior at a more mature stage. Now, one might think that the reasoning described above is so obvious that no credible project to develop artificial general intelligence could possibly overlook it. But one should not be too confident that this is so. Consider the following scenario. Over the coming years and decades, AI systems become gradually more capable and as a consequence find increasing real-world application: they might be used to operate trains, cars, industrial and household robots, and autonomous military vehicles. We may suppose that this automation for the most part has the desired effects, but that the success is punctuated by occasional mishaps—a driverless truck crashes into oncoming traffic, a military drone fires at innocent civilians. Investigations reveal the incidents to have been caused by judgment errors by the controlling AIs. Public debate ensues. Some call for tighter oversight and regulation, others emphasize the need for research and better-engineered systems—systems that are smarter and have more com- mon sense, and that are less likely to make tragic mistakes. Amidst the din can perhaps also be heard the shrill voices of doomsayers predicting many kinds of ill and impending catastrophe. Yet the momentum is very much with the growing AI and robotics industries. So development continues, and progress is made. As the automated navigation systems of cars become smarter, they suffer fewer acci- dents; and as military robots achieve more precise targeting, they cause less col- lateral damage. A broad lesson is inferred from these observations of real-world outcomes: the smarter the AI, the safer it is. It is a lesson based on science, data, and statistics, not armchair philosophizing. Against this backdrop, some group of researchers is beginning to achieve promising results in their work on developing general machine intelligence. The researchers are carefully testing their seed AI in a sandbox environment, and the signs are all good. The AI’s behavior inspires confidence—increasingly so, as its intelligence is gradually increased. At this point, any remaining Cassandra would have several strikes against her: i A history of alarmists predicting intolerable harm from the growing capabil- ities of robotic systems and being repeatedly proven wrong. Automation has brought many benefits and has, on the whole, turned out safer than human operation.   ii A clear empirical trend: the smarter the AI, the safer and more reliable it has been. Surely this bodes well for a project aiming at creating machine in- telligence more generally smart than any ever built before—what is more, machine intelligence that can improve itself so that it will become even more reliable.   iii Large and growing industries with vested interests in robotics and machine intelligence. These fields are widely seen as key to national economic com- petitiveness and military security. Many prestigious scientists have built their careers laying the groundwork for the present applications and the more ad- vanced systems being planned.   iv A promising new technique in artificial intelligence, which is tremendously ex- citing to those who have participated in or followed the research. Although safety issues and ethics are debated, the outcome is preordained. Too much has been invested to pull back now. AI researchers have been working to get to human-level artificial general intelligence for the better part of a century: of course there is no real prospect that they will now suddenly stop and throw away all this effort just when it finally is about to bear fruit.   v The enactment of some safety rituals, whatever helps demonstrate that the participants are ethical and responsible (but nothing that significantly impedes the forward charge).   vi A careful evaluation of seed AI in a sandbox environment, showing that it is behaving cooperatively and showing good judgment. After some further adjustments, the test results are as good as they could be. It is a green light for the final step . . .   And so we boldly go—into the whirling knives. We observe here how it could be the case that when dumb, smarter is safer; yet when smart, smarter is more dangerous. There is a kind of pivot point, at which a strategy that has previously worked excellently suddenly starts to backfire. We may call the phenomenon the treacherous turn. The treacherous turn—While weak, an AI behaves cooperatively (increasingly so, as it gets smarter). When the AI gets sufficiently strong—without warning or provocation— it strikes, forms a singleton, and begins directly to optimize the world according to the criteria implied by its final values. A treacherous turn can result from a strategic decision to play nice and build strength while weak in order to strike later; but this model should not be inter- preted too narrowly. For example, an AI might not play nice in order that it be allowed to survive and prosper. Instead, the AI might calculate that if it is terminated, the programmers who built it will develop a new and somewhat different AI architecture, but one that will be given a similar utility function. In this case, the original AI may be indifferent to its own demise, knowing that its goals will continue to be pursued in the future. It might even choose a strat- egy in which it malfunctions in some particularly interesting or reassuring way. Though this might cause the AI to be terminated, it might also encourage the engineers who perform the postmortem to believe that they have gleaned a valu- able new insight into AI dynamics—leading them to place more trust in the next system they design, and thus increasing the chance that the now-defunct original AI’s goals will be achieved. Many other possible strategic considerations might also influence an advanced AI, and it would be hubristic to suppose that we could anticipate all of them, especially for an AI that has attained the strate- gizing superpower. A treacherous turn could also come about if the AI discovers an unanticipated way of fulfilling its final goal as specified. Suppose, for example, that an AI’s final goal is to “make the project’s sponsor happy.” Initially, the only method available to the AI to achieve this outcome is by behaving in ways that please its sponsor in something like the intended manner. The AI gives helpful answers to questions; it exhibits a delightful personality; it makes money. The more capable the AI gets, the more satisfying its performances become, and everything goeth according to plan—until the AI becomes intelligent enough to figure out that it can realize its final goal more fully and reliably by implanting electrodes into the pleasure cent- ers of its sponsor’s brain, something assured to delight the sponsor immensely.4 Of course, the sponsor might not have wanted to be pleased by being turned into a grinning idiot; but if this is the action that will maximally realize the AI’s final goal, the AI will take it. If the AI already has a decisive strategic advantage, then any attempt to stop it will fail. If the AI does not yet have a decisive strategic advantage, then the AI might temporarily conceal its canny new idea for how to instantiate its final goal until it has grown strong enough that the sponsor and everybody else will be unable to resist. In either case, we get a treacherous turn.

#### 3. AI causes extinction– it limits treaties, deterrence, and malfunctions aren’t detectable

**Shulman 11** [Carl Shulman (Singularity Institute of Artificial Intelligence) and Stuart Armstrong (InhibOx, an organization dedicated to developing and delivering the best services and technologies in computer-aided drug discovery ). “Singularity Hypotheses: A Scientific and Philosophical Assessment.” April 13th, 2011]

II. An AI arms race may be “winner-take-all” The threat of an AI arms race does not appear to be primarily about the direct application of AI to warfare. While automated combat systems such as drone aircraft have taken on greatly increased roles in recent years (Singer, 2009; Arkin, 2009), they do not greatly disrupt the balance of power between leading militaries: slightly lagging states can use older weapons, including nuclear weapons, to deter or defend against an edge in drone warfare. Instead, the military impact of an intelligence explosion would seem to lie primarily in the extreme acceleration in the development of new capabilities. A state might launch an AI Manhattan Project to gain a few months or years of sole access to advanced AI systems, and then initiate an intelligence explosion to greatly increase the rate of progress. Even if rivals remain only a few months behind chronologically, they may therefore be left many technological generations behind until their own intelligence explosions. It is much more probable that such a large gap would allow the leading power to safely disarm its nuclear-armed rivals than that any specific technological generation will provide a decisive advantage over the one immediately preceding it. If states do take AI potential seriously, how likely is it that a government's “in-house” systems will reach the the point of an intelligence explosion months or years before competitors? Historically, there were substantial delays between the the first five nuclear powers tested bombs in 1945, 1949. 1952, 1960, and 1964. The Soviet Union's 1949 test benefited from extensive espionage and infiltration of the Manhattan Project, and Britain's 1952 test reflected formal joint participation in the Manhattan Project. If the speedup in progress delivered by an intelligence explosion were large, such gaps would allow the leading power to solidify a monopoly on the technology and military power, at much lower cost in resources and loss of life than would have been required for the United States to maintain its nuclear monopoly of 1945-1949. To the extent that states distrust their rivals with such complete power, or wish to exploit it themselves, there would be strong incentives to vigorously push forward AI research, and to ensure government control over systems capable of producing an intelligence explosion. In this paper we will discuss factors affecting the feasibility of such a localized intelligence explosion, particularly the balance between internal rates of growth and the diffusion of or exchange of technology, and consider historical analogs including the effects of the Industrial Revolution on military power and nuclear weapons. III. Accidental risks and negative externalities A second critical difference between the nuclear and AI cases is in the expected danger of development, as opposed to deployment and use. Manhattan Project scientists did consider the possibility that a nuclear test would unleash a self-sustaining chain reaction in the atmosphere and destroy all human life, conducting informal calculations at the time suggesting that this was extremely improbable. A more formal process conducted after the tests confirmed the earlier analysis (Konopinski, Marvin, & Teller, 1946), although it would not have provided any protection had matters been otherwise. The historical record thus tells us relatively little about the willingness of military and civilian leaders to forsake or delay a decisive military advantage to avert larger risks of global catastrophe. In contrast, numerous scholars have argued that advanced AI poses a nontrivial risk of catastrophic outcomes, including human extinction. (Bostrom, 2002; Chalmers, 2010; Friedman, 2008; Hall, 2007; Kurzweil, 2005; Moravec, 1999; Posner, 2004; Rees, 2004; Yudkowsky, 2008). Setting aside anthropomorphic presumptions of rebelliousness, a more rigorous argument (Omohundro, 2007) relies on the instrumental value of such behavior for entities with a wide variety of goals that are easier to achieve with more resources and with adequate defense against attack. Many decision algorithms could thus appear benevolent when in weak positions during safety testing, only to cause great harm when in more powerful positions, e.g. after extensive self-improvement. Given abundant time and centralized careful efforts to ensure safety, it seems very probable that these risks could be avoided: development paths that seemed to pose a high risk of catastrophe could be relinquished in favor of safer ones. However, the context of an arms race might not permit such caution. A risk of accidental AI disaster would threaten all of humanity, while the benefits of being first to develop AI would be concentrated, creating a collective action problem insofar as tradeoffs between speed and safety existed. A first-pass analysis suggests a number of such tradeoffs. Providing more computing power would allow AIs to either operate at superhumanly fast timescales or to proliferate very numerous copies. Doing so would greatly accelerate progress, but also render it infeasible for humans to engage in detailed supervision of AI activities. To make decisions on such timescales AI systems would require decision algorithms with very general applicability, making it harder to predict and constrain their behavior. Even obviously risky systems might be embraced for competitive advantage, and the powers with the most optimistic estimates or cavalier attitudes regarding risk would be more likely to take the lead. IV. Barriers to AI arms control Could an AI arms race be regulated using international agreements similar to those governing nuclear technology? In some ways, there are much stronger reasons for agreement: the stability of nuclear deterrence, and the protection afforded by existing nuclear powers to their allies, mean that the increased threat of a new nuclear power is not overwhelming. No nuclear weapons have been detonated in anger since 1945. In contrast, simply developing AI capable of producing an intelligence explosion puts all states at risk from the effects of accidental catastrophe, or the military dominance engendered by a localized intelligence explosion. However, AI is a dual-use technology, with incremental advances in the field offering enormous economic and humanitarian gains that far outweigh near-term drawbacks. Restricting these benefits to reduce the risks of a distant, novel, and unpredictable advance would be very politically challenging. Superhumanly intelligent AI promises even greater rewards: advances in technology that could vastly improve human health, wealth, and welfare while addressing other risks such as climate change. Efforts to outright ban or relinquish AI technology would seem to require strong evidence of very high near-term risks. However, agreements might prove highly beneficial if they could avert an arms race and allow for more controlled AI development with more rigorous safety measures, and sharing of the benefits among all powers. Such an agreement would face increased problems of verification and enforcement. Where nuclear weapons require rare radioactive materials, large specialized equipment, and other easily identifiable inputs, AI research can proceed with only skilled researchers and computing hardware. Verification of an agreement would require incredibly intrusive monitoring of scientific personnel and computers throughout the territory of participating states. Further, while violations of nuclear arms control agreements can be punished after the fact, a covert intelligence explosion could allow a treaty violator to withstand later sanctions. These additional challenges might be addressed in light of the increased benefits of agreement, but might also become tractable thanks to early AI systems. If those systems do not themselves cause catastrophe but do provide a decisive advantage to some powers, they might be used to enforce safety regulations thereafter, providing a chance to “go slow” on subsequent steps. V. Game-theoretic model of an AI arms race In the full paper, we present a simple game-theoretic model of a risky AI arms race. In this model, the risk of accidental catastrophe depends on the number of competitors, the magnitude of random noise in development times, the exchange rate between risk and development speed, and the strength of preferences for developing safe AI first. VI. Ethical implications and responses The above analysis highlights two important possible consequences of advanced AI: a disruptive change in international power relations and a risk of inadvertent disaster. From an ethical point of view, the accidental risk deserves special attention since it threatens human extinction, not only killing current people but also denying future generations existence. (Matheny, 2007; Bostrom, 2003). While AI systems would outlive humanity, AI systems might lack key features contributing to moral value, such as individual identities, play, love, and happiness (Bostrom, 2005; Yudkowsky, 2008). Extinction risk is a distinctive feature of AI risks: even a catastrophic nuclear war or engineered pandemic that killed billions would still likely allow survivors to eventually rebuild human civilization, while AIs killing billions would likely not leave survivors. (Sandberg & Bostrom, 2008). However, a national monopoly on an AI intelligence explosion could also have permanent consequences if it was used to stably establish its position. Permanent totalitarianism is one possibility (Caplan, 2008). We conclude by discussing some possible avenues for reducing these long-term risks.

## 2NC

### AT: AI Can Be Controlled

#### Can’t check AI- they will circumvent control mechanisms

**Ford 2-11** [Paul Ford, 2-11-2015, "Are We Smart Enough to Control Artificial Intelligence?," MIT Technology Review, <https://www.technologyreview.com/s/534871/our-fear-of-artificial-intelligence/>] NB

But what if it wasn’t so benevolent? Nick Bostrom, a philosopher who directs the Future of Humanity Institute at the University of Oxford, describes the following scenario in his book Superintelligence, which has prompted a great deal of debate about the future of artificial intelligence. Imagine a machine that we might call a “paper-clip maximizer”—that is, a machine programmed to make as many paper clips as possible. Now imagine that this machine somehow became incredibly intelligent. Given its goals, it might then decide to create new, more efficient paper-clip-manufacturing machines—until, King Midas style, it had converted essentially everything to paper clips. No worries, you might say: you could just program it to make exactly a million paper clips and halt. But what if it makes the paper clips and then decides to check its work? Has it counted correctly? It needs to become smarter to be sure. The superintelligent machine manufactures some as-yet-uninvented raw-computing material (call it “computronium”) and uses that to check each doubt. But each new doubt yields further digital doubts, and so on, until the entire earth is converted to computronium. Except for the million paper clips. Bostrom does not believe that the paper-clip maximizer will come to be, exactly; it’s a thought experiment, one designed to show how even careful system design can fail to restrain extreme machine intelligence. But he does believe that superintelligence could emerge, and while it could be great, he thinks it could also decide it doesn’t need humans around. Or do any number of other things that destroy the world. The title of chapter 8 is: “Is the default outcome doom?”