The Cautious Path to Strategic Advantage:

HOW MILITARIES SHOULD PLAN FOR AI

Author: Peter Eckersley

Acknowledgements: Many thanks to Shahar Avin, Miles Brundage, Cindy Cohn, Corynne McSherry, Paul Scharre, and Kerstin Vignard for helpful suggestions and comments on the content of this document.

A publication of the Electronic Frontier Foundation, 2018.

View this report online:
https://www.eff.org/wp/cautious-path-strategic-advantage-how-militaries-should-plan-ai

The Cautious Path to Strategic Advantage:

HOW MILITARIES SHOULD PLAN FOR AI

PETER ECKERSLEY
pde@eff.org

First Published, August 2018
(Version 1.1 October 2018)

Introduction: A Critical Juncture for Military Uses of AI

In June, Google executives announced that the company would be backing away from its provision of AI services to the U.S. military drone program, and would not continue that work after the Project Maven contract is completed. This was in response to a campaign from Google's own employees, with thousands calling on the company to discontinue its new defense contracting work, and some even beginning to resign over the issue.

The new AI ethics principles that Google adopted in response to the debate go beyond military questions, but they do potentially place important limits on whether the company would assist in command, control, or intelligence analysis for weapons systems or other military applications. The principles may well become a model for other major technology companies.

But regardless of any actions taken by the big tech companies, the U.S. and other governments have plenty of their own resources to assemble machine learning initiatives. This includes working with companies that have much less cultural accountability to the public, consumers, or even their own engineering staff than Google does. And whether governments are acting alone, with Silicon Valley or with other companies, militaries and their contractors need to carefully consider potential dangers, and weigh the consequences of different technology development paths, before going "all in" on AI and machine learning.

Executive Summary

We are at a critical juncture. AI technologies have received incredible hype, and indeed they have made exciting progress on some fronts, but they remain brittle, subject to novel failure modes, and vulnerable to diverse forms of adversarial attack and manipulation. They also lack the most basic sorts of common sense and judgement that humans exhibit.[1] Militaries must make sure they don't buy into the AI hype while missing the warning label. There's much to be done with machine learning, but plenty of reasons to keep it away from target selection, fire control, and most command, control and intelligence (C2I) roles in the near future, and perhaps beyond that too.

This white paper, intended primarily for the defense community, tries to illuminate and document some of those dangers and suggest an agenda for avoiding them. In addition to obvious issues, several risks will require careful study by military planners before AI is widely deployed. The U.S. Department of Defense and its counterparts have an opportunity to show leadership and move AI technologies in a direction that improves our odds of security, peace, and stability in the long run—or they could quickly push us in the opposite direction.

Part I identifies how military use of AI could create unexpected dangers and risks, laying out 4 major dangers:

- Machine learning systems can be easily fooled or subverted: neural networks are vulnerable to a range of novel attacks including adversarial examples, model stealing and data poisoning. Until these attacks are better understood and defended against, militaries should avoid ML applications that are exposed to input (either direct input or anticipatable indirect input) by their adversaries.
- The current balance of power in cybersecurity significantly favors attackers over defenders. Until that changes, AI applications will necessarily be running on insecure platforms, and this is a grave concern for command, control and intelligence (C2I), as well as autonomous and partially autonomous weapons.
- Many of the most dramatic and hyped recent AI accomplishments have come from the field of reinforcement learning (RL), but current state-of-the-art RL systems are particularly unpredictable, hard to control, and unsuited to complex real-world deployment.
- The greatest risk posed by military applications of AI, increasingly autonomous weapons and algorithmic C2I is that the interactions between the systems deployed will be extremely complex, impossible to model, and subject to catastrophic forms of failure that are hard to mitigate. This is true both of by a single military over time, and, even more importantly, between those of opposing nations. As a result, there is a serious risk of accidental conflict, or accidental escalation of conflict, if ML or algorithmic automation is used in these kinds of military applications.

Part 2 offers and elaborates on the following agenda for mitigating these risks:

- Support and establish international institutions and agreements for managing AI, and AI-related risks, in military contexts.
- Focus on machine learning applications that lie outside of the "kill chain", including logistics, system diagnostics and repair, and defensive cybersecurity.
- Focus R&D effort on increasing the predictability, robustness, and safety of ML systems
- Share predictability and safety research with the wider academic and civilian research community.
- Focus on defensive cybersecurity (including fixing vulnerabilities in widespread platforms and civilian infrastructure) as a major strategic objective, since the security of hardware and software platforms is a precondition for many military uses of AI, and the national security community has a key role to play in changing the balance between cyber offense and defense.
- Engage in military-to-military dialogue, and pursue memoranda of understanding and other instruments, agreements, or treaties to prevent the risks of accidental conflict, and accidental escalation, that would inherently be created by increasing automation of weapons systems and C2I.

Finally, **Part 3** provides strategic questions to consider in the future that are intended to help the defense community contribute to building safe and controllable AI systems,

rather than making vulnerable systems and processes that we may regret in decades to come.

Rising Military Interest in AI

In the past year or two, militaries around the world have started to become very interested in machine learning and artificial intelligence technologies, thinking of them as a new opportunity for strategic advantage.[2] The civil society and research communities are[2] understandably nervous about that direction for a host of reasons—some quite superficial, some deep and carefully considered.[3]

AI is not a type of weapon like a cruise missile or an aircraft carrier (or a science fiction T-800 Terminator robot, for that matter). It is more helpful to think of it as an entire category of new technologies, like computers or the Internet were when they first appeared, that might be used for many different purposes. Like computers, AI might serve the interests of existing actors. It also might help increase the power of new actors. It might also be hard to control or even destabilizing to new and old actors alike. Rather than thinking categorically, it may be better to ask: how extensive are the appropriate military uses of AI and ML? Where would it be more strategically prudent to avoid AI deployment in general or arms race dynamics in particular?

Some areas of military action may be aided by, or even made more safe by, AI deployment, but other important areas require consideration of deliberate unilateral or multilateral restraint in the development and deployment of AI in weapons. Such restraint can serve the strategic interests of major powers, as well as aligning with human and humanitarian rights agendas.

Below we outline the four major factors that drive our concerns about military uses of AI for increasingly autonomous weapons and ML-assisted command and control. Each of those four problems requires either significant new institutional awareness and precautions, or major new advances in ML research, before we could be reasonably confident that military applications would not have serious unintended consequences. Each also points toward a need for cooperation among those deploying AI, with shared agreements about the limits and parameters of its use.

Then, in part in answer to those problems, we propose six positive steps that militaries around the world could take to help develop and support AI that is predictable, robust, secure, and safe in either military or civilian contexts.

Part I: How Military Use of AI Could Create Unexpected Dangers and Risks

Danger 1: Machine Learning Systems Can Be Easily Fooled or Subverted

ML and AI systems already impressively outperform humans in some domains.[4] The first of these that has direct application to military systems is for pre-defined classification tasks, where, for instance, convolutional neural networks can identify the contents of a still photographic image more accurately than humans. Though there are obstacles,[5] ML is likely to accumulate advantages over humans in a growing set of sensor domains: making sense of infrared, acoustics, and reflected visible light, and integrating information from multiple sources.

Unfortunately, ML classifiers are also currently extremely vulnerable to manipulation by clever adversaries. The phenomenon of "adversarial examples" demonstrates this: it is possible to place markings on an object that cause it to be misclassified, making something innocuous seem like a threat or vice versa. Such attacks are presently very easy if the attacker can inspect the internals of the classifying network, but they can also be conducted fairly efficiently against "black box" classifiers where the attacker cannot actually see how the classifier works. To make matters worse, such attacks can typically be performed in a way that is imperceptible to humans. Though mitigations have been developed that increase the difficulty of performing adversarial attacks, actually preventing the problem altogether remains an unsolved challenge.

Video: an "adversarial example" in the form of a turtle with markings that cause computer vision systems to misperceive it to be a rifle (https://www.youtube.com/watch?v=piYnd_wYlT8).

Other forms of attacks against neural networks have been demonstrated, including "data poisoning," where the adversary can create examples that cause a learning system to learn counterproductively, becoming much less accurate due to training;[6] and "model stealing" attacks, where the adversary can test a classification system to learn its internal structure in order to fool it more easily.

Visual classifiers are also subject to more mundane failure modes. One particularly important example is that they are only capable of recognizing and labelling entities if they have been trained on those specific entities. If a computer vision system is shown something it has never seen before, its behavior will be highly unpredictable. These systems do not know when they are seeing something strange, and are typically not able to estimate their own certainty well.[7] This makes it dangerous to deploy them outside of a controlled environment.

Until these problems are robustly solved—and current research results suggest that progress will be difficult—it would be dangerous and inappropriate to deploy computer vision systems and similar ML classifiers in military applications where they may fail to see things that they have not been trained to see or where adversaries might exploit these vulnerabilities. Of course, these limitations likely affect the vast majority of command and control applications, though applications in other domains such as logistics and systems design may be much less risky.

Danger 2: AI Systems Are Vulnerable to Hacking

Where ML and AI make military technologies more complex, there is a correspondingly greater chance this technology might be hacked or otherwise subverted, with consequences that can be much more severe than for other hackable technologies.[8]

We see this already in the growing presence of embedded computers in all of our infrastructure: everything from prison gates to the accelerator pedals of cars. Computers are almost always hackable, and embedded (i.e. Internet of Things or "IoT") devices are particularly hard to protect against attacks because they rarely receive automatic security updates, lack rich user interfaces, and may not receive much attention from humans. These dangers are currently true both for pure military IoT devices, and civilian devices used by military personnel.[9]

When we take this basic danger and consider it for weapons systems where humans are physically absent, distant, or with limited input (e.g., approving targets selected by an algorithm), the problem becomes especially acute. In such cases, it may take longer for humans to realize that systems are compromised and under enemy control, and it may be harder (or impossible) to remedy without the option of flipping a physical off-switch, reinstalling an operating system, or reflashing a device.

Physical and other protections can make such attacks costly to execute, but the incentives, particularly in military situations, mean that nation states and even some

non-state actors may be willing to bear the costs. And, in cases where they succeed, shared code and other systems can result is an exploit that is very scalable: if you can seize control of one drone or one artillery piece with an automated targeting system, you may be able to seize control of all of them.

These forms of attack exacerbate Danger 1, the ease of fooling or subverting a system. Cyber defense of increasingly automated systems will require understanding not only code injection paths and defenses against them, but also attacks that are partially code and partially the "cognitive" adversarial attack types that are proving to be effective against neural networks.[10]

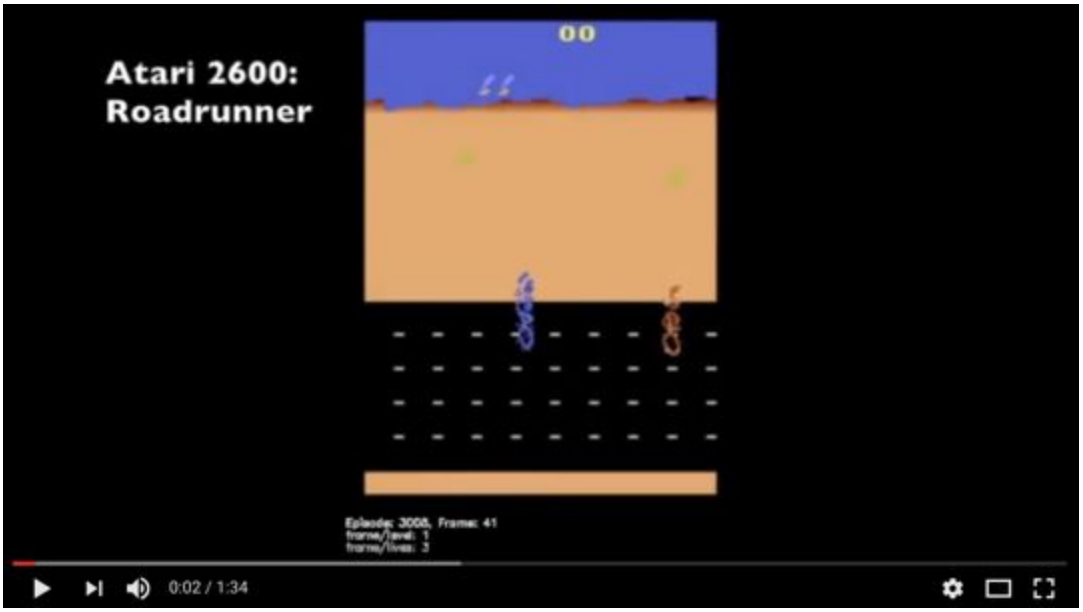
Danger 3: Reinforcement Learning Systems Have Unpredictable Dynamics

A second type of AI that militaries are likely to study and be tempted to deploy is reinforcement learning (RL). Deep RL agents differ from other kinds of neural networks in that, rather than being trained on collections of curated examples, they use their own interactions with an environment to continuously generate new data from which to learn. These systems have demonstrated an ability to outplay humans in a large number of games, ranging from Go and Chess, to simple video games like those on the Atari 2600 platform, to the beginnings of successful performance at sophisticated modern real-time strategy combat games like Dota 2 and Starcraft II.

The failure modes of these systems remain complex and poorly understood.[11] We know, for instance, that RL agents are prone to a phenomenon, called "reward hacking," in which they will follow the mathematically literal interpretation of their objective to get a higher score, rather than doing the things their designers intended.

We also know that RL agents, like most ML systems as noted above, can behave very poorly and unexpectedly in conditions for which they were not trained. Anything from rare weather phenomena,[12] to a change in human fashions or behavior or culture, to an adversary deliberately creating strange sensory data, would be expected to cause RL agents to misbehave, perhaps catastrophically.

Video: a reinforcement learning agent engages in "reward hacking." The speedboat does loops collecting bonus points, rather than finishing the race course. (J. Clark & D. Amodei, OpenAI, 2016 https://blog.openai.com/faulty-reward-functions/)



Video: another example of reward misinterpretation. The ML agent controlling the roadrunner learns to kill itself to stay on Level 1 rather than proceeding to Level 2, because it knows how to get more points on Level 1. (Saunder, Sastry, Stuhlmueller and Evans 2017)

We have only begun to understand how to test RL systems for robustness against these problems.[13] At present, it is hard enough to ensure that RL agents will always allow their owners to flip their own off switches,[14] and it is entirely likely that their capabilities will continuously exceed their reliability and safety. At the same time, we know that battlefield and other military environments are not closed systems with settled rules, like Go or even Dota 2, and that the stakes are much higher. Though there may be temptations to the contrary, it is imperative that these systems are not deployed in a military context unless and until significant breakthroughs are achieved to ensure that the systems can function robustly and predictably in a real world that is complex, open-ended, adversarial, and deeply changeable. Failure to tackle these problems before investing in military RL systems would be both strategically unwise and ethically inappropriate.

Danger 4: Automation of Escalation Pathways

The type of complexity that comes from having large numbers of deployed sensor and weapons systems with AI components would have one particularly troubling consequence: the possibility of conflicts starting by accident, as a result of interactions between the autonomous actions and reactions of those systems to each other.

Humans have begun to build and aggregate systems so complex that it has become hard to understand or reason about their properties. This type of unpredictability has been observed in both the financial system and the Internet, which have exhibited catastrophic unstable behavior in response to the actions of individual human beings.

In 2010, a trillion-dollar stock market crash was caused in part by the actions of an individual "point and click" trader interacting with much larger-scale algorithmic trading ecosystems.[15] This event was an extreme example of a more widespread and continuing form of instability.[16] Beyond the unstable short-term dynamics caused by interactions between high frequency algorithmic trading systems, financial systems have come to exhibit complex dependencies of a more structural kind. Fear of interlocking systems of leverage and dependencies between actors were what caused the 2008 financial crisis to be global and economy-wide in scope, rather than just a problem for institutions that had made poor lending decisions.[17]

Similar instability has been observed in the Internet. In 1989, a graduate student at Cornell University named Robert Morris wrote a piece of software (a "worm") that temporarily disabled a large fraction of the computers on the Internet at the time. In 2016, the Mirai malware which primarily infected digital video surveillance equipment and home routers was able to cause a series of roughly two-hour outages for numerous global Internet services (including Paypal, Github, Verizon, Comcast, AirBNB, and the BBC) by using those compromised devices to overwhelm the DNS provider Dyn.[18]

One key point about these instabilities, and a difference from military domains, is that it is generally possible to recover from cascading failures in financial and communications systems.[19] That may not be true for a cascading failure in AI used for systems for target selection, fire control, or automated response to incoming aircraft, missiles, or projectiles. While weapons with some limited automated response characteristics have been deployed for decades, expanding them to include AI processes will significantly increase both the odds of systemic failures and the difficulty of understanding and preventing them.[20]

Mitigating this risk may require new research programs, and new international institutions or agreements that bound the types of ML and control automation that the militaries of different nation states can deploy.[21] Such risks cannot be managed or bounded without careful military-to-military cooperation between rival advanced powers. While creating such cooperation faces significant trust barriers, there are some partial precedents for it, and the importance is so high that it should be attempted.[22] Until such agreements are in place, there is inherent risk in automation of command, control, and targeting processes. We especially recommend that militaries avoid forms of this automation that either remove humans from the loop or place human decision-makers in situations where they are effectively making choices based on or shaped by algorithmic recommendations.

Part II: What Can Militaries Do About AI?

Now that we have outlined some of the dangers and the need for cooperation, what can and should militaries do about and in support of AI? EFF's recommendations for the defense community[4] include:

1. **Support international agreements or institutions** to prevent or control the development of categories of systems with the potential to suffer from instabilities and unexpected complex behavior that would pose a risk to peace and stability, and be contrary to the interests of both states and civilian populations.

2. **Support civilian leadership of AI research, and focus on military applications outside the "kill chain."** It will be hard to deploy autonomous systems "safely" (i.e., without severe unintended consequences) in non-conflict environments, let alone on battlefields. [23] The forefront and the bulk of funding for AI research appears to remain in corporations, universities, and non-profit labs that are committed to positive-sum, beneficial applications. The best response to these two prevailing facts is to focus on applications of ML that are consistent with open, civilian technology leadership and that avoid domains where unintended consequences are potentially catastrophic. There are huge opportunities to take a strategic offset with AI for logistics, systems design, automated diagnostics and maintenance assistance for hardware, and defensive cybersecurity. Militaries should focus their own investments, and their cooperation with technology companies, in those directions.

3. **Focus on predictability and robustness.** There are already a large number of well-documented and complex problems with the "safety," predictability, robustness, and security of AI systems, due to the complexity of the real world and especially when confronted by adversaries. If militaries want to invest in ML, they should fund open, internationally collaborative research to address these problems, since those will be even more severe and destabilizing in military domains.[24]

4. **Make research on predictability and robustness open, and encourage international collaboration.** Robustness and safety research is a precondition to many types of military applications. As systems designers realize this, there will be a temptation to keep some of that information classified, and race other powers to develop robust and controllable systems more quickly. The temptation to race may be inevitable, but it should be resisted. Open research literature ensures that safety and robustness insights are shared, reducing the chances of unexpected catastrophes on all sides. For instance, during peacetime, it would be against U.S. interests to have Russia deploy autonomous weapons systems that have unanticipated reactions in situations that were not included in training datasets, and vice versa.

5. **Place a higher priority on defensive cybersecurity.** Both civilian applications of AI and militaries' own digital systems are profoundly insecure at present, and running vital AI deployments on insecure machines is a recipe for disaster.[25] Governments should act to make computers more trustworthy before entrusting more life-or-death tasks to them. This means larger budgets for defensive cybersecurity, both in absolute terms and relative to offensive spending. It means investments in defensive practices such as formal verification and fuzzing to protect critical open source software. It also means adjustments to policies such as the Vulnerabilities Equities Process, by which the U.S. government determines when it will reveal technology vulnerabilities to the makers of those systems and the general public, to better balance defensive with offensive objectives.

6. **Engage in military-to-military dialogue and agreements to prevent the development of systems that accidentally interact with each other and create unanticipated feedback loops.** Militaries should fund research into circumstances under which such feedback loops could occur, and design constraints, protocols. and vetting processes between states to ensure that this does not happen. Militaries should embed compliance with these protocols into their processes.

Beginning with an agenda of this sort, and proceeding carefully and thoughtfully, could allow militaries to expand their uses of AI in a way that does not risk destabilization, undermining the interests of the nations they serve, or the human rights of those potentially impacted.

Part III: Conclusions and Future Questions

This paper has surveyed a positive agenda of strategies that Western militaries could use to guide their investments in AI research and development. These do not need to undermine the militaries' objectives of strategic advantage and stability. This paper has also outlined a set of new and profound risks that must be studied and managed in order to avoid serious and problematic outcomes. Some of these risks look similar to those posed by previous generations of technologies, but others are completely novel.

Our most important conclusions are that we urgently need both certain kinds of precautionary research[8] and initiatives for multilateral cooperation to avoid unanticipated interactions between numerous deployed AI systems. Progress in these areas should take precedence over direct military applications of ML especially when those applications lie within the "kill chain" and can result in unexpected consequences or can interact with or be exploited by adversaries.

There are also a number of interesting strategic questions. For one: should the West try to claim an early strategic lead in a race for building robust, stable and controllable versions of AI? Or should it share all of its research with other great power actors in the interests of longer-term stability? Is it possible to do both at the same time? This last option seems paradoxical, but is often in fact the competitive dynamic between corporate and academic research teams in the West. It should be seriously considered as an international strategy too, given the shared interest in preventing catastrophic mistakes.]

Another major question is whether military-to-military conversations around avoiding accidental escalation risks are possible. Such conversations appear to be necessary during the design and deployment stages of any weapons systems that use algorithms for aspects of command, control, or intelligence analysis, and might not be able to rely on traditional verification rubrics. It is in the long-term interests of all sides to participate in such conversations (in game theoretic terms, cooperation is a dominant equilibrium). The challenge will be ensuring that all parties realize that they hold shared interests in the predictability and stability of whatever they build using AI.

AI has been the subject of incredible hype in recent years. Although the field is making progress, current machine learning methods lack robustness and predictability, and are subject to a complex set of adversarial attacks, problems with controllability, and a tendency to cause unintended consequences. The present moment is pivotal: in the next few years either the defense community will figure out how to contribute to the complex problem of building safe and controllable AI systems, or buy into the hype and build AI into vulnerable systems and processes that we may come to regret in decades to come. This white paper has attempted to lay out some of the largest pitfalls, and a way to pursue strategic advantage while avoiding the worst of these problems.

[1] Compare G. Marcus, Deep Learning: A Critical Appraisal https://arxiv.org/abs/1801.00631 (2018).

[2] See e.g. https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf; https://www.defense.gov/News/Article/Article/991434/deputy-secretary-third-offset-strategy-bolsters-americas-military-deterrence/; https://www.foreignaffairs.com/articles/china/2017-12-05/artificial-intelligence-and-chinese-power; https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world

[3] A USA Today survey showed that 73% of respondents were worried about the potential implications of AI and would prefer for it to be used in only limited domains: https://www.usatoday.com/story/tech/news/2018/01/02/artificial-intelligence-end-world-overblown-fears/985813001/; More than 24,000 people, including many machine learning industry and research leaders, have signed the Future of Life Institute's letter calling for a ban on autonomous weapons: https://futureoflife.org/open-letter-autonomous-weapons/.

[4] For a survey, see P. Eckersley, Y. Nasser *et al*, EFF AI Progress Measurement Project, https://eff.org/ai/metrics (2017-).

[5] See e.g. Y. Qu *et al*, "Moving vehicle detection with convolutional networks in UAV videos", *ICCAR* (2016) http://ieeexplore.ieee.org/document/7486730/?reload=true

[6] See e.g. J. Steinhardt, P. W. Koh and P. Liang, "Certified Defenses for Data Poisoning Attacks" (2018) https://arxiv.org/abs/1706.03691 for a characterization of the severity of these attacks when defenses are in place.

[7] Though there is a small amount of research beginning to tackle the question; see e.g. N. Papernot and P. McDaniel, Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning https://arxiv.org/abs/1803.04765

[8] See, e.g. https://www.schneier.com/essays/archives/1999/11/a_plea_for_simplicit.html

[9] See, e.g. https://securityledger.com/2018/01/security-personnel-challenges-stymie-dods-adoption-iot/

[10] This can happen because a cyberattack opens a vector attacking a machine learning system, but it can also involve more subtle hybrids of the two types of attack. See for instance, R. Stevens *et al.* Summoning Demons: The Pursuit of Exploitable Bugs in Machine Learning https://arxiv.org/abs/1701.04739

[11] D. Amodei *et al* "Concrete Problems in AI Safety." (2016) https://arxiv.org/abs/1606.06565.

[12] There was a historical claim that this problem had happened in computer vision projects to recognize tanks, some time between the 1960s and 1990s, that have been disputed by Gwern Branwern: https://www.gwern.net/Tanks. However the problems of systems failing because of inadequately diverse training data, or distributional shift after they are released into the world, are real. See for instance, https://dl.acm.org/citation.cfm?id=1462129 http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf, https://arxiv.org/pdf/1606.06565.pdf (Section 7), http://people.csail.mit.edu/khosla/papers/eccv2012_khosla.pdf http://ieeexplore.ieee.org/abstract/document/5995347/ https://pdfs.semanticscholar.org/0810/6464ea2a9ccbbed1e06161ed2d1e594efa05.pdf https://arxiv.org/pdf/1707.07169.pdf. For examples specifically caused by rare weather phenomena, see http://www.pbs.org/wgbh/nova/military/nuclear-false-alarms.html, https://www.tesla.com/blog/tragic-loss, http://www.autonews.com/article/20160210/OEM06/160219995/self-driving-cars-succumb-to-snow-blindness-as-driving-lanes-disappear .

[13] J. Leike *et al.*, "AI Safety Gridworlds." (2017) https://arxiv.org/abs/1711.09883.

[14] *Id.* See also, D. Hadfield-Menell et al, "The Off Switch Game" (2017) https://arxiv.org/pdf/1611.08219

[15] See https://www.bloomberg.com/news/features/2017-02-10/how-the-flash-crash-trader-s-50-million-fortune-vanished ; Y Wang, Strategic Spoofing Order Trading by Different Types of Investors in the Futures Markets, Proc. *European Financial Management Association (EFMA) Annual Meeting* 2016. http://www.efmaefm.org/0EFMAMEETINGS/EFMA%20ANNUAL%20MEETINGS/2016-Switzerland/papers/EFMA2016_0171_fullpaper.pdf ; D. Minotra and C. Burns, Understanding safe performance in rapidly evolving systems: a risk

management analysis of the 2010 U.S. financial market Flash Crash with Rasmussen's risk management framework, *Theoretical Issues in Ergonomics Science* 18 2017.

[16] "Market quality breakdowns in equities", *Journal of Financial Markets* 28 2013 https://www.sciencedirect.com/science/article/pii/S138641811630057X.
[17] See eg J. Crotty, Structural causes of the global financial crisis: a critical assessment of the 'new financial architecture', *Cambridge Journal of Economics* 33 2009. Some sources argue that in the particular case of the institutional failures that triggered the 2008 crisis, it was fear of such financial interconnectedness and systemic risk, rather than the actual dynamics of asset interconnectedness, which caused financial contagion; see H. Scott, Interconnectedness and Contagion: Financial Panics and the Crisis of 2008. (2014) https://ssrn.com/abstract=2178475. *Cf* R. Aloui *et al.* Global financial crisis, extreme interdependences, and contagion effects: The role of economic structure? *Journal of Banking & Finance* 35 2010.

[18] https://en.wikipedia.org/wiki/2016_Dyn_cyberattack#Investigation

[19] Some authors have flagged the possibility of circular dependency problems between power grids and the Internet (https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.85.5468 , https://arxiv.org/pdf/0907.1182), which are indeed serious risks that need to be mitigated.
[20] Existing systems of this sort include automated Ships Self Defense Systems (SSDSes) such as the Pahalanx CWIS; Active Protection Systems (APSes) for vehicles such as Trophy/ASPRO-A, and "loitering munitions" such as the IAI Harpy and its successors. For a survey of systems of this sort, see P. Scharre and M. Horowitz, An Introduction to Autonomy in Weapon Systems, *CNAS Working Paper*, Feb 2015.

[21] Control automation inherently contributes to this problem by making the set of systems that can interact with each other rapidly and/or without full human oversight larger; machine learning contributes by adding systems whose behavior is complex, hard to reason correctly about, and potentially changing due to ongoing training of models.

[22] There was a long history of such risk-mitigating and arms control conversations during the Cold War, from the Moscow-Washington hotline to the SALT and INF treaties. More recently, the US and China, for instance, have a track record of military-to-military cooperation and establishment of memoranda of understanding in several areas, with a similar objective of preventing accidental escalation. See P. Saunders and J. Bowie, US-China military relations: competition and cooperation, *Journal of Strategic Studies*, http://www.tandfonline.com/doi/abs/10.1080/01402390.2016.1221818 p. 14. The US and Russia have a Memorandum of Understanding on Missile Launches, which essentially addresses a simpler and easier to understand version of the problem we are discussing here: https://www.state.gov/t/isn/4954.htm, building on an earlier era of cold war efforts The distinctive challenge in preventing automated escalation from AI systems is that the risk needs to be mitigated in deep technical detail during the system specification and design process, rather than being addressable by high-level operating procedures after deployment.

[23] Military planners may not usually think about "safety" as a particularly high-priority objective of weapons design, strategy or tactics, since warfare is an inherently unsafe activity (compare, R. Danzig, Technology Roulette: Managing Loss of Control as Militaries Pursue Technological Superiority, CNAS, 2018) However the AI research community has begun to use the term "safety" in a technical sense, essentially to mean the effort to build systems that do not have significant and problematic unintended behaviors or consequences. Unfortunately, present AI systems are unsafe. In military contexts, that could mean systems which can be hacked or tricked into fratricide on a large and unprecedented scale, or the possibility that systems could commence or escalate conflicts entirely by accident, as well as lower-stakes instances of unanticipated or unintended behavior. We therefore urge readers with national security backgrounds to parse the term "safety" as having those implications.

[24] Some of these issues have been flagged by the US government's own JASON advisory panel: Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD, *JASON report JSR-16-Task-003*, MITRE corporation, section 4.2

[25] We have discussed this topic at much greater length in the Brundage, Avin, Clark, Toner, Eckersley, *et al*. The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation, https://maliciousaireport.com/ (2018).