Regulating for ‘normal AI accidents’— Operational lessons for the responsible governance of AI deployment

Matthijs M. Maas
Centre for International Law, Conflict and Crisis, Faculty of Law, Copenhagen University
Karen Blixens Plads 16, 2300 Copenhagen S, Denmark
matthijs.maas@jur.ku.dk

Abstract

New technologies, particularly those which are deployed rapidly across sectors, or which have to operate in competitive conditions, can disrupt previously stable technology governance regimes. This leads to a precarious need to balance caution against performance while exploring the resulting ‘safe operating space’. This paper will argue that Artificial Intelligence is one such critical technology, the responsible deployment of which is likely to prove especially complex, because even narrow AI applications often involve networked (tightly coupled, opaque) systems operating in complex or competitive environments. This ensures such systems are prone to ‘normal accident’-type failures which can cascade rapidly, and are hard to contain or even detect in time. Legal and governance approaches to the deployment of AI will have to reckon with the specific causes and features of such ‘normal accidents’. While this suggests that large-scale, cascading errors in AI systems are inevitable, an examination of the operational features that lead technologies to exhibit ‘normal accidents’ enables us to derive both tentative principles for precautionary policymaking, and practical recommendations for the safe(r) deployment of AI systems. This may help enhance the safety and security of these systems in the public sphere, both in the short- and in the long term.

Keywords. Ethical design and development of AI systems; AI and Law; trust and explanations in AI systems; normal accident theory

Introduction

New technologies, particularly those which are deployed rapidly across industries, or which can offer a (economic, political, military) edge, frequently disrupt previously stable international governance arrangements or power distributions. The introduction of such technology can therefore be followed by a period of uncertainty and risk, as policymakers, operators and public seek to grasp their societal and operational implications. This leads to a precarious need to balance caution against performance while exploring the ‘safe operating space’ of said critical technology.

AI: progress & applications

Artificial Intelligence is one such critical technology. In recent years, the field of artificial intelligence (AI) has advanced at a rapid pace. While AI researchers and experts still differ in their assessment of when, if ever (Plebe and Perconti 2012; Dietterich and Horvitz 2015), we may expect the achievement of ‘general’ artificial intelligence, (Baum, Goertzel, and Goertzel 2011; Armstrong and Sotala 2012; Brundage 2017; Müller and Bostrom 2016), today’s narrow artificially intelligent systems already match or outperform humans across many narrow domains. In just the past three years—to provide an incomplete list—AI systems have proven that they can meet or exceed human performance in object image recognition (Linn 2015), speech transcription and direct translation (Xiong et al. 2016; Castelvecchi n.d.; Lewis-kraus 2016). AI systems have learned how to drive (Bryant 2016); can parse paragraphs to answer questions posed (Metz 2017); recognize human faces (even in blurred pictures) and some emotions (Newman 2016); can create new encryption schemes and detect malware (Abadi and Andersen 2016; Musthaler 2016); identify crop diseases (Furness 2016); teach themselves the ancient game of go in mere days (Silver et al. 2017, 2016), and write cookbooks, news articles, music and published poetry (Mascarenhas 2016; Marshall 2017; Scholl 2015; Kleeman 2016). As a result, these systems are beginning to see wider adoption in a broad range of applications—across fields as diverse as stock markets, transport infrastructure, healthcare, agriculture, education, cybersecurity and the military.

Mapping AI Governance Challenges

Along with the immense appeal, there is a growing recognition amongst policymakers that the development and deployment of AI brings ethical, political, societal and operational risks, and will even pose systemic challenges for regulation and governance frameworks (OSTP 2016; US Senate Subcommittee on Space, Science and Competitiveness 2016; House of Commons Science and Technology Committee 2016; China's State Council 2017). These challenges can loosely be grouped in two straightforward categories—'(ab)use' and 'accident'.

AI governance challenges from (ab)use

Many of the greatest challenges posed by AI, derive from their systemic use by many actors, who deploy these systems either to enable greater legibility or facility in the exercise of power, or in a competitive context—whether economic, political, or military. Thus, domestically, the impact of introducing even 'narrow' artificial intelligence systems may result in far-reaching technological unemployment (Brynjolfsson and McAfee 2014; Hanson 2016), comprehensive erosion of privacy (Calo 2010); over-dependency on social robots (Lin, Abney, and Bekey 2011; Weng 2010); increased inequality and societal dysfunction as a result of machine bias (Crawford and Calo 2016), or as a result of (perceived) electoral manipulation, and political polarization as a result of pervasive customized 'computational propaganda' (Woolley and Howard 2016). Moreover, the automation of social-media identity theft (Bilge et al. 2009), as well as automated system-vulnerability-detection and victim-customized social-engineering attacks—as illustrated by the 'Mayhem' AI in the 2016 DARPA Grand Cyber Challenge, and by tools such as 'WifiPhisher', respectively (DARPA 2016; wifiphisher [2014] 2017)—suggest an unprecedented increase in hazards on cyberspace, from both criminal and state actors.

In a global context, meanwhile, the integration of lethal autonomous weapons systems with 'war-algorithms' (D. A. Lewis, Blum, and Modirzadeh 2016), amongst a host of other emerging AI battlefield applications (Maas, Sweijs, and De Spiegeleire 2017), raises deep ethical, legal, and operational ramifications (Nehal et al. 2016; Scharre 2016a; Roff 2014). For instance, enhanced surveillance capabilities and autonomous weapons may also disrupt established strategic landscapes, entrenching authoritarian regimes' ability to monitor dissent and deploy centralized stand-off force projection capabilities (Horowitz 2016; Horowitz, Kreps, and Fuhrmann 2016). Moreover, autonomous weapons systems enable at-cost swarming tactics, in which an adversary's defensive measures (i.e. point-defense systems) are overwhelmed by sustained simultaneous attacks. This technological innovation overturns a prevailing tactical offense-defense balance, possibly upsetting global equilibria by putting a premium on pre-emption (Rickli 2017). The deployment of dispersed autonomous systems, along with AI-enabled advances in sensing and data analysis capabilities, might even put at risk the stability of existing nuclear deterrence dyads, by rendering previously 'secure' nuclear assets (such as ballistic missile submarines) vulnerable once more (Littlefield n.d.; Courtland 2016; The Economist 2016; Lieber and Press 2017; Hambling 2016).

AI governance challenges from accident

Right along with the governance challenges posed by the 'systemic' or 'intended' deployment of AI, however, new AI governance approaches will also have to reckon with underlying operational risks from failure. While regulatory approaches or governance systems are no stranger to (industrial) accidents involving new (even robotic) technologies, they may be prone to misunderstanding the scale and intractability of the safety challenges posed by deployed AI systems. Critically, an anecdotal reliance on familiar accidents involving 'embodied' robots (such as factory robots, or autonomous cars) or on 'malfunctioning chatbots' as the type specimens for 'AI accident risks', is likely to lead us to misunderstand the causes and dynamics of cascading accidents involving networked and opaque AI systems. Worse, it may lead us to underestimate their frequency, scale, and reach.

Empirically, leading AI architectures have demonstrated risks of 'flash crashes' amongst other failure modes (Yampolskiy and Spellchecker 2016; Scharre 2016a, 35–36), and demonstrate behaviour which can be intrinsically difficult to anticipate (in the context of evolutionary programming architectures) or reconstruct and explain post-accident (in the case of deep learning and certain neural network architectures). There exists an incipient and broad research agenda examining 'concrete problems in AI Safety' (Amodei et al. 2016), to avoid unexpected and unwanted side effects as AI systems become more advanced and/or are deployed to more complex environments, but this valuable research program is at present still at an early stage.

This is a key problem, in light of the deployment and integration of increasingly capable AI systems across society—from the financial sector to the penal system, and from critical infrastructure to the military. There is a wide range of potential principles informing AI regulatory regimes, whether the precautionary principle; 'responsible research and innovation' paradigm (Stilgoe, Owen, and Macnaghten 2013; Brundage 2016); 'differential technological progress' (Bostrom 2002), or some other set of emerging norms or policy desiderata (Bostrom, Dafoe, and Flynn 2017). Yet the observation of recurring errors in AI systems suggests that these governance regimes must also consider the problem of 'normal accidents'—not as a possibility, but as an inevitability.

AI accidents as 'normal' accidents

Developed by Charles Perrow in the wake of the Three Mile Island nuclear reactor meltdown (Perrow 1984), 'normal accident theory' has been applied to understand catastrophic technological failure across a wide range of domains, ranging from the *Apollo 13*, *Challenger* and *Columbia* spacecraft accidents (Perrow 1984, 23) to false-alarms and near-accidents plaguing the US nuclear forces (Sagan 1993); and from the Air France 447 crash to the 2003 Gulf War 'Patriot fratricides'—where faulty IFF systems led semi-autonomous coalition air defenses to shoot down friendly aircraft (Scharre 2016a, 30–33), to name but a few.

'Normal accident' theory analyzes how accidents at the crux of mechanical, software, operator and organizational failures. It is this systemic perspective which makes normal accident theory so useful in understanding the hazards and failure-modes of deployed AI systems; while each new technology should of course be assessed on its own characteristics, there are also valuable lessons we can derive from our past experience with strategically powerful technologies. While on an object (or scientific) level, AI is of course a profoundly different technology from, say, nuclear weapons, on an operational level, they share key features which make these assemblages (sensors, algorithms, human operators) prone to normal accidents.

(1) AI systems are complex and opaque

Normal accident theory focuses on system applications that are complex and tightly coupled. Complex systems are those which have many interlocking parts, 'black-boxed' processing units, or *hidden interactions*—constellations of feedback loops which cannot be observed or fully understood directly or in real-time, but only imperfectly inferred, based on aggregate behavior. This makes the system more complex than can be properly understood by the human operator (Perrow 1984, 9). Moreover, there are many *common mode connections*, where it is not (immediately) clear which components have failed when there is an overall failure.

Tightly coupled means that "there is no slack or buffer or give between two items. What happens in one directly affects what happens in the other." (Perrow 1984, 89–90). For instance, in the context of the US nuclear force, scholars have recorded how the tightly interlinked web of early launch warning satellites, Chrome Dome aircraft, and command-and-control nodes, created a large scope for small technical failures or operator errors to rapidly cascade throughout the system, creating major false alarms (Sagan 1993; Schlosser 2014).

Critically, emerging AI systems meet all of the relevant criteria to exhibit 'normal accidents': many AI algorithms are 'black boxes', complex in design or operation. Like all software programs, they almost by default contain bugs—past studies have estimated that the software industry sees an average error rate of 15-50 errors per 1,000 lines of code (McConnell 2004; Scharre 2016a, 13). The problem of complexity is acute for many approaches in machine learning, in particular. These networks are intrinsically opaque—it is impossible to produce a formal proof of their behavior; they can be stochastic; and it is difficult to adequately anticipate all real-world scenarios, or to test a system's reactions to them, within a simulated sandbox (Borrie 2014, 8–9). Critically, while work on the 'interpretability' of machine learning decisions is advancing—compare DARPA's work on 'explainable' AI (DARPA, n.d.; Doran, Schulz, and Besold 2017)—this field is still in its infancy. AI systems as a result also suffer from common-mode failures, as it can be unclear where the error has originated—whether in sensor fault; underlying bias in the training datasets, adversarial input or infection by a virus, or some other component.

The opacity of leading AI systems also leads to an additional problem, in that it inhibits operator's critical ability to learn from 'near-accidents' or 'close calls'. After all, for certain AI applications, it may be unclear what would be the signature of such a 'near-accident'. We may not get many warning shots, and past performance of a system may not adequately prepare us for the scope of eventual failure.

(2) AI systems are tightly-coupled and fast

Moreover, many AI applications are tightly coupled, involving fast interactions and reaching and executing decisions at a speed that arguably exceeds the coupling of past systems (such as nuclear reactors) prone to normal accidents. The tight coupling and high operational speed are key features of AI systems which are plugged into extensive networks (such as the Internet of Things), or which operate in competitive environments such as high-frequency trading markets, in cyberspace or on the battlefield. The speed of AI operation ensures that when errors inevitably emerge, they are not just difficult to detect, but are also hard to halt in time. This suggests that having a human operator 'on-the-loop' is not always viable, if interaction speeds are high enough.

Worse, redundancies and safety measures built into an AI system can actually cause accidents. This is because features such as self-performance-monitoring sensors or -software, automated fail-safes, or behavioural tripwires (Bostrom 2014, 137) may increase the overall complexity of a system. They add more 'interacting parts' which themselves can quietly fail or react in unanticipated ways. In this way, technological safety measures may hinder problem isolation. At the same time, research has shown that automated safety systems can instill a blanket trust ('automation bias') in operators or users—a trust that may, perversely, encourage greater risk-taking by those users as a result of risk homeostasis (Cummings 2004; Borrie 2014, 12). Moreover, as demonstrated by the 'Patriot fratricides', automation bias may cause human operators who are nominally in-the-loop to nonetheless trust the system without question, authorizing even incorrect action requests by force of habit (Hawley 2011; Scharre 2016a, 31).

(3) AI trainers and operators have multiple objectives beyond safety

The technical propensity of AI systems to normal accidents (their complexity and tight coupling), will likely be exacerbated by the incentives of the principals that run them. This is because the designers, trainers, and operators of AI systems may in practice encounter multiple, conflicting organizational objectives beyond pure 'safety'.

At the level of AI developers and trainers, there may be restrictions on error feedback and on learning from incidents. This is the case both within companies (for instance, when working towards a tight software deployment deadline, and reporting apparently insignificant 'anomalies' would be an inconvenience), as well as between them (for instance, when it is feared that sharing details on security incidents may give away critical information about an AI's architecture or the initial settings of its algorithms).

On an operator level, there is a tension between automating functions to allow for rapid functioning, and decoupling & decentralizing them, to enable flexible error response. More obviously, for operators, safety often must be traded off against performance (Sagan 1993, 13; P. Lewis et al. 2014): some AIs, for instance those used to automate cybersecurity penetration testing, may be designed to come up with 'unintuitive' solutions and to test these rapidly. Of course, we necessarily and understandably accept some risk when using many technologies—possibly only an inert system would be perfectly (and knowably) safe. Yet the impact of these errors can be particularly high once AI systems are integrated in major infrastructures.

(4) Competitive pressures exacerbate operational risk of AI

Because of the above, many AI systems are likely to demonstrate at least some propensity towards normal accidents. Yet there are some exceptions.

In the first place, in contexts where the decisions made by the AI are less tightly coupled to (irreversible) physical outcomes, the system may be less susceptible to normal accidents, and risks may be modest or manageable. For instance, while algorithms trained on biased datasets are a real and pressing problem when applied to, for instance, sentencing or incarceration decisions (Kirchner et al. 2016; Corbett-Davies et al. 2016; Bolukbasi et al. 2016), the speed at which these decisions are implemented is bottlenecked—for now—by the human organizations that act upon them. As some of these decisions are not time-critical, in principle, this could give operators some leeway to spot errors or test for (statistical) discrepancies in the system's output. In such contexts, having a human-in-the-loop is not a (performance-inhibiting) safety measure, but instead a basic feature of the assemblage, and can function as a relatively effective fail-safe containing the error cascade.

In the second place, for many AI applications in society—such as in healthcare, transport or critical infrastructure—all actors involved share an interest in safe and reliable operation, which could still serve to promote the sharing of lessons and best practices even against the gradient of inter-actor rivalry.

However, neither of these caveats (uncoupling; unanimous interest in safety) applies in a competitive context, such as on physical or especially virtual battlefields. In a competitive context, a number of factors begin to exacerbate the operational risks of deployed AI systems. These factors include: (1) decisions must be made based on incomplete, 'messy', non-structured and potentially unreliable data; (2) the speed of reaction or interaction is accelerated even further; (3) there are risks of systems being hacked (Scharre 2016a, 34)—either directly, or indirectly, through spoofing or behavioural exploitation. For instance, deep neural networks have proven vulnerable to confrontation with adversarial examples (Nguyen, Yosinski, and Clune 2015), which can be generated without any privileged access to the algorithm's training data or logic; these spoofing attacks can moreover be hidden, so that they are invisible to humans. At their extreme, unexpected interactions between competing systems, especially in cyberspace, could cause unexpected escalation—a 'flash war' (Scharre 2016b), analogous to the algorithmic flash crashes observed in the financial sector.

Because of the advantage afforded by operational speed, many of the 'normal accident' problems may therefore be particularly acute in military AI applications (Borrie 2014; Scharre 2016a)—a frightening prospect, given that major powers, including the United States,(Center for Strategic and International Studies n.d.; Freedberg 2015) China (China's State Council 2017; Kania 2017b, 2017a), and Russia (Putin 2017) increasingly perceive AI as a cornerstone of their next-generation military and strategic supremacy. In the context of competitive pressures, there may occur technology race dynamics which produce a strong pressure to cut down on safety and err on the side of rapid deployment of relatively untested systems (Armstrong, Bostrom, and Shulman 2013), even when principals are nominally committed to safe and responsible development.

Implications for AI governance

In sum, it appears plausible that many AI applications may be even more susceptible to normal accidents than past 'textbook' case technologies such as nuclear power or aviation.

Moreover, normal accident theory suggests—and past and present experience in the field of cybersecurity has repeatedly borne out (Yampolskiy and Spellchecker 2016)—that such risks cannot simply be 'designed out' of the technology (at least not without giving up on many of their benefits). These operational insights into the 'normal' failure modes of AI systems matter on two levels.

In the first place, on a systemic level, they provide an overall context for understanding if, how, or when to 'trust' AI systems—and conversely, when operator 'trust' in an AI begins to turn into an accident-enabling 'automation bias'. In this way, it shows the point at which even the ideal of maintaining a human-in-the-loop does not offer the level of safety or reliability we may ascribe to it—indeed, how the illusion of reliability which it creates may in some ways make us less safe or robust once disaster does strike.

Secondly, on a governance level, the likely susceptibility of AI systems to normal accidents suggests that even if regulatory regimes can converge on concrete norms and standards to ensure that AI systems are deployed in a lawful and ethical manner, 'unforeseeable' yet inevitable accidents will emerge in their performance, putting both users and the public at risk. This throws up interesting problems for frameworks ranging from liability to disaster insurance, to name but a few.

There are, as of yet, no clear-cut answers to the question of how to 'prevent' normal accidents. The following is therefore largely speculative. Nonetheless, one upside of AI is that it is, in its applications and range of configurations, arguably more 'flexible' or customizable than, say, a nuclear power plant. As such, we may still try to derive relevant lessons for responsible AI regulation.

In the first place, legal and regulatory strategies should not solely trust in maintaining a human-in-the-loop, which (absent broader measures) at best simply sets up an operator to take the fall—to serve as 'moral crumple zone' (Elish 2016)—once accidents happen, and at worst actually creates new avenues for those errors to be introduced. Instead, regulators (and innovators) might seek to explore ways to change *how the system* is used by users—to find, wherever possible, ways to reduce either the coupling, or the opacity of the AI; to limit the system's autonomy (or speed) when deployed outside of intended environments. Promising legal approaches might also, for instance, seek to ensure and align the objectives of organizations towards safety (emphasizing the exchange of best practices and the sharing of incident reports); promote heterogeneity in deployed AI architectures and networked systems (to insulate systems from flash crashes), or to spur research into 'explainable' AI and AI safety more broadly.

Many of such interventions might be underwritten by an appeal to the precautionary principle—an emerging principle within international law (Andorno 2004)—which holds a particular relevance to the uncertain-yet-likely risks from 'AI normal accidents'.

Conclusion

This argument has sought to explore responsible and robust governance approaches for the deployment of AI. Like all tightly coupled, opaque systems, AIs will be prone to 'normal accidents', ensuring that perfect safety may not be attainable. Nonetheless, we may anticipate the particular 'field conditions' under which AIs are more or less susceptible to such errors, and try to account for these 'risk factors'. On this basis, this paper has briefly suggested tentative principles and practical recommendations for precautionary regulation.

Further research (as per usual) is needed. Such research might seek to map out in which sectors, or for which applications, AI systems are more or less likely to exhibit features priming them for normal accidents—and in which of these cases it is possible to mitigate these risk factors (by reducing opacity; by increasing 'slack') without incurring decay in overall performance. In addition, research could explore assemblages of organizational and technological innovation which might better promote graceful failures, or even function as hypothetical 'super-fail-safe' solutions (perhaps allowing the failure of system or operator in isolation, but reliably preventing the simultaneous failure of both). Finally, research might examine whether new generations of AI systems might perhaps after all be able to serve as more reliable 'monitoring and fail-safe systems', potentially by identifying 'signatures of failure' in deployed AIs. The history of normal accidents suggests that the error always gets through.

References

Abadi, Martin, and David G. Andersen. 2016. "Learning to Protect Communications with Adversarial Neural Cryptography." https://arxiv.org/pdf/1610.06918v1.pdf.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." *ArXiv:1606.06565 [Cs]*, June. http://arxiv.org/abs/1606.06565.

Andorno, Roberto. 2004. "The Precautionary Principle: A New Legal Standard for a Technological Age." *Journal of International Biotechnology Law* 1:11–19.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. 2013. "Racing to the Precipice: A Model of Artificial Intelligence Development." Technical Report 2013–1. Future of Humanity Institute.

Armstrong, Stuart, and Kaj Sotala. 2012. "How We're Predicting AI--Or Failing To." In *Beyond AI: Artificial Dreams*, edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75. Pilsen: University of West Bohemia. https://intelligence.org/files/PredictingAI.pdf.

Baum, S.D., Ben Goertzel, and Ted G. Goertzel. 2011. "How Long until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting & Social Change* 78:185–95.

Bilge, Leyla, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. 2009. "All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks." In . http://www2009.eprints.org/56/1/p551.pdf.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *ArXiv:1607.06520 [Cs, Stat]*, July. http://arxiv.org/abs/1607.06520.

Borrie, John. 2014. "Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies." 5. UNIDIR Resources. UNIDIR. http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9 (1).

———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bostrom, Nick, Allan Dafoe, and Carrick Flynn. 2017. "Policy Desiderata in the Development of Machine Superintelligence." http://www.nickbostrom.com/papers/aipolicy.pdf.

Brundage, Miles. 2016. "Artificial Intelligence and Responsible Innovation." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 543–54. Synthese Library 376. Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_32.

———. 2017. "My AI Forecasts--Past, Present, and Future (Main Post)." 2017. http://www.milesbrundage.com/1/post/2017/01/my-ai-forecasts-past-present-and-future-main-post.html.

Bryant, Ross. 2016. "Google's AI Becomes First Non-Human to Qualify as a Driver." Dezeen. February 12, 2016. https://www.dezeen.com/2016/02/12/google-self-driving-car-artficial-intelligence-system-recognised-as-driver-usa/.

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton & Company.

Calo, Ryan. 2010. "Peeping HALs: Making Sense of Artificial Intelligence and Privacy." *EJLS - European Journal of Legal Studies*, The Future of... Law & Technology in the Information Society, 2 (3). http://www.ejls.eu/6/83UK.htm.

Castelvecchi, Davide. n.d. "Deep Learning Boosts Google Translate Tool." *Nature News*. Accessed February 25, 2017. https://doi.org/10.1038/nature.2016.20696.

Center for Strategic and International Studies. n.d. "Assessing the Third Offset Strategy." Accessed March 5, 2017. https://www.csis.org/events/assessing-third-offset-strategy.

China's State Council. 2017. "A Next Generation Artificial Intelligence Development Plan." Translated by Rogier Creemers, Graham Webster, Paul Triolo, and Elsa Kania. New America Cybersecurity Initiative. https://na-production.s3.amazonaws.com/documents/translation-fulltext-8.1.17.pdf.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. 2016. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." *Washington Post*, October 17, 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

Courtland, Rachel. 2016. "DARPA's Self-Driving Submarine Hunter Steers Like a Human." IEEE Spectrum: Technology, Engineering, and Science News. April 7, 2016. http://spectrum.ieee.org/automaton/robotics/military-robots/darpa-actuv-self-driving-submarine-hunter-steers-like-a-human.

Crawford, Kate, and Ryan Calo. 2016. "There Is a Blind Spot in AI Research." *Nature News* 538 (7625):311. https://doi.org/10.1038/538311a.

Cummings, Mary L. 2004. "Automation Bias in Intelligent Time Critical Decision Support Systems." In *AIAA 1st Intelligent Systems Technical Conference*, 2:557–562. AIAA.

DARPA. 2016. "'Mayhem' Declared Preliminary Winner of Historic Cyber Grand Challenge." 2016. http://www.darpa.mil/news-events/2016-08-04.

———. n.d. "GRANT OPPORTUNITY - DARPA-BAA-16-53 - Explainable Artificial Intelligence (XAI)." Department of Defense - DARPA - Information Innovation Office. http://www.grants.gov/web/grants/view-opportunity.html?oppId=287284.

Dietterich, Thomas G., and Eric J. Horvitz. 2015. "Rise of Concerns about AI: Reflections and Directions." *Communications of the ACM* 58 (10):38–40. https://doi.org/10.1145/2770869.

Doran, Derek, Sarah Schulz, and Tarek R. Besold. 2017. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives." *ArXiv:1710.00794 [Cs]*, October. http://arxiv.org/abs/1710.00794.

Elish, M. C. 2016. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (We Robot 2016)." SSRN Scholarly Paper ID 2757236. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2757236.

Freedberg, Sydney J. 2015. "Centaur Army: Bob Work, Robotics, & The Third Offset Strategy." *Breaking Defense*, November 9, 2015. http://breakingdefense.com/2015/11/centaur-army-bob-work-robotics-the-third-offset-strategy/.

Furness, Dyllan. 2016. "AI in Agriculture? Algorithms Help Farmers Spot Crop Disease like Experts." Digital Trends. October 6, 2016. http://www.digitaltrends.com/computing/ai-crop-disease/.

Hambling, D. 2016. "The Inescapable Net: Unmanned Systems in Anti-Submarine Warfare." BASIC Parliamentary Briefings on Trident Renewal. British-American Security Information Council. http://www.basicint.org/sites/default/files/BASIC_Hambling_ASW_Feb2016_final_0.pdf.

Hanson, Robin. 2016. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford, New York: Oxford University Press.

Hawley, John K. 2011. "Not by Widgets Alone: The Human Challenge of Technology-Intensive Military Systems." *Armed Forces Journal*. http://armedforcesjournal.com/not-by-widgets-alone/.

Horowitz, Michael C. 2016. "Who'll Want Artificially Intelligent Weapons? ISIS, Democracies, or Autocracies?" *Bulletin of the Atomic Scientists*, July 29, 2016. http://thebulletin.org/who%E2%80%99ll-want-artificially-intelligent-weapons-isis-democracies-or-autocracies9692.

Horowitz, Michael C., Sarah E. Kreps, and Matthew Fuhrmann. 2016. "Separating Fact from Fiction in the Debate over Drone Proliferation." *International Security* 41 (2):7–42. https://doi.org/10.1162/ISEC_a_00257.

House of Commons Science and Technology Committee. 2016. "Robotics and Artificial Intelligence: Fifth Report of Session

2016–17." https://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf.

Kania, Elsa. 2017a. "数字化 – 网络化 – 智能化: China's Quest for an AI Revolution in Warfare." *The Strategy Bridge* (blog). June 8, 2017. https://thestrategybridge.org/the-bridge/2017/6/8/-chinas-quest-for-an-ai-revolution-in-warfare.

———. 2017b. "Great Power Competition and the AI Revolution: A Range of Risks to Military and Strategic Stability." Lawfare. September 19, 2017. https://www.lawfareblog.com/great-power-competition-and-ai-revolution-range-risks-military-and-strategic-stability.

Kirchner, Lauren, Julia Angwin, Jeff Larson, and Surya Mattu. 2016. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Kleeman, Alexandra. 2016. "Cooking with Chef Watson, I.B.M.'s Artificial-Intelligence App." The New Yorker. 2016. http://www.newyorker.com/magazine/2016/11/28/cooking-with-chef-watson-ibms-artificial-intelligence-app.

Lewis, Dustin A., Gabriella Blum, and Naz K. Modirzadeh. 2016. "War-Algorithm Accountability." SSRN Scholarly Paper ID 2832734. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2832734.

Lewis, Patricia, Heather Williams, Benoît Pelopidas, and Sasan Aghlani. 2014. "Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy." Chatham House.

Lewis-kraus, Gideon. 2016. "The Great A.I. Awakening." *The New York Times*, December 14, 2016. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

Lieber, Keir A., and Daryl G. Press. 2017. "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence." *International Security* 41 (4):9–49. https://doi.org/10.1162/ISEC_a_00273.

Lin, Patrick, Keith Abney, and George A. Bekey. 2011. *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press.

Linn, Allison. 2015. "Microsoft Researchers Win ImageNet Computer Vision Challenge." *Next at Microsoft* (blog). December 10, 2015. https://blogs.microsoft.com/next/2015/12/10/microsoft-researchers-win-imagenet-computer-vision-challenge/.

Littlefield, Scott. n.d. "Anti-Submarine Warfare (ASW) Continuous Trail Unmanned Vessel (ACTUV)." DARPA. Accessed September 19, 2016. http://www.darpa.mil/program/anti-submarine-warfare-continuous-trail-unmanned-vessel.

Maas, Matthijs, Tim Sweijs, and Stephan De Spiegeleire. 2017. *Artificial Intelligence and the Future of Defense: Strategic Implications for Small- and Medium-Sized Force Providers*. The Hague, The Netherlands: The Hague Centre for Strategic Studies. http://hcss.nl/report/artificial-intelligence-and-future-defense.

Marshall, Alex. 2017. "From Jingles to Pop Hits, A.I. Is Music to Some Ears." *The New York Times*, January 22, 2017. https://www.nytimes.com/2017/01/22/arts/music/jukedeck-artificial-intelligence-songwriting.html.

Mascarenhas, Hyacinth. 2016. "Associated Press to Expand Its Sports Coverage by Using AI to Write Minor League Baseball Articles." International Business Times UK. July 5, 2016. http://www.ibtimes.co.uk/associated-press-expand-its-sports-coverage-by-using-ai-write-minor-league-baseball-articles-1568804.

McConnell, Steve. 2004. *Code Complete: A Practical Handbook of Software Construction, Second Edition*. 2nd edition. Redmond, Wash: Microsoft Press.

Metz, Cade. 2017. "The AI Threat Isn't Skynet. It's the End of the Middle Class." *WIRED*, 2017. https://www.wired.com/2017/02/ai-threat-isnt-skynet-end-middle-class/.

Müller, Vincent C., and Nick Bostrom. 2016. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, edited by Müller, Vincent C. Berlin: Synthese Library. http://www.nickbostrom.com/papers/survey.pdf.

Musthaler, Linda. 2016. "How to Use Deep Learning AI to Detect and Prevent Malware and APTs in Real-Time." Network World. March 11, 2016. http://www.networkworld.com/article/3043202/security/how-to-use-deep-learning-ai-to-detect-and-prevent-malware-and-apts-in-real-time.html.

Nehal, Bhuta, Susanne Beck, Robin Geiss, Hin-Yan Liu, and Klaus Kress, eds. 2016. *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press.

Newman, LilHay. 2016. "AI Can Recognize Your Face Even If You're Pixelated." WIRED. 2016. https://www.wired.com/2016/09/machine-learning-can-identify-pixelated-faces-researchers-show/.

Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2015. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427–436.

OSTP. 2016. "Preparing for the Future of Artificial Intelligence." Office of Science and Technology Policy, National Science and Technology Council Committee on Technology. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

Perrow, Charles. 1984. "Normal Accidents: Living with High Risk Technologies." 1984. http://press.princeton.edu/titles/6596.html.

Plebe, Alessio, and Pietro Perconti. 2012. "The Slowdown Hypothesis." In *Singularity Hypotheses*, edited by Amnon H. Eden, James H. Moor, Johnny H. Søraker, and Eric Steinhart, 349–65. The Frontiers Collection. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32560-1_17.

Putin, Vladimir. 2017. "Открытый Урок «Россия, Устремлённая в Будущее» | Open Lesson 'Russia Looking to the Future.'" Президент России | President of the Russian Federation. September 1, 2017. http://kremlin.ru/events/president/news/55493.

Rickli, Jean-Marc. 2017. "Artificial Intelligence and the Future of Warfare." In *WEF Global Risks Report 2017*, 49. http://www3.weforum.org/docs/GRR17_Report_web.pdf.

Roff, Heather M. 2014. "The Strategic Robot Problem: Lethal Autonomous Weapons in War." *Journal of Military Ethics* 13 (3):211–27. https://doi.org/10.1080/15027570.2014.975010.

Sagan, Scott D. 1993. *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons.* Princeton: Princeton University Press.

Scharre, Paul. 2016a. "Autonomous Weapons and Operational Risk." Ethical Autonomy Project. 20YY Future of Warfare Initiative. Center for a New American Security. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf.

———. 2016b. "Flash War - Autonomous Weapons and Strategic Stability." presented at the Understanding Different Types of Risk, Geneva, April 11. http://www.unidir.ch/files/conferences/pdfs/-en-1-1113.pdf.

Schlosser, Eric. 2014. *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. Reprint edition. New York: Penguin Books.

Scholl, Zackary. 2015. "Turing Test: Passed, Using Computer-Generated Poetry." *Raspberry PI AI* (blog). January 24, 2015. https://rpiai.wordpress.com/2015/01/24/turing-test-passed-using-computer-generated-poetry/.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587):484–89. https://doi.org/10.1038/nature16961.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. 2017. "Mastering the Game of Go without Human Knowledge." *Nature* 550 (7676):nature24270. https://doi.org/10.1038/nature24270.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9):1568–80. https://doi.org/10.1016/j.respol.2013.05.008.

The Economist. 2016. "Anti-Submarine Warfare: Seek, but Shall Ye Find?" *The Economist*, August 6, 2016. http://www.economist.com/news/science-and-technology/21703360-proliferation-quieter-submarines-pushing-navies-concoct-better-ways.

US Senate Subcommittee on Space, Science and Competitiveness. 2016. *The Dawn of Artificial Intelligence*. http://www.commerce.senate.gov/public/index.cfm/2016/11/commerce-announces-first-artificial-intelligence-hearing.

Weng, Yueh-Hsuan. 2010. "Beyond Robot Ethics: On a Legislative Consortium for Social Robotics." *Advanced Robotics* 24 (13):1919–26. https://doi.org/10.1163/016918610X527220.

wifiphisher. (2014) 2017. *Wifiphisher: Automated Victim-Customized Phishing Attacks against Wi-Fi Clients*. Python. Wifiphisher. https://github.com/wifiphisher/wifiphisher.

Woolley, Samuel C., and Philip N. Howard. 2016. "Automation, Algorithms, and Politics| Political Communication, Computational Propaganda, and Autonomous Agents — Introduction." *International Journal of Communication* 10 (0):9.

Xiong, Wayne, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. "Achieving Human Parity in Conversational Speech Recognition." *ArXiv Preprint ArXiv:1610.05256*. https://arxiv.org/abs/1610.05256.

Yampolskiy, Roman V., and M. S. Spellchecker. 2016. "Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures." *ArXiv:1610.07997 [Cs]*, October. http://arxiv.org/abs/1610.07997.