Establishing the Conditions of Engagement with Machines

Dan Geer
Glenn Gaffney

Realism is dealing with the machines, people, organizations, and governance systems we have, not those we wish for.

We begin our discussion of "autonomy" with its Western meaning for the human individual: "to be autonomous is to govern oneself, to be directed by considerations, desires, conditions, and characteristics that are not simply imposed externally upon one." Autonomy is "the capacity to impose upon ourselves, by virtue of our practical identities, obligations to act."[1] Similarly, extending autonomy to machines is a partial release from external control that comes with obligations to act. That's the easy part.

Until the last decade, machines with no human in the loop had very limited repertoires of actions they could take, turning on the pump when they detected the water was rising. From that set of inherent constraints came reliability and understandability. As is obvious, we are transiting an inflection point where machines are gaining trained reasoning capacity that can allow problem-solving without a human in the loop. Even the training can be self-administered: the autonomy of self-modification (and, with it, emergent behavior – a topic to which we return below).

This leads to the set of interactions touched upon in this essay: Western principles of control, various tradeoffs, drivers of adoption, responsibilities, predictability, and recovery from faults – a list that is neither ordered nor exhaustive of the work remaining to be done. We are well past arguing over whether autonomy is coming, or that it is a national security issue.[2] However, while the transiting of the inflection point is clear and many of

Dan Geer, Senior Fellow, In-Q-Tel. Milestones: The X Window System and Kerberos (1988), the first information security consulting firm on Wall Street (1992), convenor of the first academic conference on mobile computing (1993), convenor of the first academic conference on electronic commerce (1995), the "Risk Management is Where the Money Is" speech that changed the focus of security (1998), the Presidency of USENIX Association (2000), the first call for the eclipse of authentication by accountability (2002), principal author of "Cyberinsecurity: The Cost of Monopoly" (2003), co-founder of SecurityMetrics.Org (2004), convener of MetriCon (2006-present), author of "Economics & Strategies of Data Security" (2008), and author of "Cybersecurity & National Policy" (2010). Creator of the Index of Cyber Security (2011) and the Cyber Security Decision Market (2012). Lifetime Achievement Award, USENIX Association, (2011). Expert for NSA Science of Security award (2013-present). Cybersecurity Hall of Fame (2016) and ISSA Hall of Fame (2019). Testified five times before Congress.

the national security concerns recognized, less obvious is whether we as a nation have a preferred destination point/performance design in mind. We appear instead to have grown comfortable letting market forces drive development and consumer outcomes, and only then to intervene around any undesirable outcomes and effects as they arise. This approach is unsound given China's fierce competition in pursuit of the foundational technology of the next-generation economic infrastructure and whose principles will dominate next generation infrastructure, and otherwise be embedded in technologies that will mediate our day-to-day life.

SPEED AND COMPLEXITY DRIVE THE DISTRIBUTION OF ROLES AND CONTROLS

This may be easy to say and accept on first reading but analyzing the implications is more complicated. We (humans) have long since proven that we can build systems that we cannot then understand enough to control. This should not surprise; complexity ensures emergent behavior. It has been 25 years since Dyson wrote, "Emergent behavior is that which cannot be predicted through analysis at any level simpler than that of the system as a whole. Emergent behavior, by definition, is what's left after everything else has been explained."[3] Therefore, when we cannot explain the cause-and-effect relationship of some autonomous system's choices that crashed some platform, we revert to comparing the overall safety of the system as a whole and "accept" the attendant risks. The contribution of algorithmic trading to flash crashes at the NYSE might be a recognizable example.[4] In striving for machines to learn not only during preparation for going live but also to learn as a result of having gone live, we are actively seeking emergent behavior yet not preparing for the potential consequences of that emergent behavior.

Consequently, we ask: should hands-off mathematical operations — autonomous algorithms — be treated as if they are correct-by-definition or incorrect-by-definition?

Glenn Gaffney is the Chief Strategy Officer for the NobleReach Foundation. The NobleReach Foundation is a nonprofit organization on a mission to inspire tech talent to tackle our nation's most pressing challenges. Before joining NobleReach, Mr. Gaffney served as a senior fellow at IQT, supporting the identification of and strategic investment in ready soon technology that can uniquely meet economic and national security needs. Before joining IQT in 2017, he enjoyed a 30+ year career in science, technology, analysis, and operations within the U.S. Intelligence Community. Mr. Gaffney's government service included senior positions as the Director of Science and Technology for the Central Intelligence Agency, the Deputy Director of National Intelligence for Collection, and the Associate Director of CIA for Talent.

Do we default to trust or mistrust? Given the myriad parameters, ML algorithms likely will never be 100% susceptible to coherent explanation[5] (see Fig. 1); hence the venerable strategy of trust-but-verify seems permanently unattainable. That leaves only slow, measured delegation of authority to the AI under the banner of hope or fighting fire with fire under the banner of the precautionary principle,[6] i.e., using one AI to watch another[7] in an attempt to constrain the emergent behavior of the base AI system, and the hope that collusion does not follow.[8] Exhaustively testing an autonomous system is impossible; reserving part of the training data for post-training validation runs is about all there is. Why? Because the possible outcome space is too large to explore — the only place to test is in production, which brings us back to the question of whether we treat the AI by default as correct or incorrect. Yes, the reality is more complex than that. In much of cybersecurity design, the emphasis is now on "zero trust;" every interaction between components must be challenged to prove it is wanted. What that means for autonomy is unclear; some argue that imposing upon ourselves an obligation to act includes the Golden Rule, which should also apply to autonomous systems.[9] Others say that trust is "confident anticipation backed by effective recourse,"[10] the antithesis of saying that every thinking entity is my friend.[11] What is the recourse when an autonomous system produces an unwanted result? Does that not imply a base requirement that all actions by autonomous systems must be inherently attributable in terms of cause and effect? Our desire for attribution is such that, even though the size of emergent behavior space is too large to predict, we will still define any system by the extrema that emerge from it and look to hold someone or something accountable for the "failure" — System X crashed the plane, or shut down the power grid, or launched a cyber-attack against an ally because some third party was using the ally's infrastructure to attack the US.

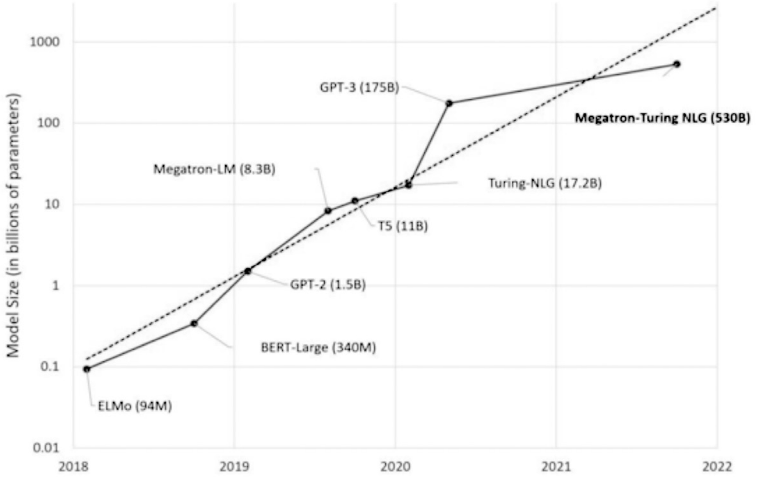

Figure 1: Growth of parameter count in Natural Language Generation (NLG) models

Speed enters all of this. First is the increasing speed of machine operations themselves. Second is the speed at which key decisions are delegated to specific machine AIs, and driven by power competition, both commercial and military. Third is the speed of proliferation of machine AI into new areas of application, and this is where the natural pace of rulemaking is the most insufficient.

Extrema are not boundable

A massive potential outcome space, which describes all non-toy machine learning (ML) systems, renders it impossible to bound what an extreme value is or might be. The more complexity in any system, the broader the range of that system's possible outcomes. This makes the probability distribution of possible outcomes have "fat tails," meaning that across the distribution of possible events, the statistical properties of the entire distribution as a whole come to be dominated by the impact of rare events. As an example, when there are one million buckets with no money in them and one bucket with $100,000,000, the expected gain you get from sticking your hand in a random bucket is $100 – an amount which you cannot actually win and, in any case, is entirely a function of that one solitary outlier value. More critically to our discussion here, for fat tails, the difference between extrema already observed and future extrema are much larger than for distributions with thin tails; the worst flood you have ever seen does not tell what is the worst flood that can happen.

Avoiding the uncharted seamount

Given the above, the mythical prudent man or government would plan for maximum damage scenarios, not for maximum likelihood scenarios. Does this mean that the introduction of autonomy must be incremental? No, insofar as the only way to test is in production and past results are no guarantee of future performance. Does this mean watchdog processes are needed (even if they, too, are massive, non-interrogable ML)? Yes, with design-space wiggle room here for whether the watchdog's autonomy includes overriding authority for the AI under watch. For the sake of resilience if no other, does this mean retaining whatever mechanism pre-dated the autonomous system now being put into operation? A strong yes if the prior mechanism can be guaranteed to remain in working order while it stands by, else no, expecting some prior mechanism to come to the rescue from unattended cold standby is a false security promise.

Mitigating downsides

As with any other substantive risk management, the key challenge is mitigating downsides. Speed, cost, efficacy, latency, side effects, and more — all figure in. Speaking broadly, planning for the loss of a meaningful asset implies planning for the availability of compensating reserve, the role insurance plays in normal affairs up to and including those where the government is the insurer of last resort. Retaining working alternatives is, in this sense, a kind of insurance. So is deployed diversity — Nature knows better than to fabricate monocultures that fail in lockstep while market forces arguably do not. While it has been repeatedly but ineffectually discussed in the setting of one technical aspect of societal digitization, namely cybersecurity, deploying thoroughgoing autonomy into societally critical roles might suggest the issuance of catastrophe bonds[12] to cover the tail risk of dependence on that autonomy. Lessons learned from managing the tail risk of the nuclear power industry under the Price-Anderson Act[13] should also inform an adroit tail risk strategy for critical autonomous processes.

Accommodation to democracy

While all politics may be local, all technology is global; hence technology policy instantiation inside an autonomous system must somehow accommodate local values. Autonomous technology suppliers sometimes assume a quasi-governmental role. Government vs. autonomy debates previously were often confined to centralized vs. decentralized administrative organization, antitrust regulation vs. "natural monopolies," or the reach of public health measures. No more. The largest tech firms now dwarf small countries[14] in economic size, number of clients,[15] and the amount of personal information stored about those clients. Rule of law observes jurisdictional boundaries, but that limitation more often than not fails to cover technology. As companies and governments deploy more autonomy, the capability set of autonomous systems must include awareness of jurisdictions in the equation. Achieving this in practical terms prompts us to quote Lessig:[16]

> Every age has its potential regulator, its threat to liberty. Our founders feared a newly empowered federal government; the Constitution is written against that fear. John Stuart Mill worried about the regulation by social norms in nineteenth-century England; his book On Liberty is written against that regulation. Many of the progressives in the twentieth century worried about the injustices of the market. The reforms of the market, and the safety nets that surround it, were erected in response.

> Ours is the age of cyberspace. It, too, has a regulator. This regulator, too, threatens liberty. But so obsessed are we with the idea that liberty means 'freedom from government' that we don't even see the regulation in this new space. We therefore don't see the threat to liberty that this regulation presents.

The regulator is code — the software and hardware that make cyberspace what it is. This code, or architecture, sets the terms on which life in cyberspace is experienced, and governs both privacy protections and censored speech. It determines whether information access is general and/or zoned. It affects who sees what or what is monitored. In a host of ways that one cannot begin to see unless one begins to understand the nature of this code, the code of cyberspace regulates.

Subjugating autonomy to democracy is no small challenge. Democracies are inefficient by design, and we need machines to do what they do best. How best can democracy thrive in an automated world? By what principles will we govern "by the people" in tandem with free-running, self-modifying algorithms? We desire clarity in understanding why control decisions are made – particularly when we do not like the outcomes of those decisions. Unless we know we have some visibility or understanding, if not transparency,[17] providing checks and balances[18] sufficient to the task is not possible. US policymakers did not foresee that surveillance would become commercially monetized[19] or that low-end job descriptions might inherently include functioning as an informant.[20] Similarly, no one should expect autonomy to play nice magically. We need to establish a way to safely exercise and test emergent behavior with some degree of public engagement and transparency. Aspects of applied techno-sociological research across critical public service systems and infrastructures must have the principal goal of establishing the design space for watchdog AI systems. All of that is before we use the word "China."

*Data as a drive*r *for autonomy*

The explosive growth in data volume has led some to suggest that DNA storage alone can accommodate the volume.[21] Even so, much data will remain at its point of collection; there is not enough bandwidth to move it all elsewhere. Lt. Col. Rhett Hierlmeier, who headed up the training center for the F-35, in an interview observed: "Standing outside the cockpit, he peers into the darkened dome and says he believes we will one day fight our enemies from inside one of these things. When I ask what that will take, he says flatly, 'Bandwidth,' which is why 'engineers are focused on things like improving artificial intelligence so planes can act with more autonomy, thus cutting down on communication bandwidth [requirements].'"[22] In this and other examples, we see that data richness is the foremost driver for algorithm autonomy.

If data volume forces distal compute nodes to require autonomy, what does that imply for cybersecurity? Authorities some years back concluded that "[t]he best approach to cybersecurity will emphasize defenses that are robust to unforeseen perturbations, evolvable in response to changing conditions, and self-repairing in the face of damage."[23] This was an early call for a second watchdog breed-type of autonomy. (Everyone, including those concerned about data interception in transit, will agree that data untransmitted is data unintercepted, which provides yet another driver toward autonomy.)

Operational reality and government responsibility

Unless an algorithm is misapplied, autonomous AI systems will usually perform better than a human at the same task. At the same time, we know that optimality and efficiency work counter to robustness and resilience.[24] We know that complexity tends to conceal interdependence, and unacknowledged interdependence is the source of black swan events. We know that the benefits of digitalization are not transitive (they do not spread to all concerned) but the risks are (and do). We know that because single points of failure require militarization wherever they underlie gross societal dependencies, frank minimization of the number of such single points of failure is a national security obligation. We know that cascade failure ignited by random faults is quenched by redundancy whereas cascade failure ignited by sentient opponents is exacerbated by redundancy. Hence, we know that preservation of uncorrelated operational mechanisms is likewise a national security obligation.[25] Once again, leaving everything up to globalized market forces will almost certainly result in serious downside outcomes for many without clarifying what constitutes acceptable costs.

*An early adopter: Autonom*y *for cybersecurity*

The need for speed in each step of the cybersecurity OODA[26] loop is growing more urgent, and that which we must protect is growing more valuable and more complex. Whether or not caused by a litany of accumulated design and implementation failures, it remains true that humans simply cannot keep up with the growing demand. Nor are they good at accepting the consequences of weakness. Cybersecurity tools must include autonomous actors. Most of us have a natural default tendency to seek (and expect) a technical solution to self-imposed problems, but we now have little choice but to center strategy on employing algorithms to do what we cannot ourselves do, which is to protect us from other algorithms. This may be inevitable, if in cybersecurity offense actually enjoys a structural advantage over defense, it might mean that wars of attrition spring up within each new theater of offense, each new dependence made critical simply by the aggregate mass adoption of the underlying technology.

To be clear, while our systems can benefit from greater autonomy in cyber security, we simultaneously must pre-determine the reasonable limits of that autonomy. All models have a tipping point, and such tipping points (vulnerabilities) need watchdog protection from sophisticated adversaries capable of exploiting those tipping points. Adversaries greatly value the ability to undermine our trust in our own data, and to redirect our autonomous agents and thereby inflict friendly fire. A paramount research grade problem here would be a solution for carefully breeding the watchdogs.

The evolution in thinking, negotiating, and training with machines

In the age of autonomy what is worse: getting the right answer for the wrong reasons, or getting the wrong answer for the right reasons? Are we troubled by the implicit de-skilling that comes with substituting autonomous algorithms for practiced, intuitive human judgment, however inferior the latter is? Langewiesche's analysis of the June 2009 crash of Air France Flight 447[27] comes to this conclusion: "We are locked into a spiral in which poor human performance begets automation, which worsens human performance, which begets increasing automation," and further, that "the effect of automation is to reduce the cockpit workload when the workload is low and to increase it when the workload is high." Put differently, as we increasingly become dependent on autonomous systems, we need to anticipate and recognize the point at which an autonomous AI becomes an irreversible necessity.

Once we cross that no-going-back point, we aren't so much flying the plane with AI as we are negotiating with an AI agent in order to fly a plane (or complete another complex task). We are already in the era of negotiation with AI rather than harnessing it as a tool we command and control, but we have yet to fully acknowledge this. In other words, it is undoubtedly essential to train humans and machines as a team. Professional certifications and regulations must ramp up to this reality. This has already begun within limited areas by firms with the resource base and drive to do so, but it is entirely locally driven. And, as is true for other high technology breakthroughs, government regulation is woefully trailing, and desperately needed.

It is possible that, during the training of the man-machine composite, the human expert can help the machine learn and become more effective in complementing human behavior. Humans and machines partnering together have already proven to be superior to machines alone in some strategy games.[28,29] In other areas, human experts have undergone retraining to learn how to better interface with the AI agent to ensure both remain on the same page for operating safely.[30] In such examples, the burden of understanding and adapting to the communication style remains with the human; the ability to reason is the distinguishing human characteristic, though for how long remains a debated question. Given the consequences of getting it wrong, the US should seriously consider the need for a "Reverse Turing Test" whereby nothing can be classified or accepted that does not either recognize that it is interfacing or working with a human and act accordingly, or at a minimum be proven to be totally subservient to the human in the loop. Indeed, this article proposes the following as a general rule that governs use of machine learning: A machine must recognize when it is interacting with a human, and we must have already chosen if and when "I'm sorry, Dave, I'm afraid I can't do that"[31] might actually be proper.

A role for government and our allies

There has been an appropriately increased focus this past year on technology innovation as part of our great power competition with China. While AI is called out as a key tech sector for competition, per se, we must recognize that AI's application within other critical tech sectors, like biotech, and in critical infrastructure systems and public services will be equally important for our overall competitiveness. The US government (USG), as part of its emphasis on innovation and economic competitiveness, must pursue an understanding of what democratic principles of digital design mean in practice. Digital proving grounds will be needed. Such proving grounds must be able to leverage national labs and other federally funded infrastructure. Integrated research teams must include public and private sector experts alike. Participation must be either compelled or heavily incentivized. There must be no confusion that this effort is a look-ahead to understand and then forestall the distribution of autonomous systems that are inadvertently anti-democratic and/or uncontrollable once deployed. Call it accountability, if you prefer, but think of it as the governance of checks and balances. Any system of trust requires a trust anchor; this effort is to construct one. So long as the autonomous decision-making is not susceptible to coherent explanation, autonomy's implicit authoritarianism[32] means operational countermeasures must be vetted at those proving grounds. Hard questions await, e.g., when may an autonomous system reproduce?

If a probative model can be established, then the USG should look to export the model to like-minded democratic allies around the world. Sharing such proving ground spaces internationally would send clear signals that alternatives to "Made in China" infrastructure and "Controlled in China" data stores are within reach. The policies around autonomy can make clear the distinctions between autocracies and democracies like few other areas of comparison. The US cannot meet the demands of the great power competition before us on our own within the short 9-10-year time frame we face. This is a time to do things *with* our allies, not *to* them.

To be adopted, allies must view the offer we make to them to be real in terms of their economic glide path. We believe the only sure way to demonstrate our commitment is to do here at home what we urge them to do ("do as we do"). If we can gain significant tech translation activity across a few critical economic areas and across several regional partners within the next 3-5 years, we believe that will prove disruptive to China and its 2030 timeline to overtake the US.

Summary of Recommendations — What the US government should establish next.

1. A national priority for dedicated, interdisciplinary, techno-sociological research on autonomous systems, including a requirement for AI-on-AI "watchdog" design and development. This research must cover autonomous system safety and fitness for use as to both industrial accidents and hostile actors. Prudence requires that all autonomous systems be considered dual use by default.

2. A national priority risk management strategy for critical autonomous processes. To motivate the best efforts of the private sector, strict liability for autonomous systems must be put in place and be explicit. Mandatory reporting thresholds for untoward and unanticipated incidents must likewise be explicit, including unarguable clarity for which agencies have the duty to receive and act upon such reports.

3. Regulations and certifications for what is acceptable practice to train humans and machines as a team. This has begun in some industries but needs to be required for critical systems drawing direction from the training of professionals who interact with complex environments, such as lawyers, licensed structural engineers, passenger aircraft pilots, certified public accountants, etc.

4. A national autonomy design criterion that, at a minimum, insists that autonomous systems recognize a human in the loop and that human's authority for interaction. Abiding by such a criterion would grant to the maker of the autonomous system those kinds of legal immunity that are proportionate to the rigor of the criteria followed. A starting point might be airworthiness certificates issued by the Federal Aviation Administration (and some other countries such as Australia).

5. Proving grounds in partnership with industry, which focus on the application of, and experience with, autonomy across tech sectors deemed critical in ongoing competition with China, critical infrastructure, and public service systems. These proving grounds can be topic-specific, and because these partnerships may include both regulator and the regulated, convenors and operators of such partnerships of such proving grounds should, as precedent has shown, be private third parties.[33] These proving grounds:

    a) Can be established across several regions of the country to engage the broadest range of Americans through inclusion, transparency, and communication in relevant work, and

    b) Should be designed and operated to provide common experience in testing and developing new risk strategies and systems using modeling and testing to explore options and develop consensus around solutions. Partnerships among the autonomy industry, government, and insurance industry should lead to new incentive models and policies for buying down the risk in key areas for rapid development, testing, and fielding.

    c) Should include experiments designed to proactively provide baselines for new regulations and certifications in team training of humans and machines.

    d) Should be enabled for next-generation data operations, establishing the necessary practices for the rapid advancement of research and tech translation into application and commercialization while providing protection from economic espionage and theft and securing the privacy of our citizens. This next-generation infrastructure will support the new economy and is as vitally important as the science, technology, and commercial enterprise it seeks to enable.

The closing question: Is autonomy a zero-sum game?

NOTES

1. "Autonomy in Moral and Political Philosophy," *Stanford Encyclopedia of Philosophy,* 2020, plato.stanford.edu/entries/autonomy-moral.
2. Defense Science Board, "Seven Defense Priorities for the New Administration," point 5, "Anticipating intelligent systems and autonomy," December 2016, dsb.cto.mil/reports/2010s/Seven_Defense_Priorities.pdf.
3. George Dyson, *Darwin Among the Machines,* Addison-Wesley, 1997.
4. Peter Bloom, "In finance the war is over, the machines won, the number of real players is under ten, and humans will never again be in the OODA loop," CNAS Catastrophic Risks Workshop, January 12, 2018.
5. As of October 11, 2021, the biggest is the Microsoft-Nvidia "Megatron-Turing Natural Language Generation" model at 530 billion parameters (www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model); the prior record holder was OpenAI's "GPT-3" at 175 billion parameters, debuting July 22, 2020, (arxiv.org/pdf/2005.14165.pdf).
6. United Nations Rio Declaration, Principle 15: "Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation." www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_CONF.151_26_Vol.I_Declaration.pdf.
7. So-called adversarial AI is to train one model and then to put that model into action with a second AI, one not trained at all on the subject matter of the first, to attempt to get desirable outcomes by feeding the first one synthesized data.
8. As it is conservative to assume that algorithms will learn to lie to us, it is unsurprising that two law professors have already suggested that price-fixing collusion amongst robot traders will be harder to detect than collusion amongst human ones – bottom line: designing a watchdog is probably harder than designing the system it exists to watch; see Ariel Ezrachi & Maurice Stucke, "When Robots Collude," May 2015, papers.ssrn.com/sol3/papers.cfm?abstract_id=2591874.
9. Sorin Adam Matei & Elisa Bertino, "Can N-Version Decision-Making Prevent the Rebirth of HAL 9000 in Military Camo? Using a 'Golden Rule' Threshold to Prevent AI Mission Individuation," *Policy-Based Autonomic Data Governance,* 2019.
10. Daniel E. Geer, "Application Security Matters," Open Web Application Security Project keynote, April 4, 2012, geer.tinho.net/geer.owasp.4iv12.txt.
11. Rudyard Kipling's poem "If" is a jewel, but this couplet captures autonomy and trust: "If neither foes nor loving friends can hurt you, If all men count with you, but none too much."
12. Federal Reserve Bank of Chicago, "Catastrophe Bonds: A Primer and Retrospective," 2018, www.chicagofed.org/publications/chicago-fed-letter/2018/405.
13. Congressional Research Service, "Price Anderson Act," 2018, crsreports.congress.gov/product/pdf/IF/IF10821.
14. As measured by total 2016 revenue, nine of the top twenty-five economic entities in the world are companies, www.weforum.org/agenda/2016/10/corporations-not-countries-dominate-the-list-of-the-world-s-biggest-economic-entities.
15. Facebook has more users than the combined population of India and China.
16. Lawrence Lessig, "Code Is Law", Harvard Magazine, January 2000, www.harvardmagazine.com/2000/01/code-is-law-html.
17. Article 15 (and others) in the EU's General Data Protection Regulation creates a right to challenge any algorithmic decision with the demand of "Why?" which cannot be answered by an uninterrogable algorithm; At what level of criticality do we require interrogability prior to legal deployment? Does this create a new variety of "informed consent" when a medical algorithm says what must be done but cannot explain why, and what does "informed" in "informed consent" then mean? www.clarip.com/data-privacy/gdpr-article-15.
18. While written with humans in mind, James Madison might well have been speaking to the issue at hand; "The accumulation of all powers … in the same hands … may justly be pronounced the very definition of tyranny," The Federalist No. 51, 1788.
19. Shoshana Zuboff, *Surveillance Capitalism,* Public Affairs Press, January 2019.
20. Elizabeth E. Joh, "A Gig Surveillance Economy," Hoover Aegis Series Paper No. 2108, November 10, 2021, s3.documentcloud.org/documents/21101558/a-gig-surveillance-economy.pdf.
21. As this working paper from the DNA Data Storage Alliance notes, global production of data storage devices is entering a period of profound and unfixable shortages; "An Introduction to DNA Data Storage," June 2021, dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf.
22. Kevin Gray, "The Last Fighter Pilot," Popular Science, December 22, 2015, www.popsci.com/last-fighter-pilot.

NOTES

23. Stephanie Forrest, Steven Hofmeyr, and Benjamin Edwards, "The Complex Science of Cyber Defense," Harvard Business Review, June 24, 2013, hbr.org/2013/06/embrace-the-complexity-of-cybe.
24. "Optimality vs. Fragility: Are Optimality and Efficiency the Enemies of Robustness and Resilience?" joint workshop by the Santa Fe Institute and Morgan Stanley, October 2, 2014; Abstract: "The goals of optimality and efficiency are pervasive in business and government, yet there is suggestive evidence that attempting to optimize a system's performance or to make it more efficient can make it more brittle, fragile, and less robust and resilient."
25. Daniel E. Geer, "A Rubicon," Hoover Aegis Series Paper No. 1801, www.hoover.org/research/rubicon.
26. The OODA loop is the repeating cycle of [observe, orient, decide, act] first developed by U.S. Air Force Col. John Boyd, military strategist.
27. William Langewiesche, "The Human Factor," *Vanity Fair*, 2014, www.vanityfair.com/news /business/2014/10/air-france-flight -447-crash.
28. "The cyborg chess players that can't be beaten," BBC Future, 2015, www.bbc.com/future/article/20151201-the-cyborg-chess-players-that-cant-be-beaten.
29. Nicky Case, "How to Become a Centaur," Journal of Design & Science, 2018, jods.mitpress.mit.edu/pub/issue3-case/release/6.
30. Gregory Travis, "How the Boeing 737 MAX Disaster Looks to a Software Developer," IEEE Spectrum, April 18, 2019, spectrum.ieee.org/how-the-boeing-737-max-disaster-looks-to-a-software-developer
31. "2001" script, 1968, where an autonomous system concludes that its own mission duty trumps human control
32. I do not say why. I only say what.
33. Daniel E. Geer & Richard Danzig, "Mutual Dependence Demands Mutual Sharing," IEEE Security & Privacy, January 2017, and in fuller treatment, Richard Danzig, *Surviving on a Diet of Poisoned Fruit*, Center for a New American Security, July 2014.