Foreign Policy at BROOKINGS

THE RISE OF THE FUTURISTS: THE PERILS OF PREDICTING WITH FUTURETHINK

ALEXANDER H. MONTGOMERY AND AMY J. NELSON

NOVEMBER 2020

EXECUTIVE SUMMARY

Policymakers, facing increasingly uncertain contemporary and future security and technology environments, are engaging in *futurethink*—using fictional scenarios to make predictions about the results of introducing artificial intelligence (AI) and other emerging technologies into these environments. Futurists engage in this process by providing scenarios to ameliorate uncertainty, drawing on a suite of tools that include simulations, worst-case planning, war-gaming, and even science fiction narratives.

A common futurethink tactic is to switch from risk-based probabilistic thinking, which is vulnerable to various decision-making pathologies, to possibilistic thinking—creatively generating scenarios outside of expected outcomes with a focus on impacts rather than probabilities. This move avoids some pathologies but is still subject to many biases and must be implemented judiciously. It can be usefully harnessed if futurists and policymakers avoid:

- rounding off probabilities, using heuristics as knowledge, and only exploring known outcomes.
- engaging in excessive deviations from reality and exotic or emotionally fraught scenarios.
- anchoring on specific scenarios, allowing embedded assumptions, and making hasty generalizations.

While still:

- being creative enough to spark new ideas.
- making explicit ideas inspired by fiction and embedding them in specific scenarios.
- seeking out expert contributions and integrating existing threats into scenarios.

INTRODUCTION

Policymakers are facing contemporary and future security and technology environments characterized by increasing uncertainty. In response, they have engaged in an approach that we call *futurethink*—using fictional scenarios to make predictions about the results of introducing artificial intelligence (AI) and other emerging technologies into these environments. Futurists are playing a key role in this process by providing scenarios that fill in gaps from missing information, drawing on a suite of tools that include simulations, worst-case planning, war-gaming, and even science fiction narratives.

This course of action would seem to make sense: predicting the future is inherently difficult, and the challenge is only exacerbated by the rapid emergence of new technologies, particularly AI, which stand to usher in entirely new ways of warfare. Moreover, national security practitioners have previously experienced catastrophic prediction failures. Policymakers have been continuously haunted by the inability to foresee the use of airplanes as weapons in the September 11 attacks, which was famously dubbed a "failure of imagination" by the authors of a congressional investigation into the attacks.[1]

The "failure of imagination" problem occurs when individuals are operating under *uncertainty*—missing information about the range of possible outcomes or the value or probability of those outcomes. The problem is particularly acute with predicting the future of AI-based conflict, for which the probabilities of different outcomes are nearly impossible to calculate. A common futurethink response to uncertainty is thus to switch from risk-based *probabilistic* thinking, which is vulnerable to various decision-making pathologies, to *possibilistic* thinking—creatively generating scenarios outside of expected outcomes with a focus on impacts rather than probabilities.[2] Possibilistic thinking can evade some pitfalls of probabilistic thinking but is subject to many biases and must be implemented judiciously.

In this paper we explore probabilistic and possibilistic approaches to uncertainty related to AI, outline their potential advantages and disadvantages, and identify common biases that hinder good prediction. We argue that creative prediction through fictional scenarios can be usefully harnessed if an active and systematic approach is taken that avoids prognostication pathologies by . . .

- avoiding the pitfalls of probabilistic thinking such as ignoring small probabilities, using heuristics, and focusing on known outcomes.
- engaging with possibilistic thinking effectively, by avoiding excessive deviations from reality, singular, or overly evocative scenarios, while making explicit ideas inspired by fiction and embedding them in specific scenarios.
- creating counterfactuals that do not make heroic assumptions about technology, organizations, or politics, while still being creative enough to spark new ideas.
- countering psychological biases that spring from a focus on individual outcomes, taken-for-granted embedded assumptions, emotionally fraught scenarios, and hasty generalizations.
- seeking out expert contributions about past and current technological trajectories and integrating existing threats into scenarios.

THE PERILS OF PROBABILISM

Until recently, the most common way for security and intelligence experts to think about risk has been probabilistic: identify a set of scenarios attached to choice options, estimate

the probability and impacts of each, rank-order the value of the choice options, and then take action to decrease potential loss and increase potential gain across scenarios. It is attractive because cost-benefit analysis can then be used to determine strategies.

The shortcomings of this approach for anticipating national security crises, however, are apparent. For one, cost-benefit calculations fail when the probabilities of potential outcomes occurring are so small that they are rounded off to zero through a process of simplification.[3] The FBI and the FAA didn't act on warnings of aircraft being used as weapons prior to the September 11 attacks because they "found the plot highly unlikely."[4] For another, probabilistic thinking presumes a knowledge not only of probabilities but also of outcomes. This leads to a focus on known problems rather than on problems that are outside of our limited imagination.

Yet the greatest shortcoming of the probabilistic approach to thinking about the future occurs when probabilities or impacts are unclear and a decision must be made anyway. Frequently, individuals rely on heuristics—decision-making shortcuts that "reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations."[5] Heuristics are generally based on preexisting knowledge about the world. Though often implicit and a recourse of last resort, they function as poor substitutes for the situational knowledge required to make high-stakes decisions, replace collection or analysis of actual information, and generate poor probability estimates. Three "classic" heuristics are described in the literature: *anchoring*, *availability*, and *representativeness*.[6] To this list we add the *affect heuristic* and *prospect theory*. All of these can produce pathologies when commissioning, using, and assessing scenarios—common across probabilistic and possibilistic thinking. These pathologies distort estimates at every step of the risk assessment chain, from intelligence analysts to planners. We walk through each heuristic in turn below.

Anchoring

Anchoring occurs when an individual's initial estimate of the probability or impact of a given risk serves as the basis for all further estimates.[7] Initial estimates that are too high or too low thus end up distorting policy since they are never re-evaluated. Anchoring is more likely to occur when a single scenario is used as a basis for planning and decision-making.[8]

However, even when multiple scenarios are used—a likely situation when agencies issue a call for futurethink scenarios—it is not uncommon for decision-makers to "anchor" to one particular scenario more than others, thus limiting rather than expanding their own mental maps. The use of multiple scenarios can also exacerbate uncertainty, making it even more difficult to determine the relative likelihoods of those scenarios, particularly when they speak to the same question or concern.

Availability Bias and Confirmation Bias

Prior experiences can result in implicit bias. One of the ways this bias manifests is in making future similar or analogous events seem more cognitively "available" and therefore seem more likely. Availability bias explains why we are always "fighting the last war," or when defense planning to counter future threats tends to resemble defense planning for the threats we know or have already seen—the very trap militaries seek to avoid in calling on futurists. However, in the case of AI, availability bias can lead to a narrow focus on extant systems that employ machine learning while novel and more potentially disruptive applications are disregarded. Moreover, when planners are presented with multiple scenarios from futurists, availability bias can lead to scenarios being rejected or accepted based on the extent to which they align with preconceptions: "A highly plausible scenario is one that fits prior knowledge well."[9]

The use of scenarios can, similarly, "prime" the imagination and make the predicted outcomes feel or appear more likely than those not predicted or outside the scope of the scenario.[10] This is known as confirmation bias—when individuals develop strongly held prior causal beliefs about a scenario or future state of the world, potential future events are judged to be more likely or more important if they resemble or conform to these preexisting beliefs.[11] Thus, exploring a given scenario runs the risk of encouraging complacency about the greater likelihood of that outcome.

Affect and Prospect

The use of scenarios for AI-driven security threats is also likely to be associated with fear and other strong emotional affects, which can cause "probability neglect."[12] Actors tend to conflate vulnerability (possibility) with risk (probability) when confronted with scenarios that evoke a subjective feeling of fear.[13] Thus the "affect heuristic" can allow highly fear-inducing scenarios to short-circuit rational processes.[14] These scenarios are likely to lead to hypervigilance around those events—and avoidance at all costs.

Prospect theory describes how people value the good or bad prospects, or outcomes, of different scenarios. Individuals tend to overweigh the value of potential losses relative to potential gains, making the losses all the more unacceptable, regardless of the probability associated with them.[15] As a result, individuals' fear of "the bad" tends to override their positive anticipation of "the good," independent of the presence of any heuristics. This can affect a policymaker's response to highly charged scenarios, allowing the fear of catastrophe to dictate decision-making processes aimed at avoiding losses—regardless of their probability.[16]

Affect heuristics and prospect theory imply that AI doomsday scenarios will pose a unique challenge to planners. Such scenarios typically include rhetoric and imagery that invoke intense emotions of fear, dread, and anxiety, and therefore crowd out other scenarios. These negative emotions can "dominate decision-making behavior, including the assessment of prospects and choice of alternatives."[17]

Representativeness

Representativeness is a cognitive bias that results in an increased propensity to assume that a single event under consideration resembles an existing, known category of events. For possibilistic thinking, this bias can result in associating particular outcomes with higher probability. The problem that representativeness poses for the use of futurethink, which tries to avoid this problem by producing scenarios consisting of multiple events, stems from the complexity of the scenarios posited, either in the way discrete events are linked or the level of detail presented.

People tend to judge multiple events to be more likely if their sequential unfolding fits more closely with the observers' preestablished worldview and experiences (representativeness). This leads to the conjunction fallacy, whereby people erroneously think that a scenario with a greater number of conditions (requiring A+B rather than just A to occur) is more likely to occur than one with fewer conditions.[18] The more detailed and vivid futurethink AI scenarios are—which is to say, the more events they contain—the more likely they will seem to be.

As *representativeness* and other heuristics illustrate, even when probability information is available, humans tend to struggle with engaging in cost-benefit analysis in an unbiased way. Indeed, we struggle at every stage of the risk calculation process, both in accurately describing the likelihood of events and identifying and quantifying potential gains and losses. By and large, humans tend to have a limited ability to intuitively estimate the probabilities or impacts of different risks. Indeed, individuals are remarkably

poor at understanding, ascribing, and interpreting the likely occurrence and impact of a given risk. Under conditions of uncertainty, when probability estimates required to make predictions about the future are missing, policymakers turn to futurethink to try to avoid these pathologies.

THE POTENTIAL OF POSSIBILISM

Futurethink requires shifting to what the sociologist Lee Clarke calls possibilistic thinking, focusing on impacts instead of probabilities.[19] Embracing uncertainty and accepting that probabilities are not estimable avoids simplification, the use of heuristics for probability estimation, and a lack of imagination. It can raise awareness of the severity of the impacts of some scenarios as well as the conditions under which they are likely to emerge, even if the probability of those conditions is difficult or impossible to estimate.

This mode of thinking is thus appropriate for events whose probabilities are thought to be low (and thus are disregarded) or are unknown (that is, conditions of uncertainty). Possibilistic scenarios are commonly explored through simulations, worst-case scenario planning, war-gaming, and science fiction narratives. The best possibilistic approaches produce generalizable lessons, even if the scenarios themselves are largely fictional. For example, the 2001 Dark Winter smallpox exercise demonstrated gaps in policymakers' understanding of biological weapons attacks, the limited number of policy responses available, a lack of surge capacity in medical facilities, state-federal conflicts in policymaking, and the need to engage civilians as participants in the response, not just as victims of the disease.[20]

Thinking "catastrophically" about AI in the way that possibilism requires—or conjuring crisis scenarios that play out in the future—is now in high demand. Interdisciplinary centers at major research universities, such as Oxford's Future of Humanity Institute or the Center for Study of Existential Risk at the University of Cambridge, have recently focused on the risks posed by AI. Likewise, both technical institutes, such as the Machine Intelligence Research Institute, and more general purpose research organizations, such as the think tank New America, are engaging in AI futurism.[21]

Scenario Solutions?

The current vogue for futurethink emerged from an awareness of our own biases and, specifically, a tendency to assume the future will look like the past. Thinking systematically and actively about fictional scenarios is inherently superior to unconsciously letting well-known scenarios shape our assessments of those trajectories. Without creative thinking we cannot predict or counter future threats that exist outside the framework of linear extrapolations from current threats, although such creativity does need to be grounded in expert knowledge of current and past technological developments. Futurethink can also provide guides for intelligence or filters for overwhelming amounts of data. In the realm of AI, this is particularly true, as the rate of technological progress and potential impacts are difficult to determine. The problem is made worse by the unpredictability of the co-evolution of AI's underlying hardware and software. "Futuristic" possibilistic scenarios can offer a kind of baseline from which to predict and hedge against negative outcomes.

This is not all that can be gained from scenario-based planning. Scenarios tend to have four major advantages: limiting bounded rationality, considering endogenous and exogenous variables, reducing stickiness, and revealing the premises of mental models.[22] Likewise, multiscenario exploratory analysis can help explore the "maximum scenario space" by combining traditional scenario planning with computer modeling and formal game theory

to reduce the cognitive issues associated with worst-case-scenario planning. These approaches can help decision-makers through effective presentation of "stories" (scenarios) in conjunction with more formal assessments.[23] This combination of scenarios, modeling, and game theory to aid in planning is used across most U.S. federally funded research and development centers and university affiliated research centers, as well as by other government contractors, such as SAIC (Science Applications International Corporation).

However, while processes exist to deal with estimating the impacts and likelihood of known scenarios, methods for generating unknown possible states of the world in a way that leads to policy-relevant applications are sorely lacking. As Richard Danzig puts it, "The propagation of scenarios, however sophisticated, broad-ranging, or insightful, does not obviate the need for strategies for coping with uncertainty."[24] Selecting which future scenarios are most applicable for policy can be difficult: In a world with an infinite number of future trajectories, how do we select for the most important ones?

The Counterfactual Future

Since the future *is* a counterfactual—a description of an alternative, plausible reality—we can use guidelines developed for counterfactual thinking. Ideally, counterfactuals represent a "minimal rewrite of history" and offer alternative "possible worlds" predicated on existing capabilities.[25] The first guideline can be adapted for futurethink as a "minimal rewrite of the future," or the need to avoid heroic assumptions about or major disruptions affecting technological (and other) trajectories. AI predictions that would violate this guideline include, for example, predictions that require violations of Moore's Law; near-term breakthroughs in the scale of quantum computing; assumptions that humankind will soon hit the "singularity;" or machine learning approaches or datasets that either don't exist or would be impossible to gather.[26] Futurethink about the security threats associated with AI is like the unwavering faith some still have in the efficacy of national missile defense: there are so many leaps of faith required for AI futurethink scenarios to function that it has become a satirical pastime to point out all of the errors.[27]

The second requirement regarding "possible worlds" stipulates that decisions made must be possible in light of real, existing capabilities. For AI-driven futurethink, "possible worlds" can be interpreted as a guideline that limits events to those that are possible, given current political and organizational capabilities. For example, assumptions about adversaries (and ourselves) perfectly and completely adopting and implementing AI capabilities are unrealistic, given the slow and sometimes abortive process of innovation and the complex assemblage of hardware and software technologies that AI requires.[28] Moreover, for the United States to deploy any AI-driven capability, we must also consider the timeline associated with procurement, including testing, evaluation, verification, and validation of new systems.

It is, however, insufficient for scenarios to minimally rewrite the future in a way that can be identified as a possible world; they must also be creative in such a way as to produce lessons beyond its narrow scope:

> This world must differ from the given in at least one way, and this one way must be sufficient to give rise to events that could not occur in our society. . . . The new idea . . . must be truly new (or a new variation on an old one) and it must be intellectually stimulating to the reader; it must invade his mind and wake it up to the possibility of something he had not up to then thought of.[29]

This broad and creative approach can be contrasted with the approach frequently taken by U.S. government entities ranging from DARPA to the Secret Service, spanning the breadth

of the Intelligence Community: futurists are tasked with exploring specific scenarios, and inevitably predictions and outcomes end up being biased toward those scenarios. This leads to the perception of an increased likelihood for those scenarios and hence to overweighting the likelihood of those outcomes. Government commissions can thus accidentally exacerbate anchoring to some scenarios while ignoring others.

Fictional Overload

Scenario generation can have significant downsides. One of the many risks is the tendency to focus on a limited number of exotic scenarios at the expense of a larger number of more commonplace ones. For example, the AI Paperclip Apocalypse—in which an AI created with the goal of manufacturing paperclips inundates the world with paperclips and ends up exterminating humanity in a war over resource control—is extraordinarily implausible but strikes a darkly humorous chord that crowds out more likely dangers, such as creating an AI that deliberately seizes power or harms others.[30] Nancy Kanwisher argues that "there is evidence that strategic priorities have in the past become distorted by overemphasizing the most extreme scenarios at the expense of less flashy but more likely ones."[31] She points out that 90 percent of RAND nuclear war scenarios in 1960 assumed a surprise attack on the United States, despite the fact that such an attack was a highly unlikely scenario. Moreover, argues Kanwisher, worst-case-scenario planning may actually increase the probability of other negative and more plausible events. For example, surprise-attack scenarios lead to placing nuclear weapons on hair-trigger alert and lead to the creation of decision-making structures geared entirely to speed, thus increasing the probability of accidental or malicious use.[32] Such scenarios "retool" the threat detection bureaucracy and apparatus to become attuned to detecting and responding to these specific worst-case scenarios. Hypervigilance around Iraq's weapons of mass destruction programs due to the "one percent doctrine"—that is, if there is a 1 percent chance that a threat is real it has to be treated as a certainty—is a tragic example of the danger of possibilistic thinking taken to its limit.[33] We should strive instead to distribute attention across a variety of scenarios, rather than focusing solely on a known (and low-probability) threat that fits into existing narratives.

An additional problem with worst-case-scenario planning is that scenarios can be influenced by prior beliefs regarding the world as well as fictional narratives. This influence can occur unconsciously or consciously, through the deliberate use of science fiction as a predictive tool. Indeed, the influence of science fiction on policy is already quite significant: by providing metaphors and conceptual frameworks for analysis and by drawing attention to potentially catastrophic outcomes, science fiction risks displacing knowledge from other sources.[34] Consumption of films with armed AI is correlated with greater opposition to autonomous weapons (since all AI is armed AI, and all armed AI is the Terminator), and science fiction is used casually in discourse to advocate on either side of the autonomous weapons debate ("What about the Matrix?" "That's science fiction, we don't need to think about that").[35] News articles on autonomous weapons tend to focus on images that are either pointlessly humanoid (often Terminator-style with glowing red eyes) or wildly incorrect (stock photos of drones), although some do show realistic prototypes (treaded vehicles with mounted cameras and weapons).

Fictional narratives have unconsciously shaped assessments of the likelihood of numerous strategic threats throughout multiple presidencies. President Ronald Reagan's determination to reduce nuclear dangers resulted from several cognitive-psychological factors: a Pentagon Single Integrated Operational Plan (SIOP) briefing, which is a form of scenario

planning, primed him to be sensitive to the risk of nuclear war. This sensitivity was enhanced by Reagan's watching *The Day After*, a film depicting a horrific vision of nuclear holocaust. Admittedly, he was predisposed to this view: Reagan's interpretation of the Bible led him to believe that Armageddon would come in the form of a nuclear holocaust.[36] As president, Reagan also inquired into the security of the U.S. nuclear command and control structure after watching *Wargames*, an early cinematic treatment of an AI worst-case scenario. Similarly, President Bill Clinton increased U.S. investment in biodefense after reading *The Cobra Event*, even though the premise of that book was biologically implausible.[37]

Not all of fiction's influence has been accidental, however. Some authors have written fiction at policymakers' requests and with the deliberate intention of shaping policy. Tom Clancy was interviewed on 9/11 as a terrorism expert because he had included an aircraft crashing into the Capitol in *Debt of Honor*. Clancy's influence dates back to Reagan, who read *Red Storm Rising* to prepare for the 1986 Reykjavik summit with Mikhail Gorbachev.[38] Peter Singer was motivated to help the United States prevail in subsequent wars driven by weapons of the future when he and August Cole wrote *Ghost Fleet*.[39]

The turn to new tools and futurism, engaging in "possibilistic" instead of "probabilistic" thinking, and using scenarios to connect the dots and fill in the gaps—in short, using imagination to preempt what would otherwise be unforeseeable catastrophe—can thus be a valuable approach. But it must identify, anticipate, and avoid pathologies that lead to risk distortion and incomplete decision-making procedures. The absence of probability estimates alone cannot prevent decision-makers from anchoring on certain scenarios, treating some alternatives as more likely than others, or distorting the weighing of impacts.

CONCLUSION: INTEGRATING THINKING, AVOIDING BIAS

Both probabilistic and possibilistic thinking can be useful for generating and evaluating scenarios. Policymakers and futurists can minimize potential pitfalls if, together, they select and elaborate on posited scenarios consciously and deliberately, rather driving scenarios through implicit narratives or heuristics. Probabilistic approaches are valuable if policymakers can avoid . . .

- implicitly or explicitly rounding off probabilities.
- substituting heuristics for knowledge.
- limiting search to known outcomes.

Futurists can aid in ameliorating some of these pitfalls by taking possibilistic approaches, but these come with their own disadvantages. To avoid them, possibilistic thinking must engage with scenarios in such a way that:

- the deviations from reality are close enough to allow for generalizable lessons.
- additional scenarios, and more commonplace ones, are considered rather than just a few exotic ones.
- metaphors, narratives, and ideas from science fiction are engaged with explicitly.
- imported ideas from fiction are embedded in specific scenarios.

Policymakers must be vigilant: under any approach, scenario consideration can suffer from a number of psychological biases. Whether decision-makers are likely to fall victim to irrationality in the use of scenarios is determined by their preexisting mental state (including bias from fear or other preexisting attitudes) and the degree of uncertainty about the event. Defense planners can make the best use of scenarios by

effectively soliciting, considering, or adjudicating them, as well as by appropriately weighting a proportional response. To wit, policymakers and defense planners must avoid . . .

- anchoring on a single scenario, probability, or impact estimate.
- taking for granted existing beliefs and knowledge about available solutions.
- being swayed by scenarios that have significant affective elements.
- assuming that individual scenarios are "representative" examples.

Beyond these biases, policymakers must ensure that scenarios are being created by experts with sufficient knowledge of the subject area. They must also keep track of existing threats, either those occurring in isolation or in combination with generated scenarios. Ideally, the futurists at hand have enough expertise in AI to effectively "connect the dots" where non-experts cannot. Even AI experts can make overly precise (and very inaccurate) predictions; Eliezer Yudkowsky of the Machine Intelligence Research Institute used his own work as an example, admitting that at one point he predicted a 90 percent chance of AI on par with human capabilities being developed between 2005 and 2025, with a peak in 2018: "This statement now seems to me like complete gibberish. Why did I *ever* think I could generate a tight probability distribution over a problem like that? Where did I even get those numbers in the first place?[40]

Because we can't figure out everything from scratch each day—unknown unknowns are infinite—we turn to futurists. In doing so, we must avoid both under- and overcorrecting prior decision-making errors. Policymakers must engage in futurethink through carefully targeted efforts to commission a diverse set of scenarios, then systematically interpret and analyze the results. For their part, futurists should ground mechanisms and events in established theories and experiences, draw on expert knowledge, and engage in careful counterfactual reasoning while avoiding heroic assumptions about adversary characteristics and response capabilities.

REFERENCES

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

Carpenter, Charli. 2016. "Rethinking the Political / -Science- / Fiction Nexus: Global Policy Making and the Campaign to Stop Killer Robots." *Perspectives on Politics* 14, no. 1: 53–69.

Cegłowski, Maciej. 2016. "Superintelligence: The Idea That Eats Smart People" (https://idlewords.com/talks/superintelligence.htm; accessed July 13, 2020).

Chermack, Thomas J. 2004. "Improving Decision-Making with Scenario Planning." *Futures* 36, no. 3: 295–309.

Clarke, Lee. 2006. *Worst Cases: Terror and Catastrophe in the Popular Imagination*. University of Chicago Press.

Clarke, Lee. 2008. "Possibilistic Thinking: A New Conceptual Tool for Thinking about Extreme Events." *Social Research* 75, no. 3: 669–90.

Connell, Louise, and Mark T. Keane. 2006. "A Model of Plausibility." *Cognitive Science* 30, no. 1: 95–120.

Daniel, J. Furman, and Paul Musgrave. 2017. "Synthetic Experiences: How Popular Culture Matters for Images of International Relations." *International Studies Quarterly* 61, no. 3: 503–16.

Danzig, Richard. 2011. "Driving in the Dark: Ten Propositions about Prediction and National Security." Washington: Center for a New American Security.

Davis, Paul K. 2012. *Lessons from RAND's Work on Planning under Uncertainty for National Security*. Santa Monica, Calif.: RAND.

Davis, Paul K., Steven C. Bankes, and Michael Egner. 2007. *Enhancing Strategic Planning with Massive Scenario Generation: Theory and Experiments*. Technical report TR-392. Santa Monica, Calif.: RAND National Security Research Division.

Dick, Philip K. 1995. "My Definition of Science Fiction." In *The Shifting Realities of Philip K. Dick: Selected Literary and Philosophical Writings*. New York: Vintage Books.

Fischer, Beth A. 2013. *The Reagan Reversal: Foreign Policy and the End of the Cold War*. University of Missouri Press.

Friedman, Benjamin H. 2011. "Managing Fear: The Politics of Homeland Security." *Political Science Quarterly* 126, no. 1: 77–106.

Gans, Joshua. 2018. "AI and the Paperclip Problem." VoxEU.org (https://voxeu.org/article/ai-and-paperclip-problem; accessed July 11, 2020).

Healey, Mark P., and Gerard P. Hodgkinson. 2017. "Making Strategy Hot." *California Management Review* 59, no. 3: 109–34.

Hodgkinson, Gerard P., and George Wright. 2002. "Confronting Strategic Inertia in a Top Management Team: Learning from Failure." *Organization Studies* 23, no. 6: 949–77.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47, no. 2: 263–91.

Kanwisher, Nancy. 1989. "Cognitive Heuristics and American Security Policy." *Journal of Conflict Resolution* 33, no. 4: 652–75.

Kean, Thomas H., Lee H. Hamilton, Richard Ben-Veniste, Bob Kerrey, Fred F. Fielding, John F. Lehman, Jamie S. Gorelick, and others. 2004. *The 9/11 Commission Report*. National Commission on Terrorist Attacks upon the United States.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books.

McDermott, Rose. 1998. *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy*. University of Michigan Press.

Montgomery, Alexander H. 2020. "Double or Nothing? The Effects of the Diffusion of Dual-Use Enabling Technologies on Strategic Stability." CISSM Working Paper. University of Maryland, School of Public Policy, Center for International and Security Studies (https://cissm.umd.edu/research-impact/publications/double-or-nothing-effects-diffusion-dual-use-enabling-technologies; accessed July 27, 2020).

Moore, George E. 1965. "Cramming More Components onto Integrated Circuits." *Electronics* 38, no. 8 (April 19).

Nichols, Thomas M. 2017. *The Death of Expertise: The Campaign against Established Knowledge and Why It Matters*. Oxford University Press.

Slovic, Paul, Melissa Finucane, Ellen Peters, and Donald G MacGregor. 2002. "Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics." *Journal of Socio-Economics* 31, no. 4: 329–42.

Sunstein, Cass R. 2002. "Probability Neglect: Emotions, Worst Cases, and Law." *Yale Law Journal* 112, no. 1: 61–107.

Suskind, Ron. 2006. *The One Percent Doctrine: Deep inside America's Pursuit of Its Enemies Since 9/11*. New York: Simon & Schuster.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, September 27, 1974, 1124–31.

Tversky, Amos, and Daniel Kahneman. 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90, no. 4: 293–315.

Wellerstein, Alex. 2019. "NC3 Decision Making: Individual versus Group Process." NAPSnet Special Report. Berkeley, Calif.: Nautilus Institute for Security and Sustainability, August 8 (https://nautilus.org/napsnet/napsnet-special-reports/nc3-decision-making-individual-versus-group-process; accessed April 8, 2020).

Whiskey Fueled Tirade. 2019. "Point/Counterpoint: Future Wars Will Be Fought with AI Robots vs. 'Microsoft Word Is Not Responding.'" *DuffelBlog* (www.duffelblog.com/2019/05/point-counterpoint-future-wars-will-be-fought-with-ai-robots-vs-microsoft-word-is-not-responding; accessed October 6, 2019).

Young, Kevin L., and Charli Carpenter. 2018. "Does Science Fiction Affect Political Fact? Yes and No: A Survey Experiment on 'Killer Robots.'" *International Studies Quarterly* 62, no. 3: 562–76.

Yudkowsky, Eliezer. 2011. "Cognitive Biases Potentially Affecting Judgement of Global Risks." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 91–119. Oxford University Press.

ENDNOTES

1 Kean and others 2004, 339–48.

2 Clarke 2006; Clarke 2008.

3 McDermott 1998, 24.

4 Quoted in Clarke 2006, 44.

5 Tversky and Kahneman 1974, 1124.

6 Tversky and Kahneman 1974.

7 Tversky and Kahneman 1974.

9 Kanwisher 1989.

9 Connell and Keane 2006, 95.

10 Healey and Hodgkinson 2017.

11 Nichols 2017, 47–69.

12 Sunstein 2002.

13 Friedman 2011.

14 Slovic and others 2002.

15 Kahneman and Tversky 1979.

16 McDermott 1998.

17 Hodgkinson and Wright 2002, 579.

18 Tversky and Kahneman 1983, 308.

19 Clarke 2006; Clarke 2008.

20 Clarke 2008, 681–83.

21 Future of Humanity Institute: www.fhi.ox.ac.uk; Machine Intelligence Research Institute: www.intelligence.org; Cambridge University, Centre for the Study of Existential Risk, "Risks from Artificial Intelligence": www.cser.ac.uk/research/risks-from-artificial-intelligence/; New America, Open Technology Institute, "What Sci-Fi Futures Can (and Can't) Teach Us About AI Policy": www.newamerica.org/oti/events/what-sci-fi-futures-can-and-cant-teach-us-about-ai-policy/.

22 Chermack 2004.

23 Davis, Bankes, and Egner 2007; Davis 2012.

24 Danzig 2011, 19.

25 Clarke 2008, 684–86.

26 Moore 1965; Kurzweil 2005; Bostrom 2014.

27 Cegłowski 2016; Whiskey Fueled Tirade 2019.

28 Montgomery 2020.

29 Dick 1995.

30 Gans 2018.

31 Kanwisher 1989, 655.

32 Wellerstein 2019.

33 Suskind 2006.

34 Daniel and Musgrave 2017.

35 Carpenter 2016; Young and Carpenter 2018.

36 Fischer 2013.

37 Daniel and Musgrave 2017, 505.

38 Daniel and Musgrave 2017, 511.

39 Daniel and Musgrave 2017, 512–13.

40 Yudkowsky 2011, 113.

ABOUT THE AUTHORS

Dr. Alexander H. Montgomery is an Associate Professor of Political Science at Reed College.

Dr. Amy J. Nelson conducted this work as part of her research with the Center for International and Security Studies at the University of Maryland, where she is a Research Associate. She is also a Fellow at the Center for the Study of Weapons of Mass Destruction at National Defense University.

All views expressed in this paper are the authors' own.

ACKNOWLEDGEMENTS

We thank Heather Roff, Chris Meserole, and an anonymous reviewer for many helpful suggestions on earlier drafts and Nicholas Winstead for excellent research assistance. The authors are equally responsible for the article; names appear in alphabetical order.

The Brookings Institution is a nonprofit organization devoted to independent research and policy solutions. Its mission is to conduct high-quality, independent research and, based on that research, to provide innovative, practical recommendations for policymakers and the public. The conclusions and recommendations of any Brookings publication are solely those of its author(s), and do not reflect the views of the Institution, its management, or its other scholars.