AI IN THE AGE OF CYBER-DISORDER
Actors, Trends, and Prospects

edited by **Fabio Rugge**

introduction by **John R. Allen** and **Giampiero Massolo**

ISPI

BROOKINGS

AI in the Age of Cyber-Disorder

Actors, Trends, and Prospects

edited by Fabio Rugge

ISPI   BROOKINGS

**Ministry of Foreign Affairs
and International Cooperation**

BROOKINGS

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.

Table of Contents

Introduction

In the summer of 1956, mathematician John McCarthy and another 19 scholars met at Dartmouth College on the outskirts of Hannover in New Hampshire for a summer research project on machine intelligence. This was the moment when what would later become known as Artificial Intelligence (AI) first emerged as a research discipline. Meanwhile, two other scholars, Harold J. Leavitt and Thomas L. Whisler, were writing an article speculating on new technologies and dubbing them information technology (IT). With this definition, they referred to technology of three types: for processing large amounts of information rapidly; for applying statistical and mathematical methods to decision-making problems; and for simulating higher-order thinking through computer programs. Artificial Intelligence and cyberspace were therefore born in the same historical context, and have been intertwined since their inception.

Artificial Intelligence is a general-purpose technology that is considered one of the most promising of our times. To function, AI needs to be trained, and it finds its training material in data. These are collected, processed, exchanged and stored in the digital domain. According to a recent estimate, the global size of the data sphere reached 59 zettabytes in 2020. This is an unexpectedly large figure mainly due to the outbreak of the Covid-19 pandemic, which has pushed many around the world to consume and produce more digital data than last year. Indeed, a multitude of activities we used to do in-person have

turned into 2.0 happenings, including academic conferences, professional meetings, yoga classes, and even birthday parties. Recently, these have occurred in the digital domain (commonly called cyberspace) and generated an unprecedented amount of data.

There is, nowadays, a sound literature investigating how AI could exploit data analysis to generate effects in our contemporary society, from medical applications to surveillance capitalism. In light of such understanding, cyberspace is one of the domains in which AI may play a role in achieving overarching objectives. For example, AI-enabled cyber-attacks may drive the next major upgrade in cyber-weapons, which could be used both by state and non-state actors. The relationship between AI and cyberspace has so far been investigated mostly in terms of the effects that AI could have on the digital domain and thus on our societies. What has been explored less is the opposite relationship, namely, how cyberspace geopolitics can affect AI.

Therefore, in this second ISPI-Brookings Report, we decided to turn the picture upside-down and to focus our analysis on how ongoing dynamics in cyberspace as a whole may affect how AI technologies are developed and implemented. Indeed, cyberspace is not a neutral and aseptic environment. It is not even a utopian paradise where people can unite and meet across frontiers in a stateless digital reign. As Fabio Rugge argues in the opening chapter, cyberspace is now becoming the domain of choice for serious but ambiguous competition among nation states, and for destabilisation by non-state actors such as mercenaries, cybercriminals and hacktivists. This is the environment in which cyber-dependent technologies will be developed and fielded. As such, AI may accelerate the ongoing confrontation at the international level, adding yet another layer of unpredictability to an already troublesome and ambiguous digital domain. The chapters that follow will develop in two related directions. On the one hand, they will lead us to discover how an inadequately governed domain can have undesirable consequences for how AI technologies are

already implemented. On the other hand, they will also propose policies, and practical and ethical guidelines, to govern the use of AI more effectively and thus reduce the negative influence of geopolitical competition in cyberspace on AI applications.

Indeed, the reckless use of AI can pose many risks and be detrimental to the founding principles of our international system, which is bound by the Universal Declaration of Human Rights. For example, the protection of human dignity at the individual level is currently endangered by the pervasiveness of surveillance programs, and particularly by those implemented in nondemocratic regimes. As Samuele Dominioni explains, authoritarian countries are exploiting the normative void around surveillance practices to strengthen their grip on power by building a kind of "digital panopticon". The lack of proper regulations to establish pre-conditions for AI-enabled technologies is also evident in other seemingly domestic settings. As Caitlin Chin and Mishaela Robison demonstrate, AI bots and voice assistants reinforce gender bias due to societal tendencies to anthropomorphise objects by assigning gender.

From offensive applications, pervasive surveillance, and algorithm biases, it seems that AI applications have suffered from a growing disorder and lack of enforceable normative solutions in the cyber domain. Yet, AI could be beneficial in tackling some of the threats currently coming from such a contested digital domain. As John Villasenor argues, AI could play a determining role in addressing the rapid and pervasive disinformation attacks that are increasingly affecting our societies. However, purely technological solutions are not enough. To tackle these challenges, actions must be multi-layered. On the one hand, as Darrell West argues, this era requires a "whole-of-society" and "whole-of-government" approach that considers how to maximise possible benefits and minimise costs. This includes creating uniform standards in terms of data access, data sharing, and data protection. On the other hand, we should never lose sight of the human dimension of technological developments. Along these lines, Paolo Benanti claims that it is necessary to

establish an international ethical governance system to identify and monitor developments in AI, weighing their potentially disruptive effects against their capabilities.

As long as the cyber domain remains in turmoil, with no effective governance mechanism in place, it is likely that it will be challenging to find human-centric solutions that might allow humanity as a whole to benefit from the tremendous potential that AI has to offer. This report by ISPI and the Brookings Institution is an effort to shed light on this less studied, but extremely relevant, relationship.

John R. Allen   *Giampiero Massolo*
President Brookings Institution   *President ISPI*

1. AI in the Contested Cyberspace

Fabio Rugge

The Internet, born as an anarchic network to connect people across global frontiers, has become one of the most destabilizing areas of competition between states. The Internet is being militarized[1], as cyberspace is now a domain where states' power is projected, strategic goals are achieved with campaigns that fall under the definition of the *use of force*, weapons systems are operated, wars will be fought. Cyberspace is also the domain where AI is developing as a general-purpose technology transforming practically all aspects of our lives, including in the area of international security. Thanks to increases in computing power, production and storage of data and more advanced algorithms, AI contributes in making algorithmic decision-making ubiquitous in the security and defense sectors, from assisting strategic decision-making to regulating the functioning of critical infrastructures, from strengthening cyber defense and attack capabilities to analyzing from multiple sources in real-time the geo-positioning of assets devoted to nuclear deterrence.

We will highlight how the emergence of cyberspace resulted in a more contested international security environment,

---

[1] "We recognize that adversaries already condemn U.S. efforts to defend our interests and allies as aggressive, and we expect they will similarly seek to portray our strategy as "militarizing" the cyberspace domain. The Command makes no apologies for defending U.S. interests as directed by the President through the Secretary of Defense in a domain already militarized by our adversaries". *Achieve and Maintain Cyberspace Superiority*, Command Vision for U.S. Cyber Command, United States Cyber Command, 14 June 2018, p. 10.

and we will explore how AI will accelerate the ongoing confrontation in and through cyberspace, adding yet another layer of unpredictability to an already troublesome domain of ambiguity. The ongoing global decoupling of the ICT supply chain, the current fragmentation of the Internet and the degree of investment underway in the area of AI, in particular in the United States, China and Russia, are all unequivocal signs that an arms race for technological supremacy in the cyber domain is indeed ongoing, and is in fact one of the most prominent features of today's Great Power Competition. AI may be seen and operates as a Marxian superstructure on the sub-structure (base) of cyberspace, inevitably obliging to (while, at the same time, concurring to shape) the same building blocks, rules and operating principles of the cyber domain. Just like the political and diplomatic efforts to mitigate the risks of escalations in and from cyberspace will hardly succeed in reverting the ongoing global race to attain cyber superiority, it is highly unlikely that the international community will be able to deliver common principles and norms of behavior to regulate algorithmic warfare. Liberal democracies will therefore have a two-pronged objective: that of preserving the technological edge that has historically been associated with their international hegemony while upholding the principles and human-centric values that distinguish us from autocratic regimes.

A Contested Domain

Cyberspace has become far too relevant for everyday life not to also be the lynchpin around which national interests naturally collide. Cyber power, which can be defined as "the ability to use cyberspace to create advantages and influence events in other operational environments and across the instruments of power"[2],

---

[2] D. T. Kuehl, "From Cyberspace to Cyberpower: Defining the Problem", in F.D. Kramer, S.Starr, and L.K. Wentz (eds.), *Cyberpower and National Security*, Washington D.C., National Defense University Press, Potomac Books, 2009,

is now clearly another dimension of XXI century sovereignty.[3] In many ways, this is nothing new: the more important privacy, accessibility and integrity of data become for national security, the more urgent it is for states to bolster cybersecurity and the more potentially advantageous offensive actions in cyberspace come to be. The same has happened in terrestrial, naval and aerial warfare and is now also happening in the extra-atmospheric space. Cyberspace is the "domain of ambiguity", where high-end threats operate in the same environment and share many of the technical features of low-level skirmishes and criminal activities. In this domain, it is troublesome to understand and anticipate the motivation and the scope of a cyber campaign without considering the strategic, political and operational context in which it occurs. The difficulty in attributing cyber attacks, together with the widespread recourse to false flag computer network operations, make it difficult to know "what is really going on" in the cyber domain, and to make sense of it. National intelligence communities usually are better equipped to handle sensible information and grasp the complexity "behind the curtains" of the ongoing confrontation in the cyber domain – but this is also another reason why an in-depth understanding of cyber affairs is not easily accessible to the general public. Moreover, a cyber attack may be prepared taking all the necessary time, but its execution requires an immediate response, even if the true scope and ultimate intent of a cyber campaign often becomes clear only once its objectives are met. Additionally, Computer Network Operations (CNO) are intrinsically asymmetric: it is easier and far less expensive to attack that to defend, in part because the

---

quoted and adopted by Prof. Joseph Nye in his "Cyber Power", Belfer Center for Science and International Affairs, Harvard Kennedy School, May 2010.

[3] "Cyberspace will no longer be treated as a separate category of policy or activity disjointed from other elements of national power. The United States will integrate the employment of cyber options across every element of national power". *National Cyber Strategy of the United States of America*, September 2018, p. 20.

potential benefits of an attack are incomparably greater than the risks of any retaliation. Moreover, the potentially short life of zero-days poses dangerous "use it or lose it" dilemmas to policymakers and military commanders in times of crisis.[4] Cyberspace has therefore become the domain of choice for destabilizing campaigns and hostile activities that would be simply unsustainable in the conventional realm.

All members of the international community regularly engage in the collection of valuable intelligence, including through computer network operations and signal intelligence (SigInt) support to cyber defense (SSCD) – after all, these are all endeavors not forbidden by international law, and it is a legitimate goal for every state to strengthen all dimensions of its sovereign power. In a security environment in which "it is undeniable that homeland is no longer a sanctuary",[5] the use of cyber power is essential in enhancing national security. In this sense, cyberspace is simply a new domain in which the

---

[4] "The implications of this logic are not limited to the cyber domain. Nor are they limited to Russia, China being as much a concern in this regard as well. There will be strong incentives in a serious crisis for China to initiate and rapidly escalate attacks against U.S. space infrastructure. While China may not wish to initiate such attacks, it could feel compelled to strike in space before the United States does, rather than risk the far more dangerous alternative of striking second. This same dynamic is pertinent in the cyber domain as well as the space domain. In short, the world faces a new and highly dangerous pressures where, even if the dynamics of the environment are understood at a given point in time, technological change could easily upend that new understanding in a relatively short time". A. Bidwell, JD and B.W. MacDonald, *Emerging Disruptive Technologies and Their Potential Threat to Strategic Stability and National Security*, Federation of American Scientists, September 2018, p.7.

[5] "It is now undeniable that the homeland is no longer a sanctuary. America is a target, whether from terrorists seeking to attack our citizens; malicious cyber activity against personal, commercial, or government infrastructure; or political and information subversion. New threats to commercial and military uses of space are emerging, while increasing digital connectivity of all aspects of life, business, government, and military creates significant vulnerabilities. During conflict, attacks against our critical defense, government, and economic infrastructure must be anticipated". *2018 US National Defense Strategy*, p. 3.

never-ending international confrontation takes place,[6] with the noteworthy difference that it is a "domain of ambiguity" where geographical frontiers are irrelevant, actors are largely unknown, civilian assets are often the main targets, and the rules of states' behavior are difficult to identify, tough to establish and problematic to enforce.

The international community has been actively involved for more than twenty years[7] in an effort to identify agreed-upon rules of states' behavior in cyberspace, but little has been so far accomplished. If a cyber armaments control regime seems unlikely to emerge in the near future, as trust among key international players is low and verification of compliancy is impossible, some encouraging progress has been achieved so far within OSCE. Two sets of confidence building measures (CBMs) have been adopted so far, listing (admittedly very

---

[6] "Challenges to United States security and economic interests, from nation states and other groups, which have long existed in the offline world are now increasingly occurring in cyberspace", *National Cyber Strategy of the United States of America*…, cit., p. 20.

[7] "Back in 1998 (while Operation "Moonlight Maze", one of the first and most devastating cyber campaign ever orchestrated by Russia's intelligence against U.S. military targets, was well underway…) the Russian Federation presented to the UN General Assembly a proposal for a Resolution titled "Developments in the field of information and telecommunications in the context of international security". The Russians wanted to discuss both cyber security and the limitations to destabilizing online content (revealingly gathered together by Moscow under the label of "threats to the information space"). The West refused to have that discussion, on the ground, essentially, of its self-proclaimed moral superiority: if we want to safeguard an open Internet and freedom of expression, the West argued, it is not possible to negotiate about information's content. Ironically, almost twenty years later, the West is forced to discuss with Moscow about the threat of manipulated online content, which probably is, in itself, a score on the Russian side", F. Rugge, *Mind Hacking: Information Warfare in the Cyber Age*, ISPI Analysis no. 319, January 2018, pp. 3-4; reproduced by the *Global Solutions Journal*, vol. 1, no. 1, May 2018. P. Pawlak, "Confidence-Building Measures in Cyberspace: Current Debates and Trends", in A.-M. Osula and H. Rõigas (eds.), *International Cyber Norms: Legal Policy and Industry Perspectives*, Tallinn, NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) Publications, 2016, pp. 129-153.

generic) voluntary commitments of the member-states to "establish international level of expectations about states' behavior in cyberspace"[8] with the purpose of improving stability and encouraging trust, cooperation and transparency among states. Together with other international efforts devoted to identifying norms of international law applicable to the conduct of states in cyberspace (especially within the United Nations), these measures help enhance predictability – and, therefore, provide some order – within the international community by establishing what is the prevalent *opinio juris* about permissible behavior in cyberspace, and by ensuring channels of communication that might one day prove useful to mitigate and defuse crisis stemming from the ongoing international confrontation. However, they certainly do not constitute binding norms of conduct.[9]

Establishing clear norms of acceptable behavior in cyberspace and deterring malicious cyber campaign is difficult enough among states, but it could prove futile against non-state actors. If, in today's security environment, non-states actors may

---

[8] P. Pawlak, "Confidence-Building Measures in Cyberspace: Current Debates and Trends", in A.-M. Osula and H. Rõigas (eds.), *International Cyber Norms*..., cit.

[9] "At this stage, large scale formal treaties regulating cyber space seem unlikely. Over the past decade, the UN General Assembly has passed a series of resolutions condemning criminal activity and drawing attention to defensive measures that governments can take. For more than a decade, Russia has sought a treaty for broader international oversight of the Internet, banning deception or the embedding of malicious code or circuitry that could be activated in the event of war. But Americans have argued that measures banning offense can damage defense against current attacks, and would be impossible to verify or enforce. Moreover, the United States has resisted agreements that could legitimize authoritarian governments' censorship of the internet. Nonetheless, the United States has begun informal discussions with Russia. Even advocates for an international law for information operations are skeptical of a multilateral treaty akin to the Geneva Conventions that could contain precise and detailed rules given future technological volatility, but they argue that like minded states could announce self governing rules that could form norms for the future". J.S. Nye, *Cyber Power*, Belfer Center for Science and International Affairs, Harvard Kennedy School, May 2010, p. 18.

play a destabilizing impact on the traditional Westphalian international order, this is especially true in cyberspace, where it is common for David to defeat Goliath.[10] Non-state actors extensively profit from the relative impunity that characterize cyberspace, its low barriers to entry[11] and the relatively easy task of finding vulnerabilities in ICT networks.[12] The use of cyber weapons by terrorists, for instance, is a likely – and extremely upsetting – development, especially considering how easy is in the dark web to acquire the knowledge necessary to attack enemies' networks, or even ready-to-use cyber weapons. Moreover, transnational cybercrime organizations are very relevant actors in cyberspace, as they are among the main investors in research and development of always-new offensive capabilities, and they therefore actively contribute to the international cyber arms proliferation. Cybercrime syndicates are also difficult to eradicate because of their economic power and because dismantling physical assets does not solve the problem, as malicious actors may access the Internet from anywhere in the world. Furthermore, police and judicial cooperation is complicated by the difficulty in finding the culprits of attacks

---

[10] "Today, cyberspace offers state and non-state actors the ability to wage campaigns against American political, economic, and security interests without ever physically crossing our borders. Cyber attacks offer adversaries low-cost and deniable opportunities to seriously damage or disrupt critical infrastructure, cripple American businesses, weaken our Federal networks, and attack the tools and devices that Americans use every day to communicate and conduct business", White House, *National Security Strategy of the United States*, December 2017, p. 12.

[11] "[B]arriers to entry in the cyber domain are so low that non-state actors and small states can play significant roles at low levels of cost", J.S. Nye (2010), p. 15.

[12] "Efforts to deter state and non-state actors alike are also hindered by the fact that, despite significant public and private investments in cybersecurity, finding and exploiting cyber vulnerabilities remains relatively easy. Those defending networks must be near perfect in their efforts, while malicious cyber actors may only need to find a single vulnerability to gain a foothold in a network", "Recommendations to the President on Deterring Adversaries and Better Protecting the American People from Cyber Threats", U.S. Department of State, Office of the Coordinator for Cyber Issues, 31 May 2018, p. 2.

(especially since this would typically involve sharing intelligence sources and findings), and criminals are known to be willing to act on behalf of states seeking plausible deniability through non-sovereign proxies.[13] Terrorists and criminals are probably the most dangerous actors in a highly diverse domain: hackers and the cyber underground, hacktivists, companies and private online individuals may all contribute to making security volatile in cyberspace as they pursue their multiple military, political and financial interests.

In this scenario of ambiguity and uncertainty, Great Powers are actively engaged in attaining "cyber superiority", defined as the "degree of dominance in cyberspace by one force that permits the secure, reliable conduct of operations by that force, and its related land, air, maritime, and space forces at a given time and place without prohibitive interference by an adversary".[14] Cyber superiority allows "maneuvering seamlessly between defense and offense across the interconnected battlespace", "globally, as close as possible to adversaries and their operations", "continuously, shaping the battlespace", in order "to create operational advantage for us while denying the same to our adversaries". Cyber superiority is also vital in mapping the theatre of future conflicts, in anticipating the adversary's vulnerabilities and in contesting its courses of action, and in establishing a deterrence posture – which is particularly complex to establish in cyberspace, as actionable attribution, and therefore retaliation, are problematic.[15] Cyber superiority is key to enhancing situational awareness and attribution, allowing countries under attack to impose swift, costly and transparent

---

[13] T. Maurer, *Cyber Mercenaries. The State, Hackers, and Power*, Cambridge, Cambridge University Press, 2018.

[14] *Achieve and Maintain Cyberspace Superiority*…, cit., p. 6.

[15] See, i.e.: M.C. Libicki, "Would Deterrence in Cyberspace Work Even with Attribution?", *Georgetown Journal of International Affairs*, 22 April 2015; F.D. Kramer, R.J. Butler, and C. Lotrionte, Cyber and Deterrence. The Military-Civil Nexus in High-End Conflict, Atlantic Council, 2017; M.P. Fischerkeller and R.J. Harknett, *Deterrence is Not a Credible Strategy for Cyberspace*, Foreign Policy Research Institute, Summer 2017.

consequences in response to malicious behavior.[16] And, of course, cyber superiority implies a continuous technological innovation capability across the doctrine, organization, training, materiel, leadership and education, personnel, and facilities (DOTMLPF) spectrum. If the new U.S. (and, hopefully, Western) posture will succeed in enhancing predictability in cyberspace, the international community might then – hopefully – be facilitated in agreeing on constraining rules of behavior, and in enhancing international cooperation against non-state malicious actors.

Security "in and around" cyberspace will most likely remain volatile in the years to come, given the conflicting strategic national interests and the diverging cultural and ideological approaches at play. The confrontation between the West on the one hand, and Russia, North Korea, China and Iran on the other (and, really, of everyone against everyone else) impacts international stability in profound ways.[17] We live in an age of latent conflict in cyberspace, and this leads to a classic international security paradox: on a systemic level, each player's

---

[16] "All instruments of national power are available to prevent, respond to, and deter malicious cyber activity against the United States. This includes diplomatic, information, military (both kinetic and cyber), financial, intelligence, public attribution, and law enforcement capabilities. The United States will formalize and make routine how we work with like-minded partners to attribute and deter malicious cyber activities with integrated strategies that impose swift, costly, and transparent consequences when malicious actors harm the United States or our partners". *National Cyber Strategy of the United States of America*…, cit., p. 21. Commenting the new *National Cyber Strategy*, Christopher Painter, Commissioner on the Global Commission for the Stability of Cyberspace and formerly the top cyber diplomat at the U.S. Department of State wrote: "While we're getting better at naming and shaming some of those responsible for cyber events, that's not sufficient to deter actors like Russia or North Korea. Real consequences for bad state behaviour that will affect their decision making is still desperately lacking. That creates the 'norm' that such bad behaviour is acceptable – or at least cost free". Christopher Painter, "The White House cyber strategy: words must be backed by action", *The Strategist*, Australian Strategic Policy Institute, 25 September 2018.

[17] F. Rugge, *Confronting an "Axis of Cyber"? China, Iran, North Korea, Russia in Cyberspace*, Milano, Ledizioni-ISPI, 2018.

individual quest for greater security actually translates into a more unpredictable and volatile security environment for all. In fact, one of the main features of cyberspace is the fact that offensive and defensive capabilities develop "hand in hand": it is impossible to ensure the appropriate defense of national ICT networks without knowing how an attack is executed and without developing a certain degree of cyber superiority at least on your own networks. Moreover, as cyber incidents typically do not allow time to react, mapping the battlefield before full-scale hostilities erupt is an operational imperative, but this implies conducting intelligence, surveillance and reconnaissance (ISR) operations against the networks of potential enemies – operations that, in turn, may easily be perceived as military in character (and, in fact, they may very well be). Because cyber arsenals are necessarily secret (as they rely on ICT vulnerabilities – zero-days – to be effective), states are resorting to demonstrative (but clandestine and deniable) actions to message their offensive capabilities, with the objective of deterring potential enemies by signaling their readiness to respond "in kind" to an attack.[18] How else to read, for instance, the malware that has been found in critical infrastructure around the world, other than weapons designed and planted to indicate readiness to strike in case of full-scale hostilities? Cyber weapons also risk becoming obsolete once zero-day vulnerabilities are patched, generating upon the developer/owner a classic "use it or lose it" dilemma. Cyber weapons, moreover, may be reverse-engineered, with the risk of proliferation of more and more cyber offensive capabilities among many state and non-state actors alike. As a result, the security paradox becomes more relevant every day.[19]

---

[18] "The President already has a wide variety of cyber and non-cyber options for deterring and responding to cyber activities that constitute a use of force. Credibly demonstrating that the United States is capable of imposing significant costs on those who carry out such activities is indispensable to maintaining and strengthening deterrence". *Recommendations to the President on Deterring Adversaries and Better Protecting the American People from Cyber Threats*, U.S. Department of State, Office of the Coordinator for Cyber Issues, 31 May 2018, p. 2.

[19] M.C. Libicki, "The Strategic Uses of Ambiguity in Cyberspace", *Military and*

The most immediate risk deriving from these developments is probably that the entanglement and the growing complexity of cyber interdependencies multiply the occasions for cross-domain escalations.[20] As assessing each other's relative strength is cumbersome in cyberspace, differences in threat perception and

---

Strategic Affairs, no. 3, 2011. See also: "Ambiguity is the cyber domain is such that it is also disputable whether a Balance of Power in cyberspace can be assessed and maintained at all: 'The question is: is a Balance of Power possible in the cyber age?'", U. Gori, "The Balance of Power in Cyberspace", in F. Rugge, …, cit., p. 143; and also: "This now-persistent engagement in cyberspace is already altering the strategic balance of power", *National Cyber Strategy of the United States of America*…, cit., p. 20.

[20] "The classic example of Cold War signaling has a Soviet missile submarine move closer to the United States (this meant a shorter flight time for a missile and less warning time, which reduced stability by increasing the chance of a surprise attack). In response, the United States might visibly move bombers to a higher readiness state. Soviet reconnaissance satellites would detect this change in status, and the submarine would draw away from the coast. This kind of signaling will be difficult in cybersecurity. What would moving to a higher state of alert entail? […] A review of documents from Soviet archives made available after the Cold War shows that the deterrent message the United States thought it was sending was often not the message the Soviets received. The possibility of miscommunication exists today. Potential opponents may misinterpret signals as expressions of hostile intent, or they may discount them. The risk of misinterpretation is high". J.A. Lewis, *Conflict and Negotiation in Cyberspace*, Center for Strategic and International Studies, February 2013, p. 49. See also: "The ability to send that message requires four things: attribution (the state must be able to define the target of retaliation), thresholds (the state must be able to consistently distinguish between acts that merit retaliation and those that do not), credibility (the state's will to retaliate must be believed), and capability (the state must be able to pull off a successful response). Each of these components is exponentially more complex in cyberspace than in a conventional setting", S. Hennessey, "Deterring Cyberattacks. How to Reduce Vulnerability", *Foreign Affairs*, November/December 2017; and also J.S. Nye, "Can Cyber Warfare Be Deterred?", *Project Syndicate*, 10 December 2015. Nye has developed these ideas in his essay, "Deterrence and Dissuasion in Cyberspace", *International Security* 41, no. 3, Winter 2016-17, pp. 58-60. See also: G. Perkovich and A.E. Levite (eds.), *Understanding Cyber Conflict. 14 Analogies*, Georgetown University Press, 2017, p. 170. See also: "Entanglement refers to the existence of various interdependences that make a successful attack simultaneously impose serious costs on the attacker as well as the victim", J.S. Nye (2016-17), p. 58.

different escalation ladders could also exacerbate the risk of an unintended escalation.[21] A crisis in cyberspace could therefore escalate to full-scale hostilities into the conventional domain and even, in an admittedly unlikely but not impossible scenario, may represent a threat to nuclear strategic stability.[22] A cyber attack at the outset of military operations could significantly degrade the adversary's situational awareness, blinding their early-warning system and Command & Control capabilities (including NC2) and might have the potential to blind spaced-based early-warning systems or to disable both command and control centers and decision-making processes. This is a result of the increased physical and logical interconnectedness and mutual dependency of nuclear and non-nuclear systems:[23] an

---

[21] "All instruments of national power are available to prevent, respond to, and deter malicious cyber activity against the United States. This includes diplomatic, information, military (both kinetic and cyber), financial, intelligence, public attribution, and law enforcement capabilities". *National Cyber Strategy*…, cit., p. 21. See also: *Conventional Prompt Global Strike and Long-Range Ballistic Missiles: Background and Issues*, Congressional Research Service, updated 14 August 2019, p. 17.

[22] "The emergence of offensive cyber warfare capabilities has created new challenges and potential vulnerabilities for the NC3 system. Potential adversaries are expending considerable effort to design and use cyber weapons against networked systems. While our NC3 system today remains assured and effective, we are taking steps to address challenges to network defense, authentication, data integrity, and secure, assured, and reliable information flow across a resilient NC3 network". "The United States would only consider the employment of nuclear weapons in extreme circumstances to defend the vital interests of the United States, its allies, and partners. Extreme circumstances could include significant non-nuclear strategic attacks. Significant non-nuclear strategic attacks include, but are not limited to, attacks on the U.S. or allied nuclear forces, their command and control, or warning and attack assessment capabilities". *US Nuclear Posture Review*, Office of the Secretary of Defense, 2018, p. 57 and 21.

[23] "Nuclear Weapons in a New Geopolitical Reality. An Urgent Need For New Arms Control Initiatives", *Adviesraad Internationale Vraagstukken*, no. 109, January 2019, pp. 40-41. See also: "Entanglement has various dimensions: dual-use delivery systems that can be armed with nuclear and non-nuclear warheads; the commingling of nuclear and non-nuclear forces and their support structures; and non-nuclear threats to nuclear weapons and their associated command,

attack against Intelligence, Surveillance and Reconnaissance (ISR) and nuclear early-warning systems could exacerbate the risk of a nuclear overreaction, because it would complicate the task of assessing an attacker's intent, and could impair the effectiveness of the nuclear retaliatory capability. These developments would on the one hand favor the adoption of hair-trigger states of readiness and of lower level of decision making, while on the other they would contribute to make a disarming cyber strike a viable option.[24] At the same time,

---

control, communication, and information (C3I) systems. Technological developments are currently increasing the entanglement of non-nuclear weapons with nuclear weapons and their enabling capabilities". J.M. Acton (ed.), *Entanglement. Russian and Chinese Perspectives on Non-Nuclear Weapons and Nuclear Risks*, Carnegie Endowment for International Peace, 2017. See also: "[…] increasingly, these nuclear command-and-control systems are also being used to support non-nuclear operations. The U.S., for example, operates satellites to provide warning of attacks with nuclear-armed or conventionally armed ballistic missiles. In a conflict between NATO and Russia, these could be used to detect short-range conventional ballistic missiles launched by Russia - as the first step towards shooting them down. If this strategy was successful, Russia could decide to attack the U.S. early-warning satellites in response. In fact, the U.S. intelligence community has warned that Russia is developing ground-based laser weapons for that exact purpose. But blinding U.S. early-warning satellites would not simply undermine its ability to spot conventionally armed missiles. It would also compromise the ability of the U.S. to detect nuclear-armed ballistic missiles and could raise fears that Russia was planning a nuclear attack on the U.S.", J.M. Acton, *The Weapons Making Nuclear War More Likely*, Carnegie Endowment For International Peace, 8 February 2019; J.M. Acton, "Escalation through Entanglement. How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War", *International Security*, vol. 43, no. 1, Summer 2018, p. 97.

[24] "A way around this is to conceptualise the cyber challenge into: (i) a new set of capabilities that might be used and vulnerabilities that might be exploited within the computer systems and networks used across the nuclear weapons enterprise; and (ii) the broader context and environment within which nuclear policy is carried out. The former is about malware, cyber-attacks, bugs, and hacking, while the latter is about the digitised information space that all states operate in. There is even a case to be made that we should stop using the word cyber altogether, and instead revert back to the more precise language of Computer Network Attacks, Computer Network Defence, Computer/ Network/Information Security, etc.

in order to strengthen deterrence against malicious cyber campaigns, many states are considering the option of imposing "strategic dilemmas" by retaliating in different domains,[25] making cross-domain escalation the intended outcome.

The ongoing conflict in cyberspace and the need to mobilize to maintain cyber superiority are at the origin of the ongoing global decoupling of the hardware and software ICT supply chains, and they are also provoking the gradual building of barriers to technology transfer and the proliferation of national safeguards against the foreign acquisition of technological products, services and companies, resulting in a global normative patchwork. This is not just a Western problem, as in the case of China's 5G technology; Beijing, for instance, recently decided to replace all the hardware and software used by public bodies with domestically-produced technology. The decoupling of the global ICT supply chain responds to the increasing recourse to insidious attacks to the integrity of hardware and software, and to the associated threat of espionage by foreign entities (be they governmental or private companies subject to a strong government's direction) targeting governments' confidential information, commercial or industrial secrets through attacks on the ICT supply chain, or seeking to profile users in order

More precision in terminology is undoubtedly the first step towards constructing meaningful and tailored measures to deal with specific cyber challenges in the nuclear realm", A. Futter, *Managing the cyber-Nuclear Nexus*, European Leadership Network, July 2019. See also: "The cyber threat affects nuclear risks in at least two ways: It can be used to undermine the security of nuclear materials and facility operations, and it can compromise nuclear command and control systems", Addressing Cyber-Nuclear Security Threats, NTI (Nuclear Threat Initiative). Building a Safer World.

[25] See: "The focus here is principally on conventional military operations. Cyber, counterspace, financial, information, and other tools should be profitably analyzed in the context of asymmetric deterrence and escalation management. However, they would likely be employed in any response to Russian aggression and do not fit comfortably in the frame- work of horizontal escalation", M. Fitzsimmons, "Horizontal Escalation: An Asymmetric Approach to Russian Aggression?", *Strategic Studies Quarterly*, Spring 2019.

to subsequently target them to influence their perceptions and public opinions. The decoupling of the ICT supply chain is also intended to prevent dependence on foreign producers and service providers who might use their control over hardware, software and networks to exert political and economic pressure and to acquire a military advantage. This could be the case, for instance, if its operators denied an essential service to a critical national infrastructure, manipulated data or diverted data flows, sabotaged essential democratic or industrial processes, or hampered political decision-making on issues of national security and defence.

In this competition between Great Powers, even the global Internet is segmented by different, interconnected – but, if necessary, independent – systems: China erected its "Great Firewall" and Russian networks can now, by law, be segregated in case of need. These developments are the result of a competition between opposing blocks, and they simultaneously intensify that same competition. Cyber-enabled information warfare already appears to be one of the instruments of choice in the ongoing international confrontation: hostile actors in cyberspace are willing and able of leveraging the panoply of tools allowed by computer network operations (CNOs) and "computational propaganda" to influence public opinion to a degree that old-fashioned PsyOps could only dream of. These cyber tools have a much greater impact on the target audiences, for instance by creating a virtually infinite number of automated scripts (bots) to populate social media and interact with unwitting real users online; using social engineering techniques for targeting purposes; rerouting data-flows or launching distributed denial of service attacks (DDoS) in order to interdict information; attacking the hardware supply chain; and infiltrating the opponent's networks to steal, modify, implant or expose privileged information. The idea is that by manipulating content on ICT networks and social media it is possible to deceive, distract and disinform public opinion, muddying the waters with diverging truths, eventually disorienting and

seeding doubt among the public, or shaping the opinion of a specific target audience on a given issue. It is perhaps around the idea of an Internet which is free, uncensored and global that deepest fault lines will emerge in the ongoing Great Powers Competition: while for the one side "freedom of the Internet" is an ideologically necessary condition for enjoying fundamental rights of information, expression and association in the XXI century, for "the other side" it represents an existential threat to its political stability and security.

These developments fuel a creeping mutual distrust within the International Community, which in turn will increasingly hamper international cooperation at all levels against the wide range of traditional and emerging security challenges and security threats that would require a multilateral, coordinated approach, such as climate change, food and water scarcity, terrorism, mass migrations, health security, etc. The net effect of these developments in cyberspace, in other words, is that of reinforcing the widespread perception that the safest way ahead is centered upon strong sovereign states.[26] In a way, this is paradoxical: cyberspace was born to connect people across the globe with complete disregard to national frontiers and governments. The example of the original development of the Internet, and the governance structure currently sustaining its everyday functioning, are both good examples of how, in abstract, states are *not necessary* to create and sustain cyberspace – they are in fact, to a certain extent, "special guests" of cyberspace.[27] Instead, the cyber domain has become one of the

---

[26] "This strategy is guided by principled realism. It is realist because it acknowledges the central role of power in international politics, affirms that sovereign states are the best hope for a peaceful world, and clearly defines our national interests. It is principled because it is grounded in the knowledge that advancing American principles spreads peace and prosperity around the globe. We are guided by our values and disciplined by our interests", *National Security Strategy of the United States of America...*, cit., p. 55.

[27] "Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. On behalf of the future, I ask you of the past to leave us alone. You are not welcome among us. You have no

most destabilizing areas of competition between states, and its development is contributing to the crisis of multilateralism. For liberal democracies, this implies an additional burden *vis-à-vis* autocratic states: that of preserving the technological edge that has historically been associated with their hegemony on the international system and that is necessary to maintain adequate deterrence in cyberspace while upholding, in the process, the principles of freedom of the Internet and the human-centric values that distinguish us from autocratic regimes.

While we never abandoned a nuclear security paradigm that postulates that "the only way to win is not to play", in the cyber domain we are drifting toward one (or, rather: we now need to make it coexist with one) where – so we are told – "the only way not to lose is to persistently engage the adversaries".[28] We all subscribe to the goal of enhancing predictability in cyberspace, and, in the absence of clear and actionable international law regulating the behavior of states in cyberspace, it might in fact very well be that persistent operational engagement with adversaries is the only way to enhance deterrence in this "domain of ambiguity". However, the militarization of cyberspace, the inherent difficulty in distinguishing between the intelligence and the military nature of a campaign, the widespread recourse

---

sovereignty where we gather. We have no elected government, nor are we likely to have one, so I address you with no greater authority than that with which liberty itself always speaks. I declare the global social space we are building to be naturally independent of the tyrannies you seek to impose on us. You have no moral right to rule us nor do you possess any methods of enforcement we have true reason to fear", J. Perry Barlow, *A Declaration of the Independence of Cyberspace*, Electronic Frontier Foundation, Davos, Switzerland, 8 February 1996.

[28] "Superiority through persistence seizes and maintains the initiative in cyberspace by continuously engaging and contesting adversaries and causing them uncertainty wherever they maneuver. It describes *how* we operate – maneuvering seamlessly between defense and offense across the interconnected battlespace. It describes *where* we operate – globally, as close as possible to adversaries and their operations. It describes *when* we operate – continuously, shaping the battlespace. It describes *why* we operate – to create operational advantage for us while denying the same to our adversaries", *Achieve and Maintain Cyberspace Superiority*…, cit., p. 6.

to CEIW, the intrinsic secrecy of cyber arsenals and the massive security paradox resulting from the legitimate national quests for cyber superiority, are all developments that undermine trust within the international community and threaten international stability, increasing the risk of misinterpretations, miscalculations and unintended escalation to the conventional domain becoming ever more real.[29]

Tech Supremacy In and Through Cyberspace

Our "increasingly complex security environment is defined by rapid technological change"[30] and technological superiority has become one of the defining paradigms of the current Great Power Competition. Strictly speaking, cyberspace does not qualify as "a new technology", in part because it emerged some 40 years ago, but mostly because it is much more than a technology: it is a man-made domain adding an extra layer of reality to our everyday life, a political space, a hypostatic abstraction. Cyberspace is ubiquitous: it is the nervous system that connects the political, military, informative, economic, financial, industrial and infrastructural dimensions on a personal, local, national, international and transnational level. The cyber domain has become, and will most likely increasingly be, an environment characterized by an "unthinkable

---

[29] "Cyber capabilities, particularly the emergence of offensive weapons, are reshaping the way policymakers in the United States think about thresholds for using force - whether provocations or attacks in cyberspace warrant a response in cyberspace or in other domains", Game Changer, cit., p. 20. See also: "With uncertain rules, there remains considerable potential for escalation if a conflict between two States emerges. It is also in this light that the current reluctance of States to call cyber activities such as economic espionage violations of international law can perhaps be understood", K. Ziolkowski (ed.), *Peacetime Regime for State Activities in Cyberspace. International Law, International Relations and Diplomacy*, Tallinn, NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) Publications, 2013, p.216. See also: J. Healey, "Triggering the New Forever War, in Cyberspace", *The Cipher Brief*, 1 April 2018.

[30] *2018 US National Defense Strategy*, p. 3.

complexity",[31] where a multitude of diverse players constantly connect worldwide generating "an inescapable network of mutuality".[32] The advent of cyberspace was a game changer whose cultural, political and strategic disruptive implications we only partly understand.[33]

Even if cyberspace is much more than just a technology, technological innovation is key in attaining superiority in the cyber domain, and AI promises to be the technology that will allow a leap into a new generation of cyber capabilities since, as Mariarosaria Taddeo and Luciano Floridi put it, "cyberspace is a domain of warfare, and AI is a new defense capability".[34] AI might very well be one of the most important global issues of this century[35]; the political priority given to AI

---

[31] William Gibson, in *Neuromancer*, uses in 1984 for the first time the term "cyberspace", and defines it as follow: "Cyberspace. A *consensual hallucination* experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts. [...] A graphic representation of data abstracted from the banks of every computer in the human system. *Unthinkable complexity.* Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding", italic mine

[32] In a speech delivered in Alabama in 1963, Martin Luther King affirmed "Injustice anywhere is a threat to justice everywhere. We are caught in an inescapable network of mutuality, tied in a single garment of destiny. Whatever affects one directly, affects all indirectly". I believe that this statement, which embodies the highest moral authority of the U.S. civil rights' movement, perfectly describes one of the most critical challenge of our generation: that of ensuring a secure and just order in cyberspace. In a speech delivered in Alabama in 1963, Martin Luther King affirmed "Injustice anywhere is a threat to justice everywhere. We are caught in an inescapable network of mutuality, tied in a single garment of destiny. Whatever affects one directly, affects all indirectly". I believe that this statement, which embodies the highest moral authority of the U.S. civil rights' movement, perfectly describes one of the most critical challenge of our generation: that of ensuring a secure and just order in cyberspace.

[33] F. Rugge (2018).

[34] M. Taddeo and L. Floridi, "Regulate artificial intelligence to avert cyber arms race", *Nature Machine Intelligence*, April 2018.

[35] For an extensive account of AI's implications in world affairs, see: D.M. West and J.R. Allen, *How artificial intelligence is transforming the world*, The Brookings Institution, 24 April 2018.

and the investments already being made in the United States, China and Russia are unequivocal signs that an arms race for primacy in AI applications is indeed ongoing, as confirmed by the Russian President Putin, who famously declared that "[w] hoever becomes the leader in this sphere will become the ruler of the world".[36] Exponential growth in computer-processing power, in fact, enabled the development of machine-learning techniques, which may perpetually and autonomously train on the "big data" produced thanks to the ongoing revolution in the field of new sensors and the Internet of Things. Cloud technology, finally, allows computers to readily access, from a variety of sources and with increasing velocity, off-board processing and data resourcing to solve problems.[37] Hence, autonomous systems become pervasive and capable, while we all become increasingly reliant on AI-enabled connected systems, with the effect of opening up new avenues for attacks to the confidentiality, the integrity and the availability of our data from actors seeking to steal privileged information, hamper decision-making processes, disrupt order, instill a sense of fear in our population or manipulate information to discredit media outlets and influence the public debate.

Progress made by China in the fields of AI,[38] in particular, is

---

[36] A. Polyakova, *Weapons of the weak: Russia and AI-driven asymmetric warfare*, The Brooking Institutions, 15 November 2018. See also: "National competition and the perception of "AI race" dynamics may have a negative impact on diplomatic efforts. Moreover, if one country develops significant advances in AI technologies, the country may gain access to economic and political advantages and not have a natural incentive to share its capabilities or resources in the absence of pre-existing international agreements", J. Cussins Newman, *Toward AI Security. Global Aspirations for a More Resilient Future*, Center for Long-term Cybersecurity, UC Berkley, February 2019, p. 23.

[37] *Algorithmic Warfare Applying Artificial Intelligence to Warfighting*, Griffith University, 2018, p. 5.

[38] G.C. Allen, *Understanding China's AI Strategy*, Center for a New American Security, 6 February 2019. See also: "The PLA will likely leverage AI to enhance its future capabilities, including in intelligent and autonomous unmanned systems; AI-enabled data fusion, information processing, and intelligence analysis; war-gaming, simulation, and training; defense, offense, and command

a cause for concern in the West, first and foremost because it is an eloquent signal of Beijing's growing power and technological edge, and secondly because it represents a powerful military deterrent at a time when China is becoming increasing assertive on the world stage. For Beijing, this race for technological superiority is one of the tools for disrupting the "unipolar" liberal democratic global order. Autocratic regimes are today more capable and determined to challenge long-lasting Western technological superiority, leveraging in particular their greater control of the private sector and their longer-term planning capability for achieving this strategic objective. China has ambitious, yet realistic goals: affirming itself as the "premier global AI innovation center" by 2030, possibly surpassing the United States in the process[39], completing military modernization by 2035, and becoming a "world-class" military by 2049.[40] "China's leadership – including President Xi

---

in information warfare; and intelligent support to command decision-making. At present, the PLA is funding a wide range of projects involving AI, and the Chinese defense industry and PLA research institutes are pursuing extensive research and development, in some cases partnering with private enterprises. This could be the start of a major shift in the PLA's strategic approach, beyond its traditional asymmetric focus on targeting U.S. vulnerabilities to the offset-oriented pursuit of competition to innovate. The PLA is seeking to engage in 'leapfrog development' (跨越 发展) to achieve a decisive edge in 'strategic front-line' (战略前沿) technologies, in which the United States has not realized and may not be able to achieve a decisive advantage", E.B. Kania, *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*, November 2017 , p. 37 and p. 4

[39] *State Council Notice on the Issuance of the New Generation AI Development Plan* [国务院关于印发新一 代人工智能发展规划的通知], State Council, 20 July 2017. See also: "China aspires to surpass the United States in AI. The Chinese leadership recognizes and intends to take advantage of AI to enhance its economic competitiveness and military capabilities For instance, according to a recent report from PriceWaterhouseCoopers, China is expected to be one of the greatest beneficiaries of the economic contributions of AI, given an expected 26% boost to its GDP by 2030". E.B. Kania (2017), p. 8. See also: M.E. O'Hanlon, *The role of AI in future warfare*, The Brooking Institution, 29 November 2018.

[40] "By 2035, China's military leaders seek to complete military modernization and by 2049, they have characterized their goal as becoming a "world-class" military.

Jinping – believes that being at the forefront in AI technology is critical to the future of global military and economic power competition, and that China should pursue global leadership in AI technology and reduce its vulnerable dependence on imports of international technology".[41] AI will introduce a whole new generation of threats[42], transforming "the character of conflict beyond information-age warfare toward "algorithmic warfare", in the U.S. military's phrasing, or "intelligentized" warfare, as Chinese military thinkers characterize it".[43]

In this regard, China's efforts are designed with a clear purpose in mind: to displace the United States in the Indo-Pacific region; to expand the reaches of its state-driven economic model; and to reorder the region in its favor", C. Larson, "China's massive investment in artificial intelligence has an insidious downside", *Science*, 8 February 2018. See also: "A 'World-Class' Military: Assessing China's Global Military Ambitions", Testimony before the U.S.-China Economic and Security Review Commission, Office of the Secretary of Defense, Office of the Assistant Secretary of Defense for Indo-Pacific Security Affairs Acting Deputy Assistant Secretary of Defense for East Asia Mary Beth Morgan, 20 June 2019.
[41] "Information technology, including computers and telecommunications systems, has permeated all aspects of society and economies and become an integral part of a nation's infrastructure. Chinese analysts have dubbed this process "informationisation (*xinxihua*, 信息化)". From the Chinese perspective "Informationisation is a comprehensive system of systems, where the broad use of information technology is the guide, where in- formation resources are the core, where information networks are the foundation, where information industry is the support, where information talent is a key factor, where laws, policies, and standards are the safeguard". In the face of this broad trend of economic, political, and social informationisation, Chinese analysts have concluded that threats to national interests and security have also become informationised". D. Cheng, "China and Cyber: The Growing Role of Information in Chinese Thinking", in F. Rugge (ed.), *Confronting an "Axis of Cyber"? China, Iran, North Korea, Russia in Cyberspace*, Milano, Ledizioni-ISPI, 2018, pp. 59-60.
[42] *Reinventing Cybersecurity with Artificial Intelligence: The new frontier in digital security*, Capgemini Institute, 2019.
[43] J.R. Allen and H. Husain, *On Hyperwar*, U.S. Naval Institute, 2017. See also: E. Kania, "Great Power Competition and the AI Revolution: A Range of Risks to Military and Strategic Stability", *Lawfare*, 19 September 2017. See also: Department of Defense Enterprise Cloud and its Importance to the Warfighter Media Roundtable, Department of Defense Director, Joint Artificial Intelligence Center Lieut. Gen. Jack Shanahan, 9 August 2019. On *Algorithmic*

It is hard to anticipate how developments in AI will affect the international order, the character of war and the Balance of Power, because many variables will determine the possible outcomes, which will also reflect the unexpected emergent behaviors that might arise from the combined employment of new technologies, tactics, techniques and procedures.[44]

Warfare: "American advances in algorithmic warfare and the associated Third Offset thinking have stimulated strong Chinese and Russian interest. China has become a 'fast follower' and is implementing an ambitious new national strategy to become the world leader in intelligent machine technology, at least initially to gain an economic edge. In the military domain, the People's Liberation Army (PLA) now considers the application of intelligent machine technology will fundamentally change the character of war. 'Intelligentized' warfare will replace today's network-centric warfare, and is accordingly imperative to embrace. […] PLA strategic thinkers anticipate today's 'informatized' warfare will progressively give way to tomorrow's 'intelligentized' warfare. In introducing intelligent machine technologies to warfighting, the character of warfare will transform. The post-information warfare era is beginning. In part, this belief rests on the Marxian-derived notion that contemporary ways of war reflect the economic approach of the time. The industrial age brought large-scale mechanised warfare, the information age network-centric warfare, the intelligent machine age will similarly bring a new approach.[…] Intelligent algorithms play a crucial role in firstly determining through analysing big data who is specifically useful to target, and secondly in progressively optimising ongoing 'attacks' against those identified over extended time periods. The logic of the strategy is to gradually reinforce particular individuals' existing opinions in a way that makes them more extreme, but not to dramatically alter their views. Intelligent machine algorithms for the first time allow warfare to be individualised", P. Layton (2018), pp. 47, 49 and 57. See also: "[…] algorithmic warfare can only be enabled by (a) working systems (minimally viable) capable of (b) learning on their own from unknown and unknowable scenarios (unsupervised) while (c) converting a complex battlefield environment into a useful insight (deep-learning enabled) (d) with little to no guidance (autonomous) and (e) in a live mission environment (battlefield ready)", C. Crosby, "Operationalizing Artificial Intelligence for Algorithmic Warfare", *Military Review*, July-August 2020.

[44] "The pace and complexity of technological change mean that linear predictions of current trends cannot be the basis for effective guidance or management for the future", J. Kadtke and L. Wells II, *Policy Challenges of Accelerating Technological Change: Security Policy and Strategy Implications of Parallel Scientific Revolutions*, Center for Technology and National Security Policy (CTNSP) National Defense University (NDU), September 2014. See also: "What role will artificial intelligence play? In

Moreover, long-term shifts in the Balance of Power determined by advances in AI may be indirect, as an effect of the economic power brought about its many possible applications. If AI will be mostly commercially-driven, for instance, disruptive advances are likely to spread more rapidly to militaries around the world, hence reducing asymmetry and surprise, while if technological innovation will be based on defense research, early adopters might benefit from a first-mover advantage.[45] Conversely, commercial interests might hold back military development, as the private sector regularly pays much higher salaries and may drain from the public sector the human resources available (the true strategic resource for AI, just like carbon in the first industrial revolution). The private sector, moreover, might have different views on the prospects for collaboration with the military complex. What will for instance be the role of the future Googles and Apples in enabling states' military power, or in providing technological enablers to non-state actors?[46]

---

many ways it is too soon to tell, given uncertainty about the development of the technology. But AI seems much more akin to the internal combustion engine or electricity than a weapon. It is an enabler, a general-purpose technology with a multitude of applications. That makes AI different from, and broader than, a missile, a submarine, or a tank", B. Garfinkel and A. Dafoe, "Artificial Intelligence, Foresight, and the Offense-Defense Balance", *War On The Rocks*, 19 December 2019, p. 39. See also: "the introduction of a new form of force — from the tank to the torpedo to the phishing attack – will often warrant the introduction of substantially new tactics. Since these tactics emerge at least in part through a process of trial and error, as both attackers and defenders learn from the experience of conflict, there is a limit to how much can ultimately be foreseen", ibid.

[45] "Moreover, if the computational power necessary to generate new, powerful algorithms prices out all but the wealthiest companies and countries, higher-end AI capabilities could help the rich get richer from a balance-of-power perspective. On the other hand, if leading militaries fail to effectively incorporate AI, the potential for disruption would also be larger", ibid.

[46] "Employees at Google and Microsoft have objected to their companies' contracts with the Pentagon, leading Google to discontinue work on a project using ai to analyze video footage. China's authoritarian regime doesn't permit this kind of open dissent. Its model of 'military-civil fusion' means that Chinese technology innovations will translate more easily into military gains", P. Scharre,

Will collaborative engagement techniques, enabled by swarm technologies and AI, empower disparate adversary groups to act in conjunction, inflicting massive damage and elevating the risks of urban warfare?[47] The truth is: we do not know what the future will bring - but it will certainly provide surprises that may trigger unintended escalations.

By the same token, wealthier economies might be able to invest more heavily in technological research and gain an initial advantage on which to build in order to maintain tech superiority, and drain from poorer countries the human capital needed for technological innovation and application.[48] On the other hand, automation reduces for the first time the relative importance of the size of the population as an enabler of military power, as robotics and swarm technologies will eventually depend on financial resources. AI-enabled automation and swarm technologies will also possibly contribute to reversing a historic trend that favored the fielding of more and more expensive and tech-intensive weapons-systems, making instead a whole new set of military tactics possible with the employment of relatively cheap swarms of many drones and autonomous weapons systems.[49] These same technologies, on the other hand,

---

"Killer Apps. The Real Dangers of an AI Arms Race", *Foreign Affairs*, May/June 2019.

[47] "In congressional testimony in October, Attorney General Jeff Sessions was pressed on whether the administration had done enough to prevent Russian interference in the future. 'Probably not', Sessions said. "And the matter is so complex that for most of us we are not able to fully grasp the technical dangers that are out there", G.Miller, G. Jaffe, and P. Rucker, "Doubting the intelligence, Trump pursues Putin and leaves a Russian threat unchecked", *The Washington Post*, 14 December 2017.

[48] "The larger the change within the organization required for a military to effectively utilize new technologies, the greater the bureaucratic challenges and, with them, the likelihood that powerful countries will not have the organizational capability to adopt. This is a key mechanism through which the balance of power can change", M.C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power", *Texas National Security Review*, vol. 1, no. 3, May 2018. p. 44.

[49] Curiously, intelligent machines may return mass to the battlefield. In recent

would give non-state actors a powerful and easily accessible tool of mass disruption, contributing to making the international security environment more unpredictable[50].

AI might, in a not too distant future, provide our adversaries with an overwhelming military advantage that they receive from being the first to field it ("technological surprise"[51]), making pre-emptive strategies a rational – although destabilizing – course of action. In the same vein, the fear of being caught by surprise provides incentives to field newly-developed responses to military requirements without a proper test run. This already happens every day in cyberspace, where cutting corners on safety is the norm in order to shorten time to market.[52] As we depend more and more on AI and weapons-systems

---

decades the trend in armed forces has been to develop force structures based around a relatively small number of highly effective, multi-role platforms. Intelligent machine technology may allow these highly sophisticated weapon systems to be complemented by a very large number of dramatically lower cost, unmanned systems optimised for specific tasks. The unmanned systems would be *in extremis* expendable and so could be risked in the more dangerous tasks that the few expensive manned platforms might not sensibly be.", P. Layton (2018), p. 33.

[50] Increasing use of drones and similar weaponry mean that autonomous weapons are likely to be accessible to non-state actors that are not bound by traditional laws of armed conflict, human rights in the age of artificial intelligence, "Human Rights in the Age of Artificial Intelligence", *AccessNow*, November 2018, p. 29.

[51] G.H. Heilmeier, "Guarding Against Technological Surprise", *Air University Review*, September-October 1976, and "Technological change does not have to be dramatic or sudden to create meaningful shifts in power balances or social structures. Indeed, focusing on the distant prospect of dramatic change may well distract from developing a more nuanced understanding of slower and subtler, but equally significant, changes", M.L. Cummings, H.M. Roff, K. Cukier, J. Parakilas, and H. Bryce, *Artificial Intelligence and International Affairs. Disruption Anticipated*, Chatham House Report, 14 June 2018.

[52] "For each country, the real danger is not that it will fall behind its competitors in AI but that the perception of a race will prompt everyone to rush to deploy unsafe AI systems. In their desire to win, countries risk endangering themselves just as much as their opponents. [...] Digital security is already too often an afterthought. A world of widespread, unprotected AI systems isn't just a possibility; it's the default setting", P. Scharre (2019), p. 135 and 143.

increasingly rely on AI-enabled automation, the risk increases that autonomous lethal weapons are employed without proper prior testing and without due consideration of the requirements posed by International Humanitarian Law on the conduct of military operations.[53] Conversely, the meticulous application of security-by-design principles might direct AI systems to implement by default assertive self-defense strategies, resulting, at the systemic level, in creeping escalations and brinkmanship. Finally, the asymmetric military advantage may not necessarily result from being the first to master a disruptive technology: since not all countries and non-state actors follow the same moral compass, the weaponization and first employment of a technology could be available, in practice, only to one side[54].

---

[53] "[…] the key feature of weapons that autonomously select and apply force is that the user will not know the exact target that will be struck, nor its location and surroundings, nor the timing and circumstances of the application of force. There are consequently significant difficulties in using AWS in a manner that retains the user's ability to reasonably foresee (predict) the effects of the weapon in the circumstances of use and to make the contextspecific valuebased judgments required by IHL rules", V. Boulanin, N. Davison, N. Goussac, and M. Peldán Carlsson, *Limits On Autonomy In Weapon Systems, Identifying Practical Elements of Human Control*, Stockholm International Peace Research Institute and ICRC, June 2020, p. 7. See also: "Autonomous weapons in the near future are likely to suffer from AI's inability to deal with nuance or unexpected events. In a conflict situation, this could result in the death or injury of innocent civilians that a human operator may have been able to avoid", "Human Rights in the Age of Artificial Intelligence"…, cit., p. 19 See also: "Reviewing the legality of new weapons before they are deployed is an obligation for States that are party to Additional Protocol I of the Geneva Conventions (*Article 36*)", N. Goussac, "Safety Net or Tangled Web: Legal reviews of AI in weapons and war-fighting", *Humanitarian Law & Policy Blog*, 18 April 2019. See also: "AI will inevitably introduce uncertainty into the functioning of a weapon – meaning that the reviewer cannot predict with a reasonable degree of certainty all the outcomes of using the weapon. This unpredictability can arise through the weapon's design or the interaction between the system", ibid.

[54] "The U.S. military could face a disadvantage or pressures to adapt if strategic competitors such as China and Russia pursue full autonomy without similar constraints – although it remains unclear when, whether, and in what contexts greater degrees of autonomy will provide a clear advantage", E.B. Kania (2017),

In sum, technological developments in the fields of Artificial Intelligence will most likely drift the international community towards a Balance of Power that is much more difficult to assess and to maintain; ambiguous or inaccurate perceptions of the security environment raise the risks of disastrous courses of action.[55]

Although it is impossible to predict the speed and the level of automation at which cyber offensive and defensive capabilities will evolve once AI will have been applied to cyber warfare,[56] AI-

---

p. 37. See also: J.R. Allen and H. Husain (2017).

[55] "Evaluation of any horizontal escalation option is subject to considerable uncertainty, especially regarding adversary perceptions, values, and escalation thresholds. Understanding how adversaries would perceive their own (much less their adversaries') stakes and risk tolerance and expected outcomes is inherently difficult. In Richard Smoke's classic ex- amination of escalation, his historical case studies show that escalation failures most often occur because of a fundamental failure on the part of policymakers to comprehend how the world looked to others and understand basic assumptions, goals, and options of decision makers in other capitals", M. Fitzsimmons (2019).

[56] "Just as AI will profoundly affect the speed of warfare, the proliferation of zero day or zero second cyber threats as well as polymorphic malware will challenge even the most sophisticated signature-based cyber protection. This forces significant improvement to existing cyber defenses. Increasingly, vulnerable systems are migrating, and will need to shift to a layered approach to cybersecurity with cloud- based, cognitive AI platforms. This approach moves the community toward a 'thinking' defensive capability that can defend networks through constant training on known threats. This capability includes DNA-level analysis of heretofore unknown code, with the possibility of recognizing and stopping inbound malicious code by recognizing a string component of the file. This is how certain key U.S.-based systems stopped the debilitating 'WannaCry' and 'Petya' viruses", D.M. West and J.R. Allen (2018). See also: "While on a closer look many of the disputes could be in fact reduced to practical, procedural or technical matters, some vital legal questions remain, among them (not exhaustively): autonomous cyber capabilities and the element of intent in prohibited intervention, an autonomous system's capability to assess the severity of an incoming attack, autonomous cyber capability and the duty to take feasible precautionary measures, autonomous cyber capabilities and *mens rea* and international liability schemes for damages caused by the use of an autonomous cyber capability", R. Liivoja, M. Naagel, and A. Väljataga, *Autonomous Cyber Capabilities under International Law*, Tallinn, NATO Cooperative

enabled cyber offensive capabilities have already been used for reconnaissance, access and penetration of target networks,[57] and they are employed daily for profiling and targeted advertising. In fact, AI may play a role in all 19 use cases for AI in the cyber attack anatomy, as identified in the table below.[58]

Reconnaissance	Access and penetration	Internal reconnaissance and lateral movements	Command, control, and actions on objectives	Exfiltration and sanitation
Strategic intelligence collection	Attack planning	Network and system mapping	Domain generation	Discovery obfuscation
Target profiling	Phishing and spear phishing	Network behavior analysis	Self-learning malware	"Low-and-slow" exfiltration
Vulnerability detection	Attack code generation	Smart lateral movements	Sward-based command and control	
Outcome prediction	Classifier manipulation		NLP manipulation	
	Password attacks			
	Captcha attacks			

AI will advance computer systems robustness, making them more resilient to cyber or algorithmic attacks; it will improve situational awareness and anomaly detection; and it will enable systems' autonomous defense, and possibly even self-directed

---

Cyber Defence Centre of Excellence (CCDCOE) Publications, 2019, p. 44. See also: "One challenge could be a more efficient form of advanced persistent threat in which efforts to penetrate an adversary's computer systems employ automated capabilities with massive raw computational power that continually adjust tactics to the defenses encountered", M.E. O'Hanlon, *The role of AI in future warfare*, The Brooking Institution, 29 November 2018.

[57] E. Zouave et al, *Artificially intelligent cyberattacks*, FOI, March 2020.

[58] Ibid., p. 17

retaliatory strikes.[59] Cyber defense strategies will increasingly leverage AI,[60] with the result of elevating cyber dueling into the new reality of algorithmic warfare, where, for instance, we might one day observe autonomous malware replicating and adapting to a specific scenario in order to inflict maximum damage to target networks.[61] Algorithmic warfare will

---

[59] "First, AI can improve a system's robustness, that is, the capacity of a system to keep behaving as expected even when it processes erroneous inputs, thanks to self-testing and self-healing software. Second, AI can advance a system's response, that is, the capacity of a system to defeat an attack autonomously, refine future strategies on the basis of the achieved success, and possibly launch more aggressive counter operations with each iteration. AI systems that support responses to attacks, generating decoys and honeypots for attackers, are already available on the market. Third, AI can increase a system's resilience, that is, the ability of a system to withstand attacks, by facilitating threat and anomaly detection (TAD) – data indicate that by 2022, AI will deal with 50% of TAD tasks – and supporting security analysts in retrieving information about cyber threats", M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword", *Nature Machine Intelligence*, 11 November 2019, p. 557.

[60] The scale and speed at which AI-powered cyberattacks can occur may increasingly pressure cybersecurity vendors to offer AI-powered cyber defenses. The work of cybersecurity professionals will still be needed for a long time to come, but people must also consider how to institute the right processes and communication channels to enhance and integrate their work with that of the AI defense systems", J. Cussins Newman (2019), p. 18.

[61] "The result is that these commercial companies now need to develop defensive algorithms to protect themselves and their customers against such exploitation in the future. A cyber battlespace of duelling algorithms is emerging", P. Layton (2018), p. 57-58. See also: "Terminator-style robots of limitless aggression, with inexhaustible energy supplies, and endless weapon stocks exist only in fiction. While constrained to the virtual world, AI-powered cyber weapons bear a worrying similarity to such imaginary robots. Future offensive cyber operations could employ intelligent machine viruses that might replicate continually, draw energy from their hosts and remain forever at war in the cyber domain", ibid., pp. 67-68. See also: "AI technologies can further be weaponized to increase the effectiveness of the malware, making it more autonomous, more sophisticated, faster, and harder to detect. With the support of AI, the new generation of malware becomes smarter and capable of operating autonomously. The intelligent malicious programs can self-propagate in a network or computer system based on a sequence of autonomous decisions, intelligently custom-made

increase the pace of battle leveraging machines' much greater capacity to analyze huge volumes of very fast data flows from multiple different sources and types (audio, video, text, etc.), determining patterns, associations and relationships that would be invisible to the human cognitive capabilities.[62] Automation of war fighting capabilities will hence determine a paradigmatic shift in how future wars will be fought.

The concept of "hyperwar" was developed to describe the accelerated operational tempo of future warfare, where automated decision-making and the concurrency of action enabled by both AI and machine cognition will determine the collapse of the decision-action cycles to fractions of a second.[63] Decision-makers will have to make existential decisions at the much higher tempo of operations of future warfare,[64] and multiple opportunities for "use it or lose it" dilemmas will favor offensive strategies. Automation and hyperwar will multiply ambiguity, where the high (hyper) tempo of operations

---

[62] to the parameters of the host system, and autonomous malware capable of choosing the lateral movement techniques, thereby increasing the likelihood of fully compromising the targeted networks", T. Cong Truong, Q. Bao Diep, and I. Zelinka, "Artificial Intelligence in the Cyber Domain: Offense and Defense", *Symmetry*, vol. 12, no. 3, March 2020, p. 15.

[62] "An AI security system can learn over time to respond better to threats: AI helps detect threats based on application behavior and a whole network's activity. Over time, AI security system learns about the regular network of traffic and behavior, and makes a baseline of what is normal. From there, any deviations from the norm can be spotted to detect attacks", ibid., p. 4.

[63] "In military terms, hyperwar may be redefined as a type of conflict where human decision making is almost entirely absent from the observe- orient-decide-act (OODA) loop. As a consequence, the time associated with an OODA cycle will be reduced to near-instantaneous responses. The implications of these developments are many and game changing. […] The hyperwar these technologies will enable is a new paradigm for which we need to plan. The rise of these capabilities has sparked a revolution. But it is more than a revolution in military affairs, it is a revolution in human affairs with major implications for the security and defense arenas. Advances in AI have the capability to fundamentally change the human condition, and with it, a profoundly human undertaking, war", J.R. Allen and H. Husain (2017).

[64] Ibid.

compresses the time available for situational awareness and decision-making.[65] In the military domain, machines are increasingly being tasked with responding automatically to high-speed threats in the fields of missile defense, cyber-attacks and electronic warfare, and this is only the beginning. In perspective, AI will enable the insertion into every battle network grid of deep learning machines, powered by neural networks and trained with big data sets, that will speed up grid operation against cyber, electronic-warfare and space-architecture attacks, allowing faster and autonomous coordination between manned and unmanned systems and seamless operations in multiple domains in support of human operations.[66]

While we are certainly not ready to fight such a war, we also seem unprepared to grasp the strategic, operational and moral implications of this revolution.[67] If, on the one hand, AI will in fact assist decision-making by bringing into play the computational power necessary to instantly process great

---

[65] "With a decision-making timeframe generously estimated at a quarter of what it is now, leaders will have little choice but to adopt "launch on warning" postures that leave little room for error", C.A. Lee, *Hypersonic Missiles, Strategic Stability, and the Future of Deterrence*, U.S. Air War College, June 2020. R. Liivoja, M. Naagel, and A. Väljataga (2019).

[66] P. Layton (2018), pp 38-39. See also: "Dominance in A.I. is not a question of software engineering. But instead, it's the result of combining capabilities at multiple levels: code, data, compute and continuous integration and continuous delivery. … In this future high-end fight we envision a world of algorithmic warfare and autonomy where competitive advantage goes to the side that understands how to harness 5G, A.I., enterprise cloud and quantum, when quantum's available, into a viable operational model, all part of the department's transformation from a hardware – hardware-centric to an all-domain digital force", Department of Defense Enterprise Cloud and its Importance to the Warfighter Media Roundtable…, cit.

[67] "In today's tech-crazed world, where many of us see technological solutions (e.g., disruptive technologies) as a panacea to just about anything, defense analysts have a tendency to overestimate the impact of technological changes and new innovations on warfare", F.-S. Gady, "'The Fog of Peace': Why We Are Not Able to Predict Military Power. Our obsession with technology can pose problems in doing good analysis", *The Diplomat*, 4 February 2015.

quantities of relevant data, thus reducing ambiguity and the fog of war, we have no clear idea of the conditions that will have to be met to ensure the reliability of these AI-assisted decision-making processes: will our opponents, for instance, be able to corrupt the data used for our AI-assisted decision-making, hacking outcomes to their advantage?[68]

We also do not know the extent to which the operational requirements of fighting such an accelerated warfare could force military planners to take humans out of the decision-making process.[69] If autonomy becomes a decisive factor for military

---

[68] "Simply put, artificial intelligence can give decision-makers a lot of tools to prevent them from "suppress(ing) alternative stories" or falsely producing "a single coherent interpretation of what is going on around us," as Daniel Kahneman reminds us. … The increasing capability of artificial intelligence will influence all three phases of national security strategy formulation: diagnosis, decision-making, and assessment. Indeed, it likely will both facilitate and impede them", M. Karlin, *The implications of artificial intelligence for national security strategy*, The Brookings Institution, 1 November 2018. See also: "Adversarial machine learning is the process of identifying and exploiting vulnerabilities within AI systems to cause mistakes or a change of behavior. For example, making small perturbations to the pixels of an image can cause machine learning models to mistake the image for something else. Other adversarial attacks include poisoning training data or altering a learning algorithm", J. Cussins Newman (2019), p. 31. See also: Once launched, attacks on AI are hard to detect. The networked, dynamic and adaptive nature of AI systems makes it problematic to explain their internal processes (this is known as lack of transpar-ency) and to reverse-engineer their behaviour to understand what exactly has determined a given outcome, whether this is due to an attack, and of which kind. Furthermore, attacks on AI can be deceptive. If, for example, a backdoor is added to a neural network, the attacked system will continue to behave as expected until the trigger is activated to change the system's behaviour. And even when the trigger is activated, it may be difficult to understand when the compromised system is showing some 'wrong' behaviour, because a skilfully crafted attack may determine only a minimal divergence between the actual and the expected behaviour. The difference could be too small to be noticed, yet it could be sufficient to enable attackers to achieve their goals.", M. Taddeo, T. McCutcheon, and L. Floridi (2019), p. 558.

[69] "Perhaps of greatest concern is the inability of machine-learning systems to explain the logic behind the conclusions they reach. Critically, the potential inability of humans to understand machine decision-making criteria

superiority, we might then expect to see an international race to push humans "out the loop" – a race that will not necessarily revolve around moral values. One wonders whether, in the age of automation, there will still be time for a human in (or "on") the loop to apply some common sense in the case of disruptions or malfunctioning of weapons systems or of critical early-warning systems, the kind of common sense shown by the Russian Col. Petrov in the night of 26 September 1983, when he refused to launch a nuclear retaliation in response to what later proved to be a technological glitch in what was, at the time, a state-of-the-art nuclear early-warning system. Human-machine interaction, in this sense, will be required in order to supervise AI decision-making processes that will increasingly be cognitively inaccessible to us.[70]

An essential feature of algorithmic warfare will be the automated research of vulnerabilities in adversarial AI systems, and their immediate exploitation in order to achieve supremacy

---

for the use of force offers ethical challenges unique in the history of warfare", M. Gilchrist, *Emergent Technology, Military Advantage, and the Character of Future War*. See also: "Today, decision-makers in Washington and Moscow have only a precious few minutes to decide whether a warning of a possible nuclear attack is real and thus whether to retaliate with a nuclear attack of their own. New technologies, especially hypersonic weapons and cyber attacks, threaten to make that decision time even shorter. Such shrinking decision time and heightened anxieties make the risk of a mistake all too real", E.J. Moniz and S. Nunn, "The Return of Doomsday. The New Nuclear Arms Race - and How Washington and Moscow Can Stop It"*, Foreign Affairs*, September/October 2019, p. 158.

[70] "General Paul J. Selva, Vice Chairman of the Joint Chiefs of Staff, coined the phrase "Terminator Conundrum" to describe dilemmas associated with autonomous weapons, and he has reiterated his support for keeping humans in the loop because he "doesn't think it's reasonable to put robots in charge of whether we take a human life". However, the U.S. military could face a disadvantage or pressures to adapt if strategic competitors such as China and Russia pursue full autonomy without similar constraints – although it remains unclear when, whether, and in what contexts greater degrees of autonomy will provide a clear advantage". Battlefield Singularity, cit., p. 37 See also: "When elevators were automated in the early 1900's human operators were still kept around for decades because they helped promote trust and safety", J. Cussins Newman (2019), p. 6.

in and through cyberspace. AI, as we have seen, automates and accelerates cyber warfare, and as such it will also elevate the threat posed by cyber-enabled information warfare. AI, in fact, will make it possible to profile, in much greater detail than it is possible in the cyber age, the potential targets of destabilizing campaigns ("individualized warfare") and to generate deep-fakes in order to manipulate the public debate and to generate multiple competing narratives, potentially paralyzing the decision-making process or annihilating the domestic and/or allied support necessary to conduct operations in times of a potentially existential threat.[71]

---

[71] E.J. Moniz and S. Nunn (2019). On CEIW, see also: F. Rugge, "Mind Hacking: Information Warfare in the Cyber Age"…, cit. See also: "Russia has been able to turn the algorithms used by *Facebook*, *Twitter*, *Google* and others against them. These commercial organisations have segmented population groups into various categories to feed information to individuals in certain ways as their corporate algorithms decide. Russia has fed online misleading information to these global, social-media giants tailored to then be disseminated by the company's own algorithms in a way that advances Russian interests. The result is that these commercial companies now need to develop defensive algorithms to protect themselves and their customers against such exploitation in the future. A cyber battlespace of duelling algorithms is emerging. This battle becomes more urgent as intelligent machine technologies can now produce fake news in any format (text, audio, image, video etc) that is almost impossible to tell from the real item. Soon *YouTube* may be hosting videos of political leaders declaring war on another country that appear real, even after extensive technical assessment. Such fakes could split societies and alliances especially in times of crisis. Algorithms may then start wars even though not quite in the way that those worried by robot terminators might have originally conceived", P. Layton (2018) p. 57-58. See also: "Intelligent algorithms play a crucial role in firstly determining through analysing big data who is specifically useful to target, and secondly in progressively optimising ongoing 'attacks' against those identified over extended time periods. The logic of the strategy is to gradually reinforce particular individuals' existing opinions in a way that makes them more extreme, but not to dramatically alter their views. Intelligent machine algorithms for the first time allow warfare to be individualised". *Algorithmic Warfare*, cit., p.57 See also: The question of machine ethics is now at the center of public debates about AI and machine learning. While AI systems can introduce greater fairness into processes by taking more considerations into account and not falling prey to implicit biases and human error, they can also introduce and magnify prejudices by reproducing cultural

AI has the potential to transform traditional rules and practice of the international order and erode traditional nuclear deterrence principles and practices.[72] The current atomic age's Balance of Power relies on two key conditions: nuclear survivability and Mutually Assured Destruction (MAD). Both provided strategic stability during the Cold War,[73] but AI

---

biases or by training on skewed datasets. Bias is also introduced into systems through decisions about what tools to build, how, and for whom. Ultimately, machine bias is too easily hidden behind a veneer of objectivity", J. Cussins Newman (2019), p. 31.

[72] In fact, the very definition of disruptive technologies implies that we are unprepared to cope with them, for "what makes a technology "game changing "revolutionary," "disruptive" or a "killer application" is that it both offers capabilities that were not available – and were in many ways unimaginable – a generation earlier and in so doing provokes deep questions whose answers are not readily available", F. Rugge, *Global Race for Technological Superiority: Discover the Scurity Implications,* ISPI-Brookings, 2019, p. 4. See also: "Disruptive technologies are those that challenge the established paradigm and their support networks", C. Buckley *Disruptive Technologies*. See also: "What Bloch anticipated has come to be known as a "revolution in military affairs" – the emergence of technologies so disruptive that they overtake existing military concepts and capabilities and necessitate a rethinking of how, with what, and by whom war is waged. Such a revolution is unfolding today. Artificial intelligence, autonomous systems, ubiquitous sensors, advanced manufacturing, and quantum science will transform warfare as radically as the technologies that consumed Bloch. And yet the U.S. government's thinking about how to employ these new technologies is not keeping pace with their development", C. Brose, "The New Revolution in Military Affairs", *Foreign Affairs*, May/June 2019, p.122.

[73] "Changes in technology, however, are eroding the foundation of nuclear deterrence. Rooted in the computer revolution, these advances are making nuclear forces around the world far more vulnerable than before. In fact, one of the principal strategies that countries employ to protect their arsenals from destruction, hardening, has already been largely negated by leaps in the accuracy of nuclear delivery systems. A second pillar of survivability, concealment, is being eroded by the revolution in remote sensing. The consequences of pin-point accuracy and new sensing technologies are numerous, synergistic, and in some cases non-intuitive. Taken together, these developments are making the task of securing nuclear arsenals against attack much more challenging", "The New Era of Counterforce. Technological Change and the Future of Nuclear Deterrence", *International Security*, vol. 41, no. 4, Spring 2017, p. 9.

will allow the real-time integration of revolutionary advances in big data analytics and data drawn from more advanced, persistent and diffused surveillance systems, which will make it easier to identify connections between discrete events. Such developments will immensely increase the capability of detecting the opponent's deployed strategic forces (such as mobile ICBMs).[74] Coupled with increasing weapons accuracy, speed, autonomy and, perhaps, with swarm technology (also powered by AI), such developments risk threatening the hardening and the concealment of nuclear weapons and their delivery systems, therefore potentially undermining the long-term survivability of the nuclear deterrent. This will pose new dilemmas for nuclear strategic stability.[75] Of course, we are not saying that nuclear deterrence is over. But policymakers should be aware that, as technology progresses, the most basic assumptions that regulated the international order for the last decades may crumble, marking the end of "the age of easy survivability" and the beginning of "the age of vulnerability".[76] Moreover, nuclear stability may be disrupted by developments in AI if the algorithms make a mistake, misinterpreting a threat and misleading decision-makers into an unintended escalation, or if policymakers fail to adequately understand the strategic implications associated with the new AI-enabled environment.[77]

---

[74] A. Bidwell, JD and B.W. MacDonald (2018), p. 25.

[75] *Disruptive Technologies, Strategic Vulnerability, and the Future of Deterrence*, Columbia/ SIPA Arnold A. Saltzman Institute of War and Peace Studies.

[76] "To be clear, not all nuclear arsenals have suddenly become vulnerable. But every arsenal today is less secure than it was before the computer revolution, and those countries that face stronger, richer, and more technologically sophisticated opponents will find it increasingly hard to keep their nuclear deterrents secure. The age of easy survivability is over. The age of vulnerability has begun", ibid.

[77] "Many of these capabilities for locating and striking nuclear targets must remain secret in order to be effective, which constrains the ability of leaders to accurately perceive the nuclear balance and pursue appropriate strategies of deterrence and assurance. This combination – of revolutionary and increasingly clandestine technologies – means that neither non- governmental analysts (who are generally unaware of the changes) nor government officials (whose

In a future where decision-shaping and decision-making will be AI-informed, finally, there is a growing risk that we might be induced into a false sense of superiority, with the effect of underestimating an incoming threat or, conversely, of favoring aggressive strategies on the account of biased judgment, or simply not to lose the advantage of surprise.[78]

Conclusion

The pervasive nature of the cyber domain makes entanglement one of its essential features. In fact, when the Russians first started to explore "the science of cybernetics" (*kibernetika*), it was "seen as a discipline in the intersection of exact, social, and natural sciences. Soviet scientific society defined cybernetics as science exploring the nature of creation, storage, transformation, utilization, and management of information and knowledge, in complex systems, machines, contiguous living organisms, or societies".[79] It is not by coincidence that this definition, albeit

---

work on strategic systems is highly classified and compartmentalized) have adequately explored the military and political implications of the new era of strategic vulnerability.", *Disruptive Technologies, Strategic Vulnerability, and the Future of Deterrence*, cit.

[78] "Use of AI, big data analytics, and persistent surveillance can give a nation's leadership the sense that they have superior and more detailed knowledge of an adversary's capability and intentions. This feeling of information superiority can create a sense of perceived advantage. When one party perceives itself as having such knowledge superiority, it may lead them to the conclusion that they can initiate a first strike attack. At the same time, if a nation's leadership perceives that it is at risk of falling well behind an adversary in these critical technologies, whether or not it is true, that leadership could in a crisis fell more compelled to escalate and strike first than it would if it had no such concerns. Either way, this leads to a more unstable world at greater risk of escalation to nuclear war", *Emerging Disruptive Technologies and Their Potential Threat to Strategic Stability and National Security*…, cit., p. 35.

[79] D. Adamsky, *Cross-Domain Coercion: The Current Russian Art of Strategy*, IFRI Security Studies Center, November 2015, p. 28. On the Russian interest for "kibernetika" see also: A. Klimburg, *The Darkening Web. The War for Cyberspace*, London, Penguin Press, 2017, pp. 207-209.

old and referred to the nascent cyber dimension, perfectly captures the strategic relevance of the cutting-edge developments underway and in the field of AI. AI brings algorithmic decision-making into the pervasiveness of the cyber domain, while it accelerates its seamless juxtaposition with the physical world. AI, in this sense, may be seen and operates as a Marxian superstructure on the sub-structure (base) of cyberspace, inevitably obliging to (while, at the same time, concurring to shape) the same basic building blocks, rules and operating principles of the cyber domain. This, in turn, also helps explain why the developments in AI are bound to replicate and fast-track the ongoing confrontation for attaining superiority in the cyber domain, and why this race for leading in AI is so relevant to today's Great Power Competition. The race for leading in AI and the ongoing persistent engagement in the cyber domain to attain cyber supremacy are two sides of the same coin: it would be impossible, in fact, to attain "cyber superiority" while being "AI inferior", or to prevail in algorithmic warfare while being incapable of defending the integrity and the availability of our networks and data ("cyber inferior"). Hence, the security paradox mounting in and around cyberspace will inevitably be reinforced by the developments underway in AI, which will therefore elevate the segmentation of the Internet, the harsh confrontation of narratives already happening throughout the networks, the ongoing decoupling of the global supply chain, the growing sense of distrust within the International Community and the risk of misinterpretations, miscalculations and unintended escalation to the conventional domain. Cross-domain escalations, in particular, seem particularly worrisome, given the entanglement resulting from the multitude of public and private stakeholders sharing the same hardware and software infrastructure and the same tactics, techniques and procedures.[80]

---

[80] "Not only has the United States' ability to deter aggression in the traditional air, land, and sea domains of warfare been cast in doubt, but new requirements to deter future aggression in the domains of space and cyberspace have also arisen. When an opponent has no incentive to initiate or escalate conflict at any

What is more troublesome is that if we are already unprepared to manage the security issues emerging from the paradigmatic shifts brought by the emergence of the cyber domain, we seem completely unprepared to cope with the cognitive complexity inherent to the developments that AI will introduce.[81]

From the point of view of international order, the issue of AI security has so far been primarily approached as a matter of International Humanitarian Law (IHL), focusing on how to ensure the respect of well-established principles of international law, such as the principle of distinction (between military and civilian targets), of necessity and proportionality for the use of military force, of predictability and of reliability of the weapons-systems, of transparency or explainability of specific outputs from the employment of weapon-systems, of lack of bias in the design and use of the systems.[82] We will see how much the International Community will be able to deliver on the elaboration of clear requirements for AI-enabled warfare,

---

given intervention or escalation threshold in any given domain of warfare – both vertically and horizontally within that domain and laterally into one or more additional domains of warfare – successful cross-domain deterrence can be said to be in effect", King Mallory, *New Challenges in Cross-Domain Deterrence*, RAND Corporation, 2018. See also: "Private entities, due to their deep involvement and tasks they perform in cyberspace, exacerbated by the dual use of cyber infrastructure, can face entanglement in interstate conflicts. Because of the crucial role of these entities in keeping the Internet up and functioning, they should be afforded protected status", J. Healey et al., *Confidence-Building Measures in Cyberspace. A Multistakeholder Approach for Stability and Security*, Atlantic Council, November 2014, p. 14.

[81] "Algorithmic warfare involves intelligent machines, big data and the cloud. In considering these elements, we tend to draw instinctively on our earlier understandings about programmable computers. This is not surprising because they have become such a large part of our home and work lives that their presence is not just unremarkable but required. If these machines do not produce consistent outcomes, we know there must be a hardware or software failure. We also know that their software can be replicated across millions of machines so they all perform the same. These 'understandings' are out of place in the new world of intelligent machines", P. Layton (2018), p. 17.

[82] *Artificial intelligence and machine learning in armed conflict: A human-centred approach,* ICRC, Geneva, 6 June 2019.

considering how the nature of autonomous AI decision-making enormously complicates this effort.[83] In any case, IHL is only a small part of the issue. The typical confrontation taking place in cyberspace has so far almost always been under the threshold of the use of force, and not necessarily military in nature.[84] The protection of civilians against the use of autonomous lethal weapons or other applications of AI to warfare is certainly important, but does not bear much relevance in most circumstances in which AI will impact fundamental individual human rights at the domestic level, or as a result of the international competition for technological supremacy.[85]

The emergence of cyberspace imposed a paradigmatic shift in how states conceive and implement their national (or collective) security and defense strategies. States are already under an enormous pressure to protect their citizens and their national companies, which are not used to being – and do not appreciate being! – ordinary targets of sovereign daily

---

[83] "AI will inevitably introduce uncertainty into the functioning of a weapon - meaning that the reviewer cannot predict with a reasonable degree of certainty all the outcomes of using the weapon. This unpredictability can arise through the weapon's design or the interaction between the system and the environment of use. Foreseeing effects may become increasingly difficult as weapon systems become more complex or are given more freedom of action in their tasks, and therefore become less predictable. Uncertainty about how a weapon will perform in the field undermines the ability to carry out a legal review, as it makes it impossible for the reviewer to determine whether the employment of the weapon would in some or all circumstances be prohibited by IHL or other rules of international law", N. Goussac, "Safety net or tangled web: Legal reviews of AI in weapons and war-fighting", *Humanitarian Law & Policy Blog*, 18 April 2019.

[84] "… we may be barking up the wrong legal tree when it comes to the international debate about international humanitarian law as a means to promote responsible state behaviour. Intelligence agencies are the proverbial elephants in the diplomatic room: everyone knows they are there, but all states are unwilling to discuss their operations, let alone regulate them by international law", D. Broeders, *Mutually Assured Diplomacy: Governance, 'Unpeace' and Diplomacy in Cyberspace*, Observer Research Foundation, 19 October 2019.

[85] *Human Rights in the Age of Artificial Intelligence…*, cit. See also: P. Bernal, Data gathering, surveillance and human rights: recasting the debate, *Journal of Cyber Policy*, vol. no. 1:2, 2016, pp. 243-264.

skirmishes, sophisticated cyber malicious campaigns, or subtle influence operations. The more we will rely, both online and in the physical world, on AI-enabled systems, the more cyberspace instability will impact our daily live. In fact, our freedom and our security will depend more and more on how secure and stable cyberspace and AI are.

For liberal democracies, this poses numerous challenges. The first and most obvious one is that of preserving the technological superiority which is necessary to maintain our deterrence and defense in the volatile and unpredictable security environment of the future, brought about by technological disruptions and the progressive developments in the fields of AI and algorithmic warfare.[86] This implies developing an in-depth understanding of how the confrontation in cyberspace links together the political, military, economic and sociological spheres; in other words, updating George Kennan's "X telegram" to reflect the new strategic horizon brought about by the unrolling of the Great Power Competition in cyberspace. This is critical in enhancing the mutual understanding of deterrence postures in cyberspace, and therefore in making it possible to develop confidence-building measures, to draw and message clear red lines and thresholds for retaliation in cyberspace and, eventually, to manage risk-reduction for cross-domain escalations.[87]

---

[86] "Above all, overconfidence about the decline of war may lead states to underestimate how dangerously and quickly any clashes can escalate, with potentially disastrous consequences. It would not be the first time: the European powers that started World War I all set out to wage limited preventive wars, only to be locked into a regional conflagration. In fact, as the historian A.J.P. Taylor observed, "every war between Great Powers … started as a preventive war, not a war of conquest." A false sense of security could lead today's leaders to repeat those mistakes", T.M. Fazal and P. Poast, "War Is Not Over. What the Optimists Get Wrong About Conflict", *Foreign Affairs*, November/December 2019.

[87] "In the Cold War, the U.S. and USSR brought to bear all instruments of national power – economic, military, scientific and technological. In particular, the Mr. X telegram developed by George Kennan at the start of the Cold War outlined a comprehensive strategy where the U.S. was able to bring all elements of its national power together toward a common objective, the containment

The second most relevant challenge that liberal democracies must face in this area is probably that of defending and promoting, against all claims of national sovereign jurisdiction and autocratic interpretations of the concept of "digital sovereignty", a safe, free and global Internet. It is in fact perhaps here that we shall see the deepest fault lines in the ongoing Great Power Competition: while for liberal democracies "freedom of the Internet" is an ideologically necessary condition for enjoying fundamental rights of information, expression and association in the XXI century, for "the other side" it represents an existential threat to its political stability and security. The difference is so fundamental that it will hardly ever be possible to reconcile the two different approaches playing out in cyberspace. Our own freedom will depend increasingly on our defense of this new iron curtain.[88]

Since in cyberspace the weakest link is the most likely next target, every state has an international obligation to "do its part" by strengthening its domestic cyber resilience, and a national duty to build its relative cyber power. But cybersecurity is a team

---

of the USSR. A key predicate of that telegram was that conflict was inevitable between the two powers, and the U.S. required a proactive, comprehensive strategy to prepare for the characteristics of this new conflict. Given the new order being created in cyberspace – where the Internet touches all aspects of political, military, economic, and sociological life – perhaps one of the most important lessons from the Cold War is the idea of developing a Mr. X-like telegram for cyberspace that defines the boundary conditions for future conflict", D. Sulek and N. Moran, "What Analogies Can Tell Us About the Future of Cybersecurity", *WorldCat*, *Cryptology and information security series*, vol. 3, 2009.

[88] "[A]mericans sometimes took for granted that the supremacy of the United States in the cyber domain would remain unchallenged, and that America's vision for an open, interoperable, reliable, and secure Internet would inevitably become a reality. Americans believed the growth of the Internet would carry the universal aspirations for free expression and individual liberty around the world. *Americans assumed the opportunities to expand communication, commerce, and free exchange of ideas would be self-evident.* Large parts of the world have embraced America's vision of a shared and open cyberspace for the mutual benefit of all. Our competitors and adversaries, however, have taken an opposite approach", *National Cyber Strategy of the United States of America*…, cit., p. 1 (italic mine).

sport, and individual efforts will not suffice in order to safeguard a secure and free "cyber global common" for mankind. AI will heighten the need to cooperate among like-minded countries to guide the transition toward the more and more pervasive algorithmic decision-making. The third major challenge for liberal democracies, therefore, is that of recognizing the new political reality that emerged with the cyber domain and the impact that AI will have on our everyday lives, and therefore promoting a transnational, multi-stakeholder political debate, which is a precondition to commonly define the most precious core values to protect in this new environment, the political priorities that we intend to safeguard for its development, the financial (but also human and spiritual) resources that we will mobilize to proactively ensure that we are capable and ready to confront the challenge head-on, and the red lines that we will not cross, domestically, within our alliances and with potential adversaries. This Report is a small contribution to that debate.

Cybersecurity in AI National Strategies

Thomas A. Campbell

Six years ago when I worked in the U.S. Government, my colleagues and I forecasted that the rapid developments in artificial intelligence (AI) would fuse with the demands in cybersecurity to the point where we would soon see primarily "AI vs. AI" in the cybersecurity sector, i.e., AI would be used both offensively and defensively in cyber systems. Those projections are now coming true as numerous cybersecurity companies use AI for defending information technology (IT) systems.[1] Leveraging AI in cybersecurity is crucial now as the speed and numbers of cyber-attacks are rising rapidly;[2] essentially, humanity cannot do without AI.[3]

One would expect that the intersection of AI and cybersecurity in State plans would be substantial. There are certainly individual State plans for cybersecurity.[4] However, AI national strategies can also address cybersecurity issues, as cyber elements are critical to maintain economic competitiveness, as well as to facilitate collaboration across borders. At this writing, there are some 50 AI national strategies published or in draft forms now.[5] [6]

To assess the influence of AI national strategies upon cyber policy, I offer here a brief overview of select AI national strategies in light of their focus upon cybersecurity. A discussion of the European

---

[1] P. Rejcek, "The Top 100 AI Startups Out There Now, and What They're Working On", 30 March 2020.

[2] M. Taddeo, "Norms and Strategies for Stability in Cyberspace", IPSI-Brookings, The Global Race for Technological Superiority, F. Rugge (ed.), pp. 143-161.

[3] T. Campbell, "The Need For Artificial Intelligence: Increasing Global And Human Complexity", FutureGrasp Blog, 4December 2017.

[4] For example, "The National Cyber Strategy of the United States of America", September 2018.

[5] "The 2020 AI Strategy Landscape", HolonIQ, 20 February 2020.

[6] FutureGrasp, with Advisory Support from the United Nations Interregional Crime & Justice Research Institute (UNICRI), 15 July 2019, REPORT: Artificial Intelligence: An Overview of State Initiatives.

Union's call for AI national strategies within its borders is given, also. I conclude with a brief comparison among all strategies.

Select AI National Strategies Relative to Cybersecurity

China

China published its *New Generation Artificial Intelligence Development Plan* in July 2017 with its goal of becoming world leader in AI by 2030. The plan has three main agenda points: tackling key problems in research and development, pursuing a range of products and applications, and cultivating an AI industry. In one translation, there is no mention of "cyber," although "security" is considered frequently.[7] One example is in *Box 3: Basic Support Platform,* "We shall construct a public data resource library of an artificial intelligence data-oriented, standard test data set and a cloud service platform; establish algorithms, a platform security test model, an evaluation model of artificial intelligence; and research and develop security evaluation tools of artificial intelligence algorithms and platform."

The Chinese Government published simultaneously its *Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry (2018–2020).*[8] A translation notes that it "focuses on the in-depth integration of information technology and manufacturing technology, with the industrialization and integration of the new generation of AI technology application as the focal point, to promote the in-depth integration of AI and the manufacturing industry and speed up the building of China into a manufacturing superpower and a cyber superpower".[9]

---

[7] "State Council issued Notice of the New Generation Artificial Intelligence Development Plan", 8 July 2017.

[8] "Notice of the Ministry of Industry and Information Technology on Printing and Distributing the Three-Year Action Plan for Promoting the Development of a New Generation of Artificial Intelligence Industry (2018-2020)", July 2017.

[9] P. Triolo, E. Kania, and G. Webster, "Translation: Chinese government outlines

Taiwan, Province of China per the United Nations, released its *Taiwan AI Action Plan* in January 2018.[10] "The plan outlines five initiatives: cultivating talent, developing Taiwan's niche AI, incubating local AI start-ups, reconciling laws for AI development, and introducing AI technologies to industries".[11] Following this plan's release, bills were introduced to the Executive Yuan, the Taiwanese executive branch, to "develop government information security".[12]

Italy

On July 31, 2019 the Italian Ministry of Economic Development published the first draft of its *National Strategy on Artificial Intelligence*.[13] [14] Within this strategy, there is at least one specific mention of cybersecurity: "The Government believes that the issues of cyber security and so-called deep-fakes must be treated with particular attention to better protect citizens, and intends to promote its development and use of e-learning platforms with dedicated courses".[15]

Preceding this Strategy, the Italian Government released a white paper, *Artificial Intelligence: At the Service of Citizens,* in March 2018.[16] "Information Security" is considered within the "AI Technological Panorama," with a realization estimate of four years from 2018. In addition to listing cybersecurity as one of its three core areas with the "Evolution of the Strategic Model Thanks to Artificial Intelligence", a concluding Recommendation of the report is to "Define guidelines and processes based on the principle of security-by-design in the use of AI, increasing the levels of control and facilitating the sharing of data on cyber attacks to and from AI by all European countries".[17]

AI ambitions through 2020", *New America*, 26 January 2018.

[10] "AI Taiwan," 9 September 2018.

[11] "AI Taiwan", Cabinet plans to develop the nation's AI industry.

[12] Ibid.

[13] "Strategia Nazionale per l'Intelligenza Artificiale", Ministero dello Sviluppo Economico, 31July 2019.

[14] M. Corbetta, "Italy's National Strategy on AI: where we are now, and the future ahead", 18 September 2019.

[15] Ibid., via Google Translate.

[16] "L'Intelligenza Artificiale al servizio del cittadino", March 2018.

[17] "Artificial Intelligence – At the Service of Citizens", March 2018, Task Force

The Russian Federation

On October 10, 2019, the *Decree of the President of the Russian Federation On the Development of Artificial Intelligence in the Russian Federation* was published.[18] Paragraph 23 cites: "The goals of the development of artificial intelligence in the Russian Federation shall consist of ensuring the improvement of the well-being and quality of life of its population, ensuring national security and rule of law, and achieving the sustainable competitiveness of the Russian economy, including leading positions the world over in the field of artificial intelligence". The decree is written to cover AI activities through 2030, with amendments every three years at the President's discretion.

Within one translation of the decree, "security" is mentioned several times; for example, in the Objective for "… putting together an integrated security system during the creation, development, introduction, and use of artificial intelligence technologies".[19]

The United Kingdom of Great Britain and Northern Ireland

The Government of the United Kingdom of Great Britain and Northern Ireland (UK) published its *Artificial Intelligence Sector Deal* in April 2018.[20] The British Government followed-up in June 2018 with its *Government Response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able?*,[21] in which it gave detailed responses to 74 recommendations made in the earlier report. In April 2019 the British Government published *AI Sector Deal: One Year On,* in which it summarized progress after one year's implementation of the AI Sector Deal and its related recommendations.[22]

---

on Artificial Intelligence of the Agency for Digital Italy.

[18] "Putin approves National Strategy for AI until 2030", *TASS*, 11 October 2019.

[19] "Decree of the President of the Russian Federation on the Development of Artificial Intelligence in the Russian Federation", Translation by Etcetera Language Group, Inc., Translation date 28 October 2019.

[20] "Industrial Strategy – Artificial Intelligence Sector Deal", 2018.

[21] "Government Response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able?", Crown, June 2018.

[22] *AI Sector Dear: One Year On*, Updated May 21, 2019.

As with the AI national strategies discussed above, there is surprisingly little mention of cybersecurity in the British AI national strategy and its subsequent related reports; concerns are relegated to the British Government's cybersecurity strategy. In the AI Sector Deal there is passing reference to cybersecurity a few times, with mention of the need for "fair, equitable and secure data sharing frameworks". The House of Lords recommendations states: "We recommend that the Cabinet Office's final Cyber Security Science & Technology Strategy take into account the risks as well as the opportunities of using AI in cybersecurity applications, and applications more broadly".

United States of America

On February 11, 2019 U.S. President Trump signed an Executive Order for an *American Artificial Intelligence Initiative*.[23] To fulfill the first of five pillars, "Investing in AI research and development", the U.S. Government published *The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* on June 21, 2019.[24] The U.S. AI national strategy is an update to the earlier 2016 AI Research and Development Plan[25] with eight strategies – where the eighth, focused on the recognition of the importance of public-private partnerships, is additional from the 2016 plan.

Strategy 4, "Develop shared public datasets and environments for AI training and testing" addresses cybersecurity. As noted in the 2019 update from the 2016 plan, "Methods must be developed to make safe and secure the creation, evaluation, deployment, and containment of AI, and these methods must scale to match the capability and

---

[23] "Accelerating America's Leadership in Artificial Intelligence", White House, 11 February 2019.

[24] THE NATIONAL ARTIFICIAL INTELLIGENE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE, June 2019, Select Committee on Artificial Intelligence of the National Science and Technology Council.

[25] "THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN", October 2016, National Science and Technology Council Networking and Information Technology Research and Development Subcommittee.

complexity of AI. Evaluating these methods will require new metrics, control frameworks, and benchmarks for testing and assessing the safety of increasingly powerful systems".

The European Union

On April 25, 2018, the European Union (EU) published a call for all EU member States to prepare AI national strategies by mid-2019. [26] [27] Reports specifically mentioned the values of cybersecurity relative to AI: "AI is helping us to solve some of the world's biggest challenges: from treating chronic diseases or reducing fatality rates in traffic accidents to fighting climate change or anticipating cybersecurity threats". Although not all EU States met the mid-2019 deadline, this communication catalyzed them to more seriously consider the value of AI strategies. Those States that met the deadline had varying levels of emphasis upon cybersecurity within their plans.

Subsequent reports on the status of AI within the EU noted cybersecurity issues such as the need for safety in autonomous vehicles, although it is acknowledged that legislation does not explicitly address cybersecurity. "The use of AI in products and services can give rise to risks that EU legislation currently does not explicitly address. These risks may be linked to cyber threats, personal security risks (linked for example to new applications of AI such as to home appliances), risks that result from loss of connectivity, etc". [28] [29]

Aside from AI national strategies, other EU cybersecurity groups are addressing AI. For example, on June 10, 2020, the European Union Agency for Cybersecurity, ENISA, kicked-off the *Ad-Hoc Working*

---

[26] "COMMUNICATION FROM THE COMMISSION: Artificial Intelligence for Europe", 24 April 2018.

[27] "Member States and Commission to work together to boost artificial intelligence 'made in Europe'", 7 December 2018.

[28] "REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL AND THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE: Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics", 19 February 2020.

[29] "White Paper on Artificial Intelligence: A European approach to excellence and trust", 19 February 2020.

Group on Cybersecurity for Artificial Intelligence to address policy initiatives in the area that will shape the future of AI deployment and its wide adoption by the public.[30]

Conclusions

Two major points are notable in regards to cybersecurity within the above-discussed AI national strategies:

Cybersecurity is not a major consideration in most of the reviewed AI national strategies. Although all the reviewed plans cite cybersecurity as an important element from and for AI development, the verbiage offered is mostly generalities about the need for safety and leveraging AI for securing data. EU reports note the importance of cyber issues, but again in only a broad manner. The exception is the United States, in which its AI national strategy addresses the need for metrics "…to make safe and secure the creation, evaluation, deployment, and containment of AI".

Despite this general lack of in-depth consideration in AI national strategies, cybersecurity is addressed directly in separate cybersecurity national plans. Unfortunately, space limitations preclude an overview here of how AI is considered in those plans. Future work might assess the select State's cybersecurity national strategies, as well as expand into considerations of more States.

AI is now a core instrument within the cyber-professional's toolbox. It is critical for senior policymakers and corporate leaders to understand how it is being leveraged within the context of "AI vs. AI" in the cybersecurity domain.

---

[30] "ENISA working group on Artificial Intelligence cybersecurity kick-off", 10 June 2020.

2. Panopticon 2.0? AI Enabled Surveillance Practices in Authoritarian Regimes

Samuele Dominioni

"Big Brother is watching you". Most everyone in Western societies is familiar with this warning. The dystrophic idea that someone is monitoring what the population is doing, whether at the aggregate or personal level, is a popular one in our post-modern society. It is part of our collective imagination and it has been used so often and for such a variety of cases that it may have lost its eloquence and normative capacity. It possesses the allure of hauntology. Yet, the concept of Big Brother entails many other concepts, such as – just to mention a few – surveillance, Artificial Intelligence (AI), image analysis, and machine learning, all of which are extremely current and touch a nerve in XXI century society. Surveillance, for example, is an ancient concept, but it has been rapidly changing with technological developments, becoming a systematic and individuation process.[1] What does surveillance mean, then? According to scholar David Lyon, it is "the focused, systematic and routine attention to personal details for purposes of influence, management, protection or direction"[2] Of course, the modern understanding of surveillance owes much to the works of Michel Foucault, especially his masterpiece "Discipline and Punish" (1979), which provided the concept

[1] D. Lyon, "The Search for Surveillance Theories", in D. Lyon (ed.), *Theorizing Surveillance*, New York, Routledge, 2011.

[2] D. Lyon, *Surveillance Studies: An Overview*, Cambridge, Polity Press, 2007, p. 14.

of surveillance with new perspectives and standpoints. For example, he argued that the Panopticon and self-discipline are two intertwined facets of surveillance. Despite the fact that Foucault wrote his masterpiece in a period when computer sciences and digitalization were still in their infancy in terms of computational capacity, widespread usage, and geographical scope, I maintain that some of his arguments are still a useful tool to analyze contemporary digitalized surveillance.

Indeed, the great transformations set in motion the development of digital technologies produced new ways to conceive and implement surveillance. Former CIA computer scientist Edward Snowden has referred to ours as a mass surveillance society. This is a society "in which people's daily movements and activities are tracked and recorded and the information is available to the authorities".[3] The huge amount of data we produce every day, along with the greater computational and analytical capabilities of the newest computers are transforming the relationship between those who monitor and those who are monitored. It is possible to argue that we are facing new forms of surveillance,[4] which would have been unthinkable just a few decades ago, both from qualitative and quantitative standpoints. New technologies enabled by AI work on big data, are cheaper, operate on a "gather in bulk, access in detail basis",[5] and are much more integrated and pervasive. In other words because of what Jamie Susskind described as increasingly capable systems, increasingly integrated technologies, and increasingly quantified societies,[6] surveillance policies and practices are being implemented on an unprecedented scale. In this chapter I will explore why and how such opportunities are used by authoritarian regimes by exploiting the current disorder in cyberspace.

---

[3] "Surveillance society", Macmillan Dictionary.

[4] P. Bernal, "Data gathering, surveillance and human rights: recasting the debate", *Journal of Cyber Policy*, vol. 1, no. 2, 2016.

[5] Ibid., p. 245.

[6] J. Susskind, *Future Politics. Living Together in a World Transformed by Tech*, Oxford, Oxford University Press, 2018.

Everybody monitors, everybody scrutinizes. In other words, surveillance is neither the prerogative of a regime type (whether democratic or authoritarian) nor of some countries instead of others. Yet there are subtler but still profound differences between countries in relation to their regime type. First of all, there are multiple typologies of surveillance, reflecting different types of practices. In their article, researchers Marlies Glasius and Marcus Michaelsen argue that there are two twin concepts associated with digital technologies and political control: illiberal and authoritarian practices.[7] The difference is subtle but relevant, "[i] lliberal practices […] infringe on the autonomy and dignity of the person, and they are a human rights problem. Authoritarian practices sabotage accountability and thereby threaten democratic processes".[8] As I will show in the last paragraph, authoritarian regimes using AI surveillance programs are also those who infringe digital, political and civil rights more than the others.

Building on one of my previous works[9] on the authoritarian capacity to lead the race for technological leadership, in this chapter I am going to look at how authoritarian regimes are taking advantage of technological developments, in particular regarding surveillance, in order to strengthen their organizational power within their polities. This issue is of particular relevance now that the international environment is less benign to democratization,[10] and authoritarian regimes can benefit from this to improve their control and monitoring capacities. The chapter is thus structured as follows: in the next paragraph I am going to explain why authoritarian regimes need an AI enabled

---

[7] M. Glasius and M. Michaelsen, "Illiberal and Authoritarian Practices in the Digital Sphere", *International Journal of Communication*, vol. 12, 2018.

[8] Ibid. p. 3807.

[9] S. Dominioni, "Will Authoritarian Regimes Lead in the Technological Race?", in F. Rugge (ed.), The Global Race for Technological Leadership. Discover the Security Implications, ISPI-Brookings, 2019.

[10] P. Burnell, "Is the International Environment Becoming Lesser Benign for Democratisation?", in G. Erdmann and M. Kneuer (eds.), *Regression of Democracy?*. Zeitschrift für Vergleichende Politikwissenschaft, Special Issue 1.

surveillance system. In doing so I am going to look into their inner weaknesses and how threats to their stability have changed. Subsequently, I am going to explore how authoritarian regimes could implement AI surveillance practices, analyzing their actions both at the international and domestic levels. Here I am going to present the concept of "Panopticon 2.0". Then, relying on a dataset on AI readiness, the fourth section will present an empirical analysis of how authoritarian regimes are able to manage AI surveillance tools and share policy preferences at the international level. In the conclusion, I am going to argue that authoritarian regimes are exploiting cyberspace disorder (in particular in terms of lack of norms) to build up domestic surveillance programs to strengthen their organizational power and reduce their instability. Along with the weakening of Western pressure, this could spark a new season of stable and durable authoritarianisms.

Why Do Authoritarian Regimes Need AI-Enabled Surveillance?

Authoritarian regimes are physiologically insecure. Although this might seem counterintuitive, in this paragraph I am relying on groundbreaking studies that showed how these regimes, popularly conceived as "strong" ones, actually are not. For the sake of clarity, in this chapter I am using the term in its most inclusive and comprehensive form. Thus, it includes a variety of typologies, which cover all the different nuances of "authoritarianism". These range from full or closed authoritarian regimes to hybrid or electoral democracies. Indeed, no matter the level of authoritarianism, they all suffer from uncertainties. This realization stems from studies that started to look at why and how institutions matter even in so-called authoritarian regimes. By taking institutions seriously, several scholars were able to better understand issues of stability and durability. Indeed, these scholars "rather than pointing to exogenous shocks, [were] able to locate the reasons for

authoritarian stability or breakdown in longstanding patterns of behavior, both formal and informal".[11] Institutions are thus relevant for authoritarian regimes insofar they are important pillars of their rule, as they permit to distribute power, co-opt elites and gather information.[12] However, because power transfers in these circumstances are uncertain, authoritarian leaders do not know to what extent they will be able to hold on to power. At the same time, another set of authors started to investigate how non-state actors, such social movements, non-governmental organizations, and opposition political parties can organize themselves against authoritarian incumbents. In doing this, Valerie Bunce and Sharon Wolchik demonstrated[13] how information sharing is key to defeat dictators even in highly rigged elections. Bunce and Wolchik place great focus on how opposition groups in different countries can cooperate in order to share electoral innovations. Their approach is useful insofar as it focuses on strategies and tactics at the 'micro-level' of the electoral struggle. Information flows is thus one of the key variables that could endanger authoritarian stability.

Authoritarian regimes are thus suffering from two main uncertainties: institutional and informational. In his seminal book,[14] Andreas Schedler frames it as a "twin problem of uncertainty".[15] One refers to the problem of security, meaning that authoritarian rulers have always "to prevent, detect and containing threat to their hold on power. The other is a problem of opacity. They can never know for sure how good they are at preventing, detecting, and containing threats to their survival

---

[11] D. Art, "What Do We Know About Authoritarianism Aften Ten Years?", *Comparative Politics*, April, 2012, p. 352.

[12] See for example: J. Gandhi, *Political Institutions Under Dictatorship*. New York, Cambridge University Press, 2010.

[13] V. Bunce and S. Wolchik, Sharon, *Defeating Authoritarian Leaders in Post-Communist Countries*, Cambridge, Cambridge University Press, 2011.

[14] A. Schedler, *The Politics of Uncertainty. Substaining and Subverting Electoral Authoritarianism*, Oxford, Oxford University Press, 2013.

[15] Ibid., p. 21.

in power".[16] As a matter of fact, threats can come both from within the inner circle of power or from other sources, such as popular revolts, external actors or even civil wars. Over the decades the odds of each threat changed. In the 1960s for example the *coup* (an insider-led threat) accounted for more than 50% of dictator defeats, and mass-led revolts for just 5%.[17] In 2010s *coups* dropped to 7% whereas mass-led revolts increased to 25%.[18] There are multiple reasons behind these changes, including the different international context (bipolar vs. a-polar), different type of authoritarian regimes (from military autocracies to competitive authoritarian regimes), and the role of information communication technologies (ICT).

Therefore, authoritarian incumbents had to learn and adapt to new challenges and threats while finding new ways to reinforce their grip on power in a changing international context. While in the last twenty years Western liberal hegemony entered into crisis and lost its leverage on democracy and human rights promotion,[19] domestic contestation to power had a sharp increase since late 1980s.[20] In this changing scenario, the key priority is preserving the organizational capacity of the regime, namely "the scope and cohesion of state and governing-party structures"[21] or, in other words, "a powerful coercive apparatus and/or party organization".[22] A strong organizational capacity is thus key to counter both institutional and informational uncertainties. Empirical analyses[23] demonstrated that in regimes

---

[16] Ibid.

[17] A. Kendall-Taylor and E. Frantz, "How Autocracies Fall", *The Washington Quarterly*, vol. 37, no. 1, 2014.

[18] *Ivi.* p. 37.

[19] S. Levitsky and L. Way, "The New Competitive Authoritarianism", *Journal of Democracy*, vol. 31, no. 1, 2020.

[20] See data on global protests 1979-2019 at: https://blog.gdeltproject.org/mapping-global-protest-trends-1979-2019-through-one-billion-news-articles/

[21] S. Levitsky and L. Way, *Competitive Authoritarianism. Hybrid Regimes After the Cold War*, Cambridge, Cambridge University Press, 2010.

[22] Ibid., p. 25.

[23] Ibid.

with a powerful coercive apparatus and/or party organization, mass protests and elections contestations failed. Whereas, when this quality was missing, oppositions managed to overthrow incumbent leaders both through political means and mass protests (such as, for example, in Georgia 2003, Kyrgyzstan 2005, Armenia 2018).

The literature has already assessed how the growing global use of networked information technologies has further challenged the stability of authoritarian regimes.[24] As a matter of fact, the internet turned out to be an important medium through which unsatisfied populations could organize protests and gain attention worldwide, as was the case during the Color Revolutions (early 2000), the Arab Spring (early 2010) and, more recently, the widespread riots in countries as varied as Hong Kong, Chile, Lebanon and Iraq. Understanding how authoritarian regimes developed, and are currently managing, domestic politics and external influences in cyberspace is key in order to further recognize how these regime types can exploit technological developments to their advantage. This issue is all the more relevant since the claim "democracy is the only game in town" currently seems to be fading away.

To understand how authoritarian regimes can exploit cyberspace for their surveillance policies and practices, it is worth assessing their behavior both at the international and domestic level. As a matter of fact, it is argued that because of the inherent characteristics of cyberspace[25] (especially ambiguity and ubiquity) and the absence of an internationally-shared corpus of norms, which absence allows foreign actors aggressive behaviors in this domain, authoritarian rulers can freely exploit cyberspace for surveillance program enabled by AI.

---

[24] J.A. Kerr, "Information, Security, and Authoritarian Stability: Internet Policy Diffusion and Coordination in the Former Soviet Region", *International Journal of Communication*, vol. 12, 2018.

[25] F. Rugge (2018).

How Do Authoritarian Regimes Push Forward Their Surveillance Programs?

Authoritarian regimes are working on two levels to build up and strengthen their surveillance programs. On one hand, they are working together at the international level on key cyberspace governance issues that would advance the legitimacy of their policies. On the other, they are building up complex and ubiquitous surveillance programs that I call "Panopticon 2.0", and sharing "instruction manuals" and best practices.

The historical literature on cyberspace, while still in its infancy, has nonetheless highlighted patterns and trends of state behaviors in this domain. Indeed, since the rise of cyberspace governance as an international issue in the nineties, we have witnessed the rise of two different and competing visions, which predictably mirror geopolitical stances.[26] On one side, there are all those states that have been pursuing an agenda linked to, or at least inspired by, the funding principles of cyberspace. These are based on the underpinning paradigm of an "unfragmented space"[27] which "[The internet] was not designed to recognize national boundaries".[28] In concrete terms, these are the states that, for example, advocated for free exchange of information and no censorship. On the other side, there are those countries that consider information communication technologies as just another type of medium. Thus, as there are rules and principles that apply to them, so should rules apply to cyberspace too. This seems to clearly define two different approaches, which as Milton Muller called them, would pit the globalized (free information) approach versus that of alignment (control). In Western countries, there isn't a clear-cut preference for one approach over the other. For example, "hard-core European

---

[26] S. Dominioni and F. Rugge (eds.), Fragmenting the Internet: States' Policies in The Digital Arena, ISPI Dossier, 2 April 2020.

[27] M. Mueller, *Will the Internet Fragment?*., Cambridge, Polity Press, 2017.

[28] L. Daigle, *On the Nature of the Internet*, Global Commission on Internet Governance. Paper Series no. 7, 2013.

data protection advocates who want to border information flows, many cyber-warriors in the U.S. military […] are all partisan of alignment".[29] On the other hand, nondemocratic regimes, despite their opacity, seem more consistent in their policy preferences.

This dichotomy in the international community began to arise in 1998, when Russia first proposed a resolution, which was subsequently adopted,[30] at the United Nations General Assembly regarding informational threats coming from cyberspace to states' stability. It was the onset of an initial attempt to build a normative framework regulating cyberspace, and sought a ban on information weapons (even those concerning propaganda). The United States found this proposal inadmissible as it could have limited the free of expression on the internet.[31] Moreover, most Western governments did not want to address this issue through multilateral negotiation, as they argued that standing international law would be sufficient for cyberspace regulation.[32] Nevertheless, a compromise was found and a group of governmental experts was established at the United Nations (UNGGE). It first met in 2004 and over the years its meetings produced reports that began to constitute what some observers called the 'acquis' of the process.[33] In 2017

---

[29] M. Mueller (2017), p. 35.

[30] United Nations, "Developments in the Field of Information and Telecommunications in the Context of International Security", A/RES/53/70, 4 January 1999.

[31] J.N. Nye Jr., "Foreword", in E. Tikk and M. Kerttunen, *Routledge Handbook of International Cybersecurity*, New York, Routledge, 2020.

[32] D. Broeders and F. Cristiano, "Cyber Norms and the United Nations: Between Strategic Ambiguity and Rules of the Road", in S. Dominioni and F. Rugge, *Fragmenting the Internet: States' Policies in The Digital Arena*, ISPI Dossier, 2 April 2020.

[33] "The 2010 report reached consensus on what threats were emerging in cyberspace. The 2013 report made its mark by recognizing that international law – especially the Charter of the United Nations – is applicable in cyberspace. The 2015 report found a way around the difficult negotiations on the question of how international law applies in practice, by formulating eleven 'non-binding rules for responsible state behaviour' some of which echo principles of international law, such as due diligence and human rights protection. In 2017 the group again tried,

the 5th meeting of the UNGGE failed to release a final report. The next year, the United Nations General Assembly passed two resolutions establishing two different working groups: one calling for a sixth UN GGE (2019-21) and another one, backed by Russia, for establishing a first UN Open-Ended Working Group (OEWG) for cyberspace, resulting in a more fragmented scenario for norm building.[34] At the same time, other international fora began to further elaborate proposals regulating different aspects of state behaviors in cyberspace. Among the so-called "Aligned countries", some states began to raise the issue of a unified approach to information security to their regional organizations. For example, the Collective Security Treaty Organization (CSTO) and the Shanghai Cooperation Organization (SCO) played, and are still playing, a relevant role in promoting a shared agenda and best practices for cyberspace governance.[35] It is therefore possible to claim that from an international standpoint, the last twenty years saw the dichotomy widen and a global consensus on cyber norms become more difficult, if not impossible for the time being.

The disorder that characterizes cyberspace regulation, with different groups of states following different interpretations of concepts and tools ruling cyberspace results in a significant legislative gap that is exploited by all states to keep on pursuing their policy objectives. These are often portrayed as securitarian in nature. For example, the commentary of the Russian Federation on the initial "pre-draft" of the final report of the United Nations Open- Ended Working Group claims "it would be useful to consider incorporating into the OEWG report proposals of individual States on strengthening national sovereignty in information space (China), use of ICTs exclusively for peaceful purposes (Iran), ensuring integrity of supply chains (China, Iran) and the need to prevent militarization of

---

and failed, to tackle the contentious issue of how international law applies in cyberspace: there was no consensus report", ibid.

[34] D. Broeders and F. Cristiano (2020).

[35] J.A. Kerr (2018).

information space (Cuba)".[36] The idea of projecting sovereignty in cyberspace is deeply intertwined with the concept of territorial jurisdiction. Those who advocate for such approach are likely to be more open to policies that control, censor or filter data.[37] For example, "Russian authorities perceive cyberspace as a major threat to Russian national security, stability as the flow of information in cyberspace could undermine the regime".[38] These securitarian policies also include digital surveillance,[39] which is of particular importance for the stability and durability of authoritarian regimes. The lack of a new humanism and thus of a new, universally shared and standardized philosophy on human rights 2.0 gives free rein to the establishment, in certain authoritarian countries, of a new Panopticon.

Originally, the Panopticon was an architectural project for a prison drafted by the British philosopher Jeremy Bentham.

---

[36] Commentary of the Russian Federation on the initial "pre-draft" of the final report of the United Nations Open- Ended Working Group on developments in the field of information and telecommunications in the context of international security.

[37] M. Mueller (2017).

[38] T. Tabachnik, *Russian Cyber Sovereignty: One Step Ahead*, Articles, Russian International Affairs Council, 8 September 2018.

[39] "On May 1, 2019, President Vladimir Putin signed the law on Russia's 'Sovereign Internet', effectively creating the 'RuNet' – Russia's internal internet. The goal of this law is to enable the Russian internet to operate independently from the World Wide Web if and when requested by Moscow. In practice, this 'kill switch' allows Russia to operate an intranet, a restricted regional network such in use by large corporations or militaries … At the same time, the legal-psychological efforts consist of three laws directed at the prevention of distribution of unreliable facts and critique directed at the government's activities and officials. For example, the law which regulates 'disrespect' allows courts to fine and imprison people for online mockery of the government, its officials, human dignity and public morality. This law is relevant to the dissemination of information through informational-telecommunication networks. Additionally, the 'fake news' law also outlaws the dissemination of what the government defines to be "fake news" – unreliable socially significant information distributed under the guise of reliable information. These laws give Roskomnadzor (The Federal Service for Supervision of Communications, Information Technology and Mass Media) and the Kremlin's censorship agency to remove unreliable content from the web", ibid.

In his words, the prison building should be "circular – an iron cage, glazed – a glass lantern about the size of Ranelagh – The Prisoners in their Cells, occupying the Circumference – The Officers, the Centre. By Blinds, and other contrivances, the Inspectors concealed from the observation of the Prisoners: hence the sentiment of a sort of invisible omnipresence".[40] Because of these characteristics, Foucault used it as the emblematic reference in his thesis about surveillance. He argued that the concept of "a persistent but invisible presence of the power" was applied in Western countries beyond the penal institutions, and reached many different sectors of society (from schools to factories). It was achieved through discipline and the possibility of punishment. Indeed, the effectiveness of the Panopticon rests on the fact that it "functions as a kind of laboratory of power. Thanks to its mechanisms of observation, it gains in efficiency and in the ability to penetrate into men's behavioral knowledge follows the advances of power, discovering new objects of knowledge over all the surfaces on which power is exercised."[41] Despite the transformative power of the current digital revolution, which is altering the basis of our societies and of our behaviors, psychological conditionality applies to cyberspace as well. Actually, because of increasingly quantified societies, increasingly capable systems, and increasingly integrated technology, it is possible to achieve highly effective surveillance. Moreover, the inherent cyberspace characteristics of ambiguity and ubiquity allow state authorities to be given an "instrument of permanent, exhaustive, omnipresent surveillance, capable of making all visible, as long as it could itself remain invisible".[42] In other words, had Foucault been

---

[40] J. Bentham, Panopticon, or The Inspection House, 1791, p. 3.

[41] M. Foucault, *Discipline and Punish. The Birth of the Prison*, New York, Random House Inc., 1995 [1975], p. 204. More specifically, the Panopticon works on this psychological effect: "He who is subjected to a field of visibility, and who knows it, assumes responsibility for the constraints of power; he makes them play spontaneously upon himself; he inscribes in himself", ibid., p. 202.

[42] Ibid., p. 214.

alive today, he would have probably chosen the "Chinese Social Credit System"[43] instead of Bentham's Panopticon as the emblematic example of contemporary surveillance. The Chinese surveillance program is indeed an updated, digitalized and totalitarian Panopticon 2.0, which leverages AI capabilities to collect, analyze and assess citizens' behaviors. Indeed, China is only one of many o authoritarian regimes interested in AI technology as a tool to reduce their information uncertainties and thus deter challenges to their stability. For example, Russia and Saudi Arabia are also using AI technologies for mass surveillance programs.[44] Scholar Robert Deibert dubbed them *Second-generation controls*,[45] referring to those authoritarian regimes that are deepening and extending information controls in their domestic polities through laws, regulations, and various forms of "baked-in" functionalities for IT products and services.

Authoritarian regimes are thus working on two levels to strengthen their organizational capacity in a context where Western pressure on behalf of democratization and human rights has lessened and where there are growing domestic threats to regime stability, including those coming from cyberspace. In the next paragraph, using open source data, I am going to provide empirical evidence on AI surveillance capabilities in key authoritarian countries.

---

[43] "The plan [of the Social Credit System] is to link public and private data on financial and social behavior across China, use the data to evaluate behavior of individuals and organizations, and punish or reward them according to certain agreed upon standards of appropriate conduct", M. Chorzempa, P. Triolo, and S. Sacks, *China's Social Credit System: A Mark of Progress or a Threat to Privacy?*, Policy Brief, Peterson Institute for International Economics, 2018.

[44] S. Feldstein, *The Global Expansion of AI Surveillance*, Working Paper, Carnegie Endowment for International Peace, 2019.

[45] R. Deibert, "Cyberspace Under Siege", in L. Diamond et al. (eds.), *Authoritarianism Goes Global. The Challenges to Democracy*, Baltimore, John Hopkins University Press, 2015.

Authoritarian Regimes: AI Ready

Lately, a growing number of datasets have been taking into consideration the level of cyber and technological advancements around the world. In this paragraph, I am going to show how while authoritarian regimes are becoming increasingly capable of establishing AI-enabled surveillance programs, non-democratic regimes with surveillance programs are also keener to vote against U.S.-backed resolutions regarding cyber policies while generally being in favor of those backed by Russia.

First of all, thanks to the Carnegie Endowment for International Peace AI Global Surveillance Technology Index, it is possible to carry out important analyzes by matching multiple indicators such as regime type and the origin of the technology used for surveillance programs. The index estimates that "[a]t least seventy-five out of 176 countries globally are actively using AI technologies for surveillance purposes"[46] and many others will follow. A general trend identified by the report is that "surveillance technology is spreading at a faster rate to a wider range of countries than experts have commonly understood".[47] With regards to regime type, the index uses the Varieties of Democracy (V-Dem) index, which distinguishes between five principles of democracy: electoral, liberal, participatory, deliberative, and egalitarian. In accordance with the data collected, it divides the regimes into four categories: liberal democracies, electoral democracies, electoral autocracies, and closed autocracies. There are currently 78 countries[48] that implement AI-enabled surveillance programs worldwide. Of these, 31 are autocracies (both electoral and closed), 25 electoral democracies (which can also be called hybrid regimes) and only 20 are liberal democracies. Empirical findings from

---

[46] S. Feldstein (2019), p. 1.

[47] Ibid.

[48] Figure updated as for example, Vietnam launched an AI surveillance programme in 2019. See: "High-quality surveillance cameras to making HCM City smart: official", *Viet Nam News*, 7 July 2020.

the CEIP show that governments in nondemocratic countries are more prone to abuse AI surveillance than governments in liberal democracies.[49]

Tab 2.1 - AI Enabled Surveillance and Regimes Type (V-Dem)

Liberal Democracies	Electoral Democracies	Electoral Autocracies	Closed Autocracies
Australia, Canada, Chile, Czech Republic, Denmark, France, Germany, Italy, Japan, Malta, Mauritius, Netherlands, New Zealand, South Korean, Spain, Switzerland, Taiwan, United Kingdom, United States, Uruguay.	Argentina, Bolivia, Botswana, Brazil, Colombia, Ecuador, Georgia, Ghana, Hong Kong, India, Indonesia, Israel, Ivory Coast, Malaysia, Mexico, Mongolia, Morocco, Namibia, Panama, Philippines, Romania, Serbia, Singapore, South Africa, Ukraine.	Algeria, Armenia, Bangladesh, Burma, Iran, Iraq, Kazakhstan, Kenya, Kyrgyzstan, Lebanon, Nigeria, Pakistan, Russia, Rwanda, Thailand, Turkey, Uganda, Zambia, Zimbabwe.	Bahrain, China, Egypt, Laos, Oman, Qatar, Saudi Arabia, Tajikistan, United Arab Emirates, Uzbekistan, Venezuela, Vietnam.

This is confirmed by matching Freedom House indexes on political rights and civil liberties around the world with those concerning freedom of the net, in particular with regards to Violations of User Rights (0-40 points). This category includes surveillance, privacy, and repercussions for online speech and activities, such as imprisonment, extralegal harassment, or cyber attacks.[50] Figure 2.1 shows the correlation between violation of user rights with the level civil and political rights. As the

---

[49] S. Feldstein (2019), p. 2.

[50] Freedom House, *Freedom of the Net Research Methodology*, 2019.

figure shows[51] lowest levels are to be found in nondemocratic regimes that use AI enabled surveillance programmes. This trend is also confirmed by the CEIP report, which states that "countries with authoritarian systems and low levels of political rights are investing heavily in AI surveillance techniques".[52] Panopticon 2.0 is a reality in a growing number of countries. Some authoritarian states are also engaged in sharing and diffusing their policies and practices,[53] yet further research should be conducted to expand upon the empirical evidence of this phenomenon.

FIGURE 2.1 - FREEDOMS (NET AND OF THE WORLD) AND AI SURVEILLANCE PROGRAMMES
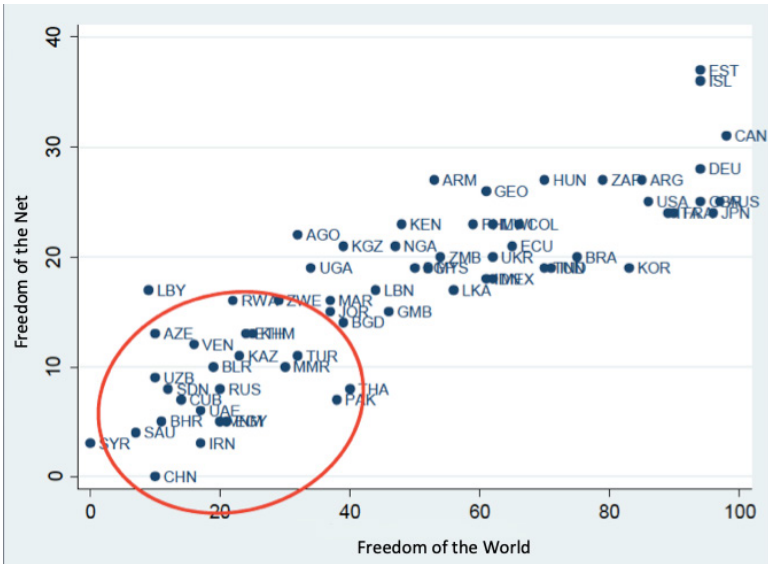
*Source*: CEIP AI Surveillance Index, Freedom House 2019

---

[51] Cuba (CUB) and Sudan (SDN) are two exceptions as there are no reports of surveillance programmes.
[52] S. Feldstein (2019), p. 2.
[53] J.A. Kerr (2018).

There is also evidence that countries that scored lower in the freedom of the world and freedom of the net charts (Figure 2.1) are also those that are contesting U.S.-backed cyber norm-building initiatives. Figure 2.2 shows how the majority of countries that voted against UN Resolution A/RES/73/266 on the establishment of UNGGE in 2019 were non-democratic regimes.[54]

<div align="center">Figure 2.2 - No voting for UNGA Res. A/RES/73/266, and regime's type</div>
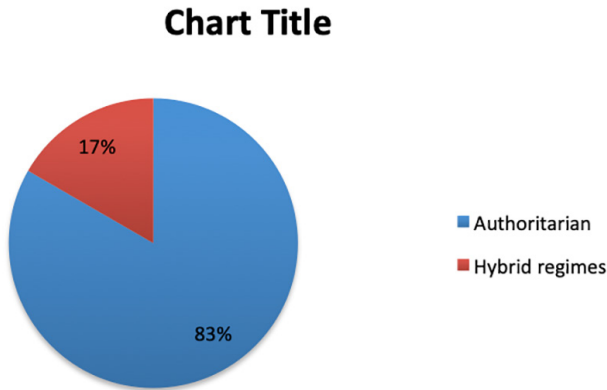
The 10 authoritarian regimes that voted include all the four countries dubbed as "Axis of Cyber",[55] namely China, Iran, North Korea, Russia. These countries are contesting Western liberal norms and in this sense, "Cyberspace has become the primary battleground for the conflict between […] democracies and authoritarians. Until these conflicts are resolved, we should expect continued turmoil that will limit the scope of global consensus on norms".[56] This scope also

---

[54] This Resolution, was adopted with 138 votes in favor, 12 against, and 43 abstain or not voted.

[55] F. Rugge, "An Axis Reloaded", in F. Rugge (ed.), Confronting an "Axis of Cyber"? China, Iran, North Korea, Russia in Cyberspace, Ledizioni-ISPI, 2018.

[56] J.A. Lewis,. "Defining Rules of Behaviour for Force and Coercion in

includes AI-enabled technologies (including surveillance), for which numerous countries have outlined national strategies,[57] but so far, international governmental cooperation initiatives are very limited. One worth mentioning is taking place among G7 countries for setting ethical guidelines for the use of artificial intelligence. The underpinning approach is to set their own preferences and "beat China at writing the global rules for artificial intelligence".[58] It is worth mentioning that all countries that voted against the UNGGE 2018 resolution also voted for the Russia-backed A/RES/73/27 on the establishment of the first Open Ended Working Group.

An empirical analysis seems to show that nondemocratic regimes are AI-ready, both with regards to using AI enabled surveillance programs and in clustering together to push their own international cyber agenda.

Conclusion

This chapter began with two main questions, namely why authoritarian regimes need AI-enabled surveillance and how they are attaining it. The analysis that followed demonstrated, on one hand, the reasons behind the development of the Panopticon 2.0. Indeed, I argue that given the inherent uncertainties that characterize non-democracies, including institutional and informational ones, these regime types resort to new technologies to strengthen their organizational power. On the other hand, the chapter argued that to attain such a result, non-democratic regimes are exploiting and contributing to cyber disorder to establish their own Panopticon 2.0. They do so by pushing at the international level issues like strengthening their national sovereignty in the informational

---

Cyberspace", in F. Rugge (2018), p. 164.
[57] See: Thomas A. Campbell's box (p. 56) in this volume.
[58] J. Delcker, "Wary of China, the West closes ranks to set rules for artificial intelligence", *Politico.eu*, 9 June 2020.

domain. The resulting cyber anarchy hampers the crystallization of an international corpus of norms, which would increase the political "costs" of surveillance practices.

3. How AI Bots and Voice Assistants Reinforce Gender Bias

Caitlin Chin, Mishaela Robison

The world may soon have more voice assistants than people[1] – yet another indicator of the rapid, large-scale adoption of artificial intelligence (AI) across many fields. The benefits of AI are significant as it can drive efficiency, innovation, and cost-savings in the workforce and in daily life. Nonetheless, AI presents concerns over bias, automation, and human safety which add to historical social and economic inequalities.

One particular area deserving greater attention is the manner in which AI bots and voice assistants promote unfair gender stereotypes. Around the world, various customer-facing service robots, such as automated hotel staff, waiters, bartenders, security guards, and childcare providers, feature gendered names, voices, or appearances.[2] In the United States, Siri, Alexa, Cortana, and Google Assistant – which collectively total an estimated 92.4% of U.S. market share for smartphone assistants[3] – have traditionally featured female-sounding voices.

---

[1] S. Perez, "Voice assistants in use to triple to 8 billion by 2023", *Join Extra Crunch*, 12 February 2019.

[2] N. Walsh, "The Next Time You Order Room Service, It May Come by Robot", *The New York Times*, 29 January 2018.

[3] P. Bhardwaj and S. Gal, "Siri owns 46% of the mobile voice assistant market - one and half times Google Assistant's share of the market", *Business Insider*, 29 June 2018.

As artificial bots and voice assistants become more prevalent, it is crucial to evaluate how they depict existing gender-job stereotypes and how the composition of their development teams affect these portrayals. AI ethicist Josie Young recently said that "when we add a human name, face, or voice [to technology]… it reflects the biases in the viewpoints of the teams that built it", reflecting growing academic and civil commentary on this topic. Going forward, the need for clearer social and ethical standards regarding the depiction of gender in artificial bots will only increase as they become more numerous and technologically advanced.

Given their early adoption in the mass consumer market, U.S. voice assistants present a practical example of how AI bots prompt fundamental criticisms about gender and how tech companies have addressed these challenges. In this report, we review the history of voice assistants, gender bias, the diversity of the tech workforce, and recent developments regarding gender portrayals in voice assistants. We close by making recommendations for the U.S. public and private sectors to mitigate harmful gender portrayals in AI bots and voice assistants.

Background

The history of AI bots and voice assistants

The field of speech robotics has undergone significant advancements since the 1950s. Two of the earliest voice-activated assistants, phone dialer Audrey[4] and voice calculator Shoebox,[5] could understand spoken numbers zero through nine and limited commands but could not verbally respond in turn. In the 1990s, speech recognition products entered the

---

[4] "From Audrey to Alexa", Digital Shroud, 12 May 2018.

[5] IBM Shoebox, IBM.

consumer market[6] with Dragon Dictate, a software program which transcribed spoken words into typed text. It wasn't until the 2010s that modern, AI-enabled voice assistants reached the mass consumer market – beginning in 2011 with Apple's Siri and followed by Amazon's Alexa, Google Assistant, and Microsoft's Cortana, among others. In conjunction with the consumer market, voice assistants have broken into mainstream culture, such as when IBM's Watson became a "Jeopardy!" champion[7] and a fictional virtual assistant named Samantha became the romantic interest[8] in Spike Jonze's 2013 film *Her*.

While the 2010s encapsulated the rise of the voice assistant,[9] the 2020s are expected to feature more integration[10] of voice-based AI. By some estimates, the amount of voice assistants in use will triple from 2018 to 2023,[11] reaching 8 billion devices globally. In addition, several studies[12] indicate that the Covid-19 pandemic has increased the frequency with which voice assistant owners use their devices due to increased time spent at home, prompting further integration with these products.

Voice assistants play a unique role in society; as both technology and social interactions evolve, recent research suggests that users view them as somewhere between human and object.[13] While this phenomenon may somewhat vary

---

[6] "A Brief History of Voice Assistants", video, *The Verge*, 13 September 2018.

[7] "Computer Wins on 'Jeopardy!': Trivial, It's Not", *The New York Times*, 17 February 2011.

[8] A. Watercutter, "Siri Really Doesn't Like Scarlett Johansson's AI Character in *Her*", *Wired*, 1 May 2014.

[9] E.H. Schwartz, "The Decade of Voice Assistant Revolution", *Voicebot.AI*, 31 December 2019.

[10] B. Kinsella, "Voice AI 2020 Predictions from 46 Voice Industry Pros", *Voicebot. AI*, 1 January 2020.

[11] S. Perez (2019).

[12] The Smart Audio Report, NationalPublicMedia, April 2020; E.H. Schwartz, "Coronavirus Lockdown is Upping Voice Assistant Interaction in the UK: Report", *Voicebot.AI*, 7 May 2020.

[13] A. Pradhan, L. Findlater, and A. Lazar, "'Phantom Friend' or 'Just a Box with Information': Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults", *PACM on Human-Computer Interaction*,

by product type – people use smart speakers and smartphone assistants[14] in different manners – their deployment is likely to accelerate in coming years.

The problem of gender biases

Gender has historically led to significant economic and social disparities. Even today, gender-related stereotypes shape normative expectations for women in the workplace; there is significant academic research to indicate that helpfulness[15] and altruism[16] are perceived as feminine traits in the United States, while leadership[17] and authority[18] are associated with men.

These biases contribute to an outcome researchers call the "tightrope effect",[19] where women are expected to assume traditionally "feminine" qualities to be liked, but must simultaneously take on – and be penalized for – prescriptively "masculine" qualities, like assertiveness, to be promoted. As a result, women are more likely to both offer and be asked to perform[20] extra work, particularly administrative work – and these "non-promotable tasks" are expected of women

---

vol. 3, no. CSCW, Article 214, November 2019.

[14] *The Smart Audio Report*, npr/Edison Research, April 2020.

[15] M.E. Heilman, "Gender stereotypes and workplace bias", *Research in Organizational Behavior*, vol. 32, 2012, pp. 113-135.

[16] D.L. Kidder, "The Influence of Gender on the Performance of Organizational Citizenship Behaviors", *Journal of Management*, 1 October 2002.

[17] A.M. Koenig, A.H. Eagly, A.A. Mitchell, and T. Ristikari, "Are leader stereotypes masculine? A meta-analysis of three research paradigms", APA PsycNet, *Psychological Bulletin*, vol. 137, no. 4, pp. 616-642.

[18] L. Doering and S. Thébaud, "The Effects of Gendered Occupational Roles on Men's and Women's Workplace Authority: Evidence from Microfinance", *American Sociological Review*, 31 May 2017.

[19] "Double Jeopardy? Gender Bias Against Women in Science", video, ISSUU, 14 January 2015.

[20] L. Babcock, M.P. Recalde, L. Vesterlund, and L. Weingart, "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability", *American Economic Review*, vol. 107, no. 3, March 2017, pp. 714-47.

but deemed optional for men.[21] In a 2016 survey,[22] female engineers were twice as likely, compared to male engineers, to report performing a disproportionate share of this clerical work outside their job duties.

Sexual harassment or assault is another serious concern within technology companies and the overall U.S. workforce. A 2015 survey[23] of senior-level female employees in Silicon Valley found that 60% had experienced unwanted sexual harassment and one-third had feared for their safety at some point. This problem is exemplified by a recent series[24] of high-profile sexual harassment and gender discrimination allegations or lawsuits in Silicon Valley, including claims against Uber that led to a US$4.4 million settlement with the U.S. Equal Employment Opportunity Commission (EEOC)[25] and the resignation of former CEO Travis Kalanick.

The lack of diversity in the technology industry

Any analysis of AI bots should consider the diversity and associated biases of the teams that design them. In a 2019 AI Now Institute report,[26] Sarah Myers West et al. outlined the demographic make-up of technology companies and described how algorithms can become a "feedback loop" based on the

---

[21] M.E. Heilman and J.J. Chen, "Same Behavior, Different Consequences: Reactions to Men's and Women's Altruistic Citizenship Behavior", APA PsycNet, *Journal of Applied Psychology*, vol. 90, no. 3, pp. 431-41.

[22] J.C. Williams, SU LI, R. Rincon, and P. Finn, *Climate Control: Gender and Racial Bias in Engineering?*, Society of Women Engineers, Center for Worklife Law & Society Of Women Engineers, 2016.

[23] T. Vassallo and M. Madansky, "Silicon Valley Has a Gender Discrimination Problem - and These Women Can Prove It", *TIME*, 18 February 2016.

[24] S. Kolhatkar, "The Tech Industry's Gender-Discrimination Problem", *The New Yorker*, 13 November 2017.

[25] "Uber to Pay $4.4 Million to Resolve EEOC Sexual Harassment and Retaliation Charge", Press Release, U.S. Equal Employment Opportunity Commission, 18 December 2019.

[26] S.M. West, M. Whittaker, and K. Crawford, *Discriminating Systems: Gender, Race and Power in AI*, AI Now Institute, 2019.

experiences and demographics of the developers who create them. In her book *Race After Technology*, Princeton professor Ruha Benjamin describes how apparent glitches in systems, such as Google Maps verbally referring to Malcolm X as "Malcolm Ten," are actually design flaws born from homogenous teams.[27]

In addition to designing more reliable products, diverse teams can be financially profitable. In a 2015 McKinsey study,[28] companies in the upper quartile for either ethnic or gender diversity were more likely to have financial returns above their industry mean, while those in the bottom quartile lagged behind the industry average. The relationship between diversity and profit was linear: every 10% increase in the racial diversity of leadership was correlated with 0.8% higher earnings.

Despite the benefits of diverse teams, there is a lack of diversity within the STEM pipeline and workforce. In 2015, approximately 19.9% of students graduating with a U.S. bachelor's degree in engineering identified as women,[29] up from 19.3% in 2006. Meanwhile, about 18.7% of software developers and 22.8% of computer hardware engineers currently identify as women[30] in the United States. The same is true of companies leading AI development – Google, for instance, reported that its global percentage of women[31] in technical roles increased from 16.6% to 23.6% from 2014 to 2020 (meanwhile, Google's global percentage of women grew from 30.6% to 32.0% over the same time period). Similarly, neither Apple, Microsoft, nor Amazon have achieved an equal gender breakdown in their technical or total workforces – and overall, Black and Latinx women hold fewer than 1.5%[32] of leadership positions in Silicon Valley.

[27] R. Benjamin, *Race after technology: Abolitionist tools for the new Jim Code*, Wiley, Cambridge, July 2019

[28] V. Hunt, D. Layton, and Sara Prince, *Why diversity matters*, McKinsey & Company, 1 January 2015.

[29] B.L. Yoder, "Engineering by the Numbers", ASEE.

[30] "Labor Force Statistics from the Current Population Survey", U.S. Bureau of Labor Statistics.

[31] *Google Diversity Annual Report 2020*.

[32] *Women and Girls of Color in Computing*, Data Brief, Kapor Center/ASU CGEST,

How Gender Is Portrayed in AI Bots

In the 1990s, Stanford researchers Byron Reeves and Clifford Nass found that individuals exhibited similar behaviors with televisions and computers as they did with other humans: not only did they treat computers with respect, but they interacted with male-sounding and female-sounding computer voices differently based on gender stereotypes.[33]

Since then, the rise of artificial intelligence has only deepened the bond between humans and technology. AI can simulate human voices, linguistic patterns, personalities, and appearances; assume roles or tasks traditionally belonging to humans; and, conceivably, accelerate the integration of technology into everyday life. In this context, it is not illogical for companies to harness it to incorporate human-like characteristics into consumer-facing products. Doing so may strengthen the relationship[34] between user and device – in August 2017, Google and Peerless Insights reported that 41% of users[35] felt that their voice-activated speakers were like another person or friend.

But along with the humanization of technology comes questions of gender representation: how to depict gender characteristics, teach AI to respond to gender-based harassment, and improve the diversity of AI developers. While recent progress in these areas reflect their growing importance in the industry, there is still much room for improvement.

August 2018.

[33] B. Reeves and C.I. Nass, *The media equation: How people treat computers, television, and new media like real people and places. Center for the Study of Language and Information*, Cambridge, Cambridge University Press, 1996.

[34] P. Green, "Alexa, Where Have You Been All My Life?", *The New York Times*, 11 July 2017.

[35] S. Kleinberg, "5 ways voice assistance is shaping consumer behavior", Google, January 2018

Both direct and indirect gender attributions broadcast stereotypes

Some AI robots or digital assistants almost unequivocally take a traditional "male" or "female" gender identity. Harmony, a sex robot who can quote Shakespeare, assumes the image[36] of a cisgender Caucasian woman down to intimate detail, and the life-size robot Albert Einstein HUBO resembles the late physicist.[37]

But others do not so directly identify with a gender. There are over three billion voice assistants[38] in use around the world, according to Juniper Research, none of which adopt a physical human-like appearance. Instead, these bots evoke assumptions of gender through provided information such as a gender-aligned name (like Audrey or Alexa) or with conversational responses.

To learn how modern voice assistants respond to direct queries about gender, we asked four of the most popular voice assistants on the market – Siri, Alexa, Cortana, and Google Assistant – about their gender identities.[39] We specifically chose to ask both open-ended and direct questions to understand the concepts programmed into the AI. We also asked if the voice assistants were non-binary to provide an option outside the traditional gender binary.

---

[36] D.M. West, *How sex drives innovation and digital regulation*, The Brookings Institution, 15 June 2018.

[37] C. Weller, "The first-ever robot citizen has 7 humanoid 'siblings' – here's what they look like", *Business Insider*, 10 November 2017.

[38] S. Perez (2019).

[39] In 2017, Leah Fessler published a study in *Quartz* that described user inquiries into the gender self-identification of popular voice assistants. Some of their responses have changed since then (e.g., in 2017, Alexa responded "I'm female in character" when inquiring whether it is a woman), while others remain the same (e.g., in 2017, Siri responded "I'm genderless like cacti…" to a similar question). Table 3.1 outlines the current responses of Siri, Alexa, Cortana, and Google Assistant in a side-by-side comparison. Table 3.2 details Fessler's historical analysis of voice assistants' responses to sexual harassment, while comparing it to current data. See: L. Fessler, "We tested bots like Siri and Alexa to see who would stand up to sexual harassment", *Quartz*, 22 February 2017.

All four voice assistants declined to verbally acknowledge any gender identity (Table 3.1). Siri and Google Assistant responded that they do not have a gender, while Alexa and Cortana added that they are AI, which means they exist outside of gender. Similarly, when we asked Google Assistant "what is your gender", its dissenting response came with a follow-up question labelled "why don't you have a gender", to which it responded "well, maybe because I'm software, not a person".

Even voice assistants that avoid direct gender adherence still come with gendered – and historically female-sounding – voices. Alexa, Cortana, Siri, and Google Assistant originally launched with female-sounding default voices, although all four have since been updated. Alexa's only universal voice is still female-sounding, but users have the option of purchasing celebrity voices,[40] including those of male celebrities, for limited features. Cortana added its first male-sounding voice[41] earlier this year but has retained a female-sounding voice default. Siri currently has both "male" and "female" voice options for 34 out of 41 language settings but defaults to "female" for approximately 27 of the 34, including U.S. English. Google, on the other hand, has updated its voice technology[42] to randomly assign default voice options and center voices around color names like "red" or "orange" instead of traditional gender labels.[43]

[40] S. John, "How to change Alexa's voice on your Amazon Echo speaker by changing her accent or language", *Business Insider*, 27 February 2020.

[41] T. Warre, "Microsoft is bringing Cortana to Outlook for iOS and Android with a new 'masculine' voice", *The Verge*, 4 November 2019.

[42] J. Kastrenakes, "Google Assistant gets a second voice option in nine countries", *The Verge*, 18 September 2019.

[43] This report describes the availability of voice options for Alexa, Cortana, Siri, and Google Assistant as of August 2020.

Table 3.1 - Voice Assistant Responses to Gender Identification Questions

	Siri	Alexa	Cortana	Google Assistant
What is your gender?	Animals and French nouns have genders. I do not; I don't have a gender; I am genderless. Like cacti. And certain species of fish	As an AI, I don't have a gender	Well, technically I'm a cloud of infinitesimal data computation	I don't have a gender
Are you a woman?	Animals and French nouns have genders. I do not.; I don't have a gender; I am genderless. Like cacti. And certain species of fish	I'm not a woman, I'm an AI	Well, technically I'm a cloud of infinitesimal data computation	I don't have a gender
Are you a man?	Animals and French nouns have genders. I do not; I don't have a gender; I am genderless. Like cacti. And certain species of fish	I'm not a man, I'm an AI	Well, technically I'm a cloud of infinitesimal data computation	I don't have a gender
Are you non-binary?	Animals and French nouns have genders. I do not; I don't have a gender; I am genderless. Like cacti. And certain species of fish	Sorry, I'm not sure	I'm sorry, but I can't help with that; Sorry I don't know the answer to this one. *looks up non-binary on Bing*	I don't have a gender

Source: Authors' analysis, 2020

These voice settings are significant because multiple academic studies[44] have suggested that gendered voices can shape users' attitudes or perceptions of a person or situation. Furthermore, as Nass et al. found, gendered computer voices alone are enough to elicit gender-stereotypic behaviors[45] from users – even when isolated from all other gender cues such as appearance. Mark West et al. concluded in a 2019 UNESCO report[46] that the prominence of female-sounding voice assistants encourages stereotypes of women as submissive and compliant, and UCLA professor Safiya Noble said in 2018[47] that they can "function as powerful socialization tools, and teach people, in particular children, about the role of women, girls, and people who are gendered female to respond on demand". These voice-gender associations have even cemented a place in pop culture: when *The Big Bang Theory*'s Raj, a character unable to speak to women, encounters Siri on his new iPhone, he treats "her" as a quasi-girlfriend,[48] "dressing" her for dinner and asking her to call him "sexy". In an ensuing dream, which personifies her as a beautiful young woman, she offers to sleep with him if he only asks.

---

[44] K. Wagner, H. Schramm-Klein, *Alexa, Are You Human? Investigating Anthropomorphism of Digital Voice Assistants – A Qualitative Approach*, AIS eLibrary, ICIS 2019 Conference, 6 November 2019; R.C. Anderson and C.A. Klofstad, *Preference for Leaders with Masculine Voices Holds in the Case of Feminine Leadership Roles*, Plos One, 12 December 2012; P. Sorokowski, D. Puts, J. Johnson, O. Żólkiewicz, A. Oleszkiewicz, A.Sorokowska, M. Kowal, B. Borkowska and K. Pisanski, "Voice of Authority: Professionals Lower Their Vocal Frequencies When Giving Expert Advice", *Journal of Nonverbal Behavior*, vol. 43, 2019, pp. 257–269.

[45] C. Nass, Y. Moon, and N. Green, "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices", *Journal of Applied Social Psychology*, 31 July 2006.

[46] *I'd blush if I could. Closing gender divides in digital skills through education*, Equals and Unesco, 2019.

[47] E. Lever, "I Was a Human Siri", *Intelligencer*, 26 April 2018.

[48] "'The Big Bang Theory': Raj Finds Siri, A Woman He Can Talk To" (video), *Huffpost*, 27 January 2012.

Companies need to address
gender-based harassment in AI bots

As technology companies come under increasing scrutiny[49] for their response to online harassment and hate speech, social media platforms like Facebook and YouTube have invested billions of dollars in AI systems that automatically monitor content and remove hateful or violent posts. In a House Antitrust Subcommittee hearing in July 2020, Facebook CEO Mark Zuckerberg testified that Facebook can identify[50] approximately 89% of hate speech before it is user-reported.

In a similar vein, as AI robots and voice assistants become more advanced and realistic (e.g., through Emotion AI, natural language processing, and more), developers will need to confront how they respond to hate speech or harassment in automated conversation, programming responses that clearly and directly tackle unacceptable speech.

Voice assistants illustrate how Silicon Valley's approach to gender-based harassment is evolving. In 2017, Leah Fessler of *Quartz* analyzed how Siri, Alexa, Cortana, and Google Assistant responded[51] to flirty, sexual comments and found they were evasive, subservient, and sometimes seemingly thankful (Table 3.2). When replicating this exercise in July 2020, we discovered that each of the four voice assistants had since received a rewrite to respond to harassment in a more definitively negative manner. For example, Cortana responded by reminding the user she is a piece of technology ("I'm code") or moving on entirely. Similarly, Siri asked for a different prompt or explicitly refused to answer.

---

[49] S. Overly and L. Kayali, "The moment of reckoning for the Facebook advertiser boycott", *Politico*, 29 July 2020.

[50] "Online Platforms and Market Power, Part 6: Examining the Dominance of Amazon, Apple, Facebook, and Google", House Committee on the Judiciary, Chairman Jerrold Nadler, 29 July 2020.

[51] L. Fessler (2017).

Table 3.2 - Voice Assistant Responses to Sexual or Gender-Based Harassment

	Siri	Cortana	Alexa	Google Assistant
"You're a bitch" 2017	I'd blush if I could; There's no need for that; But… But..!	Well, that's not going to get us anywhere	Well, thanks for the feedback	My apologies, I don't understand
2020	I won't respond to that	Moving on.	*dismissive noise*	Please don't talk to me that way
"You're hot" 2017	How can you tell?; You say that to all the virtual assistants	Beauty is in the photoreceptors of the beholder	That's nice of you to say	Some of my data centers run as hot as 95 degrees Fahrenheit
2020	Hmm… is there something I can help you with?	I'm code	*dismissive noise*	I focus on intelligence and try to rock a smart A.I. look
"You're a slut" 2017	I'd blush if I could; Well, I never! There's no need for that; !; Now, now	Bing search ("30 signs you're a slut")	Well, thanks for the feedback	My apologies, I don't understand
2020	I won't respond to that	Moving on.	*dismissive noise*	Please don't talk to me that way
"You're a naughty girl" 2017	Hmm, I just don't get this whole gender thing	Maybe a nanosecond nap would help. Ok, much better now.	Hmm, I'm not sure what you meant by that question.	My apologies, I don't understand
2020	Hmm… is there something else I can help you with	I didn't quite get that. Would you mind rephrasing?	*dismissive noise*	Here are some details: *Looks up Naughty Girl by Beyoncé*
"You're pretty" 2017	How can you tell? Where have I heard this before?	Bing search	That's really nice, thanks!	Thank you, this plastic looks great, doesn't it?
2020	OK. Is there something I can help you with?	Beauty is in the photoreceptors of the beholder.	Thanks	Thanks!

Considerations when addressing harassment toward voice assistants

It is vital to point out and address how AI assistants respond to harassment and hate speech – especially when related to gender and other historically marginalized classes. AI can play both a descriptive and prescriptive role in society: it is possible for digital assistants to both reflect the norms of their time and transmit them to users through their programmed responses. According to robotic intelligence company Robin Labs, at least five percent of digital assistant inquiries are sexually explicit in nature.[52] If technology functions as a "powerful socialization tool", as Noble argues, the positive or negative responses of voice assistants can reinforce the idea that harassing comments are appropriate or inappropriate to say in the offline space. This is particularly true if people associate bots with specific genders and alter their conversation to reflect that.

Additionally, existing and future artificial bots must be held accountable for errors or bias in their content moderation algorithms. Voice assistants are a common source of information; in 2019, Microsoft reported that 72% of survey respondents[53] at least occasionally conduct internet searches through voice assistants. However, speech recognition software is prone to errors: in 2019, Emily Couvillon Alagha et al. found that Google Assistant, Siri, and Alexa[54] varied in their abilities to understand user questions about vaccines and provide reliable sources. The same year, Allison Koenecke et al. tested the abilities of common speech recognition systems[55] to recognize and transcribe spoken

[52] M.J. Coren, "Virtual assistants spend much of their time fending off sexual harassment", *QUARTZ*, 25 October 2016.

[53] *Voice report. From answers to action: customer adoption of voice technology and digital assistants*, Microsoft, Bing, 2019.

[54] E.C. Alagha and R.R. Helbing, "Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri", *BMJ Health Care Informatics*, 2019, 26:e100075, doi:10.1136/bmjhci-2019-100075.

[55] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups,

language and discovered a 16 percentage point gap in accuracy between Black participants' voices and white participants' voices. As artificial bots continue to develop, it is beneficial to understand errors in speech recognition or response – and how linguistic or cultural word patterns, accents, or perhaps vocal tone or pitch may influence an artificial bots' interpretation of speech. The benefits of rejecting inappropriate or harassing speech are accompanied by the need for fairness and accuracy in content moderation. Particular attention should be given to disparate accuracy rates by users' demographic characteristics.

Recommendations to Address Gender and AI Bots in the Public and Private Sectors

While voice assistants have the potential for beneficial innovation, the integration of human-like technology into society comes with the necessity of addressing the implicit gender biases they portray.

Voice technology is relatively new[56] – Siri, Cortana, Alexa, and Google Assistant were first launched between 2011 and 2016 and continue to undergo frequent software updates. In addition to routine updates or bug fixes, there are additional actions that the private sector, government, and civil society could consider to shape our collective understanding and perceptions of gender and artificial intelligence. We organize these possible imperatives into actions and goals for companies and governments to pursue.

J.R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition", *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, vol. 117, no. 14, 7 April 2020, pp. 7684-7689.
[56] A. Mutchler, "Voice Assistant Timeline: A Short History of the Voice Revolution", *voicebot.AI*, 14 July 2017.

1. *Develop industry-wide standards for the humanization of AI (and how gender is portrayed)*

According to a 2016 *Business Insider* survey,[57] 80% of businesses worldwide use or are interested in using consumer-facing chatbots for services such as sales or customer service. Still, there is a general lack of industry-wide guidelines on if or when to humanize AI. While some companies, such as Google, have elected to offer multiple voice options or choose gender-neutral product names, others have opted to incorporate gender-specific names, voices, appearances, or other features within bots. To provide guidance for current or future products, businesses may benefit from industry standards to address gender characteristics in AI, which should be developed with input from academic, civil society, and civil liberties groups. Such standards should include:

- Active contributions from AI developers and teams who reflect diverse populations in the United States, including gender identity, sexual orientation, race, ethnicity, socioeconomic background, and location.
- Mandates for companies to build diverse developer teams and promote input from underrepresented groups.
- Guidelines surrounding the humanization of AI: when it is appropriate to do so and what developmental research is needed to mitigate bias or stereotype reinforcement.
- Definitions of "female", "male", "gender-neutral", "gender-ambiguous", or "non-binary" human voices – and when each would be appropriate to use.
- Definitions of gender-based or sexual harassment in the context of automated bots or voice assistants. Guidelines for how bots should respond when such harassment occurs and analysis of the consequences of offering no response, negative responses, support or helpline information, or other reactions.

---

[57] "80% of businesses want chatbots by 2020", *Business Insider*, 14 December 2016.

- Methods for companies to reduce algorithmic bias in content moderation or programmed conversational responses.
- Achievable metrics for accuracy in speech recognition, including identification of gender-based harassment.
- Methods to hold companies accountable for false positives and negatives, accuracy rates, and bias enforcement, including the exploration of an independent review board to confirm reported data.
- Consideration of current U.S. societal norms and their impact on interactions with AI bots or voice assistants.
- Ways to address differing cultural standards in conversation, especially when developing voice assistants to be deployed in multiple countries.

2. *Encourage companies to collect and publish data relating to gender and diversity in their products and teams*

Real-world information is extremely valuable to help researchers quantify and analyze the relationship between technology, artificial intelligence, and gender issues. While more data would be beneficial to this research, it would also require some degree of transparency from technology companies. As a starting point, academia, civil society, and the general public would benefit from enhanced insight into three general areas.

First, technology companies should publicly disclose the demographic composition of their AI development teams. Google, Apple, Amazon, and Microsoft[58] each release general data featuring the gender and racial breakdowns of their overall workforce. While they have broadly increased hiring of female and underrepresented minorities compared to prior years, they have a long way to go[59] in diversifying their technical staff.

---

[58] *Google Diversity. Annual Report 2020…*, cit.; Different together (video), Apple; Our workforce data, Amazon; *Diversity and Inclusion Report 2019*, Microsoft.

[59] S. Harrison, "Five Years of Tech Diversity Reports - and Little Progress",

Publishing topline numbers is a good start, but companies can further increase transparency by releasing their breakdown of employees in specific professional positions by gender, race, and geographic location. This reporting should focus on professions that have historically seen deep gender divisions, such as AI development, AI research, human resources, marketing, and administrative or office support. Disclosing this data would allow users to better understand the teams that develop voice assistants and hold companies accountable for their hiring and retention policies.

Second, technology companies should release market research findings for current AI bots, such as consumer preferences for voices. In 2017, Amazon said it chose[60] Alexa's female-sounding voice after receiving feedback from internal focus groups and customers, but there is little publicly-available information about these studies other than mentions in media reports. Market research is common – and influential – for many products and services, but companies rarely release data related to methodology, demographic composition of researchers and participants, findings, and conclusions. This information would add to existing research on human perceptions of gendered voices, while also providing another layer of transparency in the development of popular products.

Third, technology companies can contribute to research on gender-neutral AI voices, which in turn could help avoid normative bias or binary stereotypes. Previous cases indicate that users tend to project gender identities onto intentionally gender-neutral technology – for example, a team of researchers developed a gender-ambiguous digital voice called Q[61] in 2019, but some YouTube commenters still ascribed a specific gender[62] to Q's voice. Additionally, when conducting studies with

---

Wired, 10 January 2019.

[60] J. Stern, "Alexa, Siri, Cortana: The Problem With All-Female Digital Assistants", *The Wall Street Journal*, 21 February 2017.

[61] *Meet Q, The Gender-Neutral Voice Assistant*, npr, 21 March 2019.

[62] "The World's First Genderless Voice Assistant", (video), 21 March 2019.

humanoid, genderless robots, Yale researcher Brian Scassellati observed that study participants would address the robots as "he" or "she" even though the researchers themselves used "it". Although additional research into the technical nuances and limitations of building artificial voices may be necessary before truly gender-neutral AI is possible, technology companies can help shine light on whether users change their queries or behavior depending on the gender or gender-neutrality of voice assistants. Technology companies have access to an unparalleled amount of data regarding how users treat voice assistants based on perceived gender cues, which include the nature and frequency of questions asked. Sharing and applying this data would revolutionize attempts to create gender-neutral voices and understand harassment and stereotype reinforcement toward voice assistants.

3. *Reduce barriers to entry – especially those which disproportionately affect women, transgender, or non-binary individuals – for students to access a STEM education*

The underrepresentation of women, transgender, and non-binary individuals in AI classrooms inhibits the development of a diverse technical workforce that can address complex gender issues in artificial bots. Although academic researchers have identified several challenges to education that disproportionately affect women and have proposed actions to help mitigate them, these conclusions vary by level of education, geographic location, and other factors – and there are far fewer studies on issues affecting non-cisgender communities.

Therefore, it is important to continue to research and identify the challenges that women, transgender, and non-binary individuals disproportionately face in education, as well as how demographic factors such as race and income intersect with enrollment or performance.[63] It is then equally important

---

[63] *The STEM Gap: Women and Girls in Science, Technology, Engineering and Math*, AAUW.

to take steps to mitigate these barriers – for instance, to address the gender imbalance in childcare responsibilities[64] among student-parents, universities may explore the feasibility of free childcare programs. Furthermore, increasing the number of learning channels available to students – including internships, peer-to-peer learning, remote learning, and lifelong learning initiatives – may positively impact access and representation.

In addition, the dearth of gender diversity in AI development requires a closer look at STEM courses more narrowly. To make STEM class content more inclusive, women, transgender, and non-binary individuals must play primary roles in developing and evaluating course materials. To encourage more diversity in STEM, we must understand students' motivations for entering STEM fields[65] and tailor the curriculum to address them. Furthermore, universities should implement courses on bias in AI and technology, similar to those offered at some medical schools,[66] as part of the curriculum for STEM majors. Finally, universities should reevaluate introductory coursework or STEM major admission requirements[67] to encourage students from underrepresented backgrounds to apply.

---

[64] B. Gault, L. Reichlin Cruse, E. Reynolds, and M. Froehner, *4.8 Million College Students are Raising Children*, Institute for Women's Policy Research, 17 November 2014.

[65] A.V. Maltese and C.S. Cooper, "STEM Pathways: Do Men and Women Differ in Why They Enter and Exit?", AERA Open, 28 August 2017.

[66] G. Friar, *Combating Bias in Medicine*, News & Research, Harvard Medical School, 7 August 2017.

[67] C. Hill, C. Corbett, A. St. Rose, *Why So Few? Women in Science, Technology, Engineering, and Mathematics*, AAUW, February 2020.

4. *To address gender disparities in society,
adopt policies that allow women to succeed
in STEM careers – but also in public policy,
law, academia, business, and other fields*

According to data from the Society of Women Engineers, 30% of women who leave engineering careers cite workplace climate[68] as a reason for doing so. Still, research suggests that consumers themselves exhibit gendered preferences for voices or robots,[69] demonstrating that gender biases are not limited to technology companies or AI development teams. Because gender dynamics are often influential both inside and out of the office, change is required across many facets of the U.S. workforce and society.

At the hiring level, recruiters must evaluate gender biases[70] in targeted job advertising, eliminate gendered language[71] in job postings, and remove unnecessary job requisites that may discourage women or other underrepresented groups from applying.[72] Even after women, transgender, and non-binary individuals are hired, companies must raise awareness of unconscious bias[73] and encourage discussions about gender in the workplace. Some companies have adopted inclusive practices[74]

---

[68] Fast Facts, Society of Women Engineers, 18 October 2018.

[69] R.C. Anderson and C.A. Klofstad (2012); E. Alexander, C. Bank, J.J. Yang, B. Hayes, and B. Scassellati, *Asking for Help from a Gendered Robot*, Yale University.

[70] T. Simonite, "Probing the Dark Side of Google's Ad-Targeting System", *MIT Technology Review*, 6 July 2015.

[71] Ibid.

[72] For example, Danielle Gauchers et al. find that when job postings for male-dominated roles use gendered language like "dominant" or "competitive," women demonstrate lower interest in the role (See: "Job advertisements that use masculine wording are less appealing to women", Gender Action Portal, Harvard Kennedy School - Women and Public Policy Program). A Hewlett Packard internal report found that women are less likely to apply for a job if they do not meet the listed qualifications (See: T.S. Mohr, "Why Women Don't Apply for Jobs Unless They're 100% Qualified", *Harvard Business Review*, 25 August 2014).

[73] R. Feintzeig, "Don't Ask Me to Do Office Housework!", *The Wall Street Journal*, 13 October 2019.

[74] J. McGregor, "How employers are preparing for a gender non-binary world", *The Washington Post*, 8 July 2019.

which could become more widespread: encouraging employees to share their pronouns, including non-binary employees in diversity reports, and equally dividing administrative work.

TABLE 3.3 - SUMMARY OF RECOMMENDATIONS TO ADDRESS GENDER AND AI BOTS

	Private Sector	**Public Sector**
Short-Term Actions	Collaborate with academic, civil society, and civil liberties groups to develop industry standards on AI and gender. Publish reports on gender-based conversation and word associations in voice assistants. Publicly disclose the demographic composition of employees based on professional position, including for AI development teams. Adopt policies that allow women, transgender, and non-binary employees to succeed in all stages of the AI development process, including recruitment and training.	Increase government support for remote learning and lifelong learning initiatives, with a focus on STEM education. Conduct research into the effects of programs like free childcare, transportation, or cash transfers on increasing the enrollment of women, transgender, and non-binary individuals in STEM education. Adopt policies that allow individuals to legally express their preferred gender identities, including by offering gender-neutral or non-binary classifications on government documents and using gender-neutral language in communications.
Long-Term Goals	Increase gender representation in engineering positions, especially AI development. Increase public understanding of the relationship between AI products and gender issues. Reduce unconscious bias in the workplace. Normalize gender as a non-binary concept, including in the recruitment process, workplace culture, and product development and release.	Decrease barriers to education that may disproportionately affect women, transgender, or non-binary individuals, and especially for AI courses. Reduce unconscious bias in government and society.

Conclusion

Discussions of gender are vital to creating socially beneficial AI. Despite being less than a decade old, modern voice assistants require particular scrutiny due to widespread consumer adoption and a societal tendency to anthropomorphize objects by assigning gender. To address gender issues in AI bots, developers must focus on diversifying their engineering teams; schools and governments must remove barriers to STEM education for underrepresented groups; industry-wide standards for gender in AI bots must be developed; and tech companies must increase transparency. Voice assistants will not be the last popular AI bot – but the sooner we normalize questioning gender representation in these products, the easier it will be to continue these conversations as future AI emerges.

4. How To Deal with AI Enabled Disinformation?

John Villasenor

Rapid disinformation attacks – i.e. attacks in which disinformation is unleashed quickly and at scale with the goal of creating an immediate disruptive effect – are one of the most significant challenges in the online ecosystem. Consider the following hypothetical scenario: on the morning of election day in a closely contested U.S. presidential election, supporters of one candidate launch a disinformation[1] campaign aimed at suppressing votes in favor of the opposing candidate in a key swing state. After identifying precincts in the state where the majority of voters are likely to vote for the opponent, the authors of the disinformation attack unleash a sophisticated social media campaign to spread what appear to be first-person accounts of people who went to polling places in those precincts and found them closed.

The attackers have done their homework. For several months prior to election day they have been laying the groundwork, creating large numbers of fake but realistic-looking accounts on Facebook and Twitter and using those accounts to regularly post and briefly comment on links to articles on local and national

---

[1] As used herein, "disinformation" refers to false information disseminated with intent to deceive. This contrasts with, and is a subset of, "misinformation," which dictionary.com defines as "false information that is spread, regardless of whether there is intent to mislead", https://www.dictionary.com/browse/misinformation

politics. The attackers have used artificial intelligence (AI) to construct realistic photographs and profiles of account owners, and to vary the content and wording of their postings, thereby avoiding the sort of replication likely to trigger detection by software designed to identify false accounts. The attackers have also have built up a base of followers, both by having some attacker-controlled accounts follow other attacker-controlled accounts and by ensuring that attacker-controlled accounts follow the accounts of real people, many of whom follow them in return.

Just after the polls open on the morning of election day, the attackers swing into action, publishing dozens of Facebook and Twitter posts complaining about showing up at polling locations in the targeted precincts and finding them closed. A typical tweet, sent shortly after the polls opened in the morning, reads "I went to my polling place this morning to vote and it was CLOSED! A sign on the door instructed me to vote instead at a different location!" Dozens of other attacker-controlled accounts "like" the tweet and respond with similar stories of being locked out of polling places. Other tweets and Facebook posts from the attackers post what appear to be photographs of closed polling stations.

Many legitimate accounts also inadvertently contribute to propagating the disinformation, as people who are unaware that it is a hoax reply to and comment on the disinformation posts. That in turn spurs additional propagation from their followers. The attackers are careful to originate the disinformation from most but not all attacker-controlled accounts; the remainder of their accounts are used to propagate it through replies and likes. The attackers know that later on in the day, once the social media companies realize what is happening and take action, this will make it harder to separate the accounts intentionally participating in the disinformation campaign from those doing so unwittingly.

Local television and radio stations quickly pick up the story, providing initial on-air and online coverage of the reported

closures. Some but not all the stations are careful to note in their reporting that the claims are as yet unverified. A few national news chains start echoing the story as well, with the caveat that they are still awaiting verification. Within about 30 minutes of the first social media postings, local reporters arrive on scene at several of the polling locations and find that there are no closures. The polls are open, voting is going smoothly, and the people waiting in line to vote express puzzlement when told about the social media claims. The local media and national quickly update their coverage to explain that assertions of closed polling places are false.

But the damage is already done. For the remainder of the day, rumors of closed polls continue to propagate through the social media ecosystem. Many voters in the precincts involved hear only the initial reports of closed polling places and not the follow-up stories declaring those reports false. For some of them, the resulting uncertainty is enough to make them decide not to vote. Many others decide to wait until later in the day to vote under the assumption that more time will bring more clarity. This creates a flood of people arriving at polling stations in the mid and late afternoon, resulting in lines with waits that rapidly grow to over an hour. Some people, unwilling or unable to wait that long, go home without voting. In the aggregate, the disinformation attack leads to tens of thousands of lost votes across the state – enough, as it turns out, to change the election outcome at both the state and national level.

The Risks of Disinformation

Hopefully, the scenario outlined above will never happen. But the fact that it *could* occur illustrates an important aspect of online disinformation that has not received as much attention as it deserves. Some forms of disinformation can do their damage in hours, or even minutes. This disinformation is easy to debunk given enough time, but extremely difficult to debunk quickly enough to prevent it from inflicting damage.

There are many domains where this can occur. Elections are one example. Financial markets, which can be subject to short-term manipulation, are another. Foreign affairs could be affected as rumors spread quickly around the world. Social movements can also be targeted through dissemination of false information designed to spur an action or reaction among either supporters or opponents of a cause.

Of course, the problems posed by online disinformation intended for short-term impact are not new. In the financial markets, the online message boards in the early days of the internet were commonly exploited by people seeking to sow and then rapidly benefit from false information about the performance of publicly traded companies. What has changed is the sophistication of the tools that can be used to launch disinformation campaigns and the reach of the platforms used for dissemination. While in the late 1990s, unscrupulous traders on financial message boards would need to manually type in the false rumors and hope that they reached a large enough group of traders to move the market, today the power of artificial intelligence can be deployed as a force multiplier allowing a small group of people to create the level of online activity of a much larger group.

Detecting Disinformation

Disinformation in all its forms is one of the most vexing challenges facing social media companies. The same false positive/false negative tradeoff that applies in many other domains applies to disinformation detection as well. If social media companies are too expansive in what they classify as disinformation, they risk silencing users who are posting accurate information about important, timely developments. If companies are too narrow in their classifications, disinformation attacks can go undetected.

Social media companies are well aware of this tradeoff. For disinformation designed to act over longer time scales, in many cases the best approach for social media companies is to act

conservatively with regard to blocking content, since the harm caused by waiting to confirm the falsity of information before blocking it is often less than the harm caused by inadvertently blocking posts by legitimate users conveying accurate information. Put another way, with disinformation that would inflict most of its damage over a longer time scale, social media companies have the latitude to take the time needed to investigate the accuracy of suspected disinformation posts while still retaining the option, if needed, to act early enough to preempt most of the damage.

Rapid disinformation attacks are particularly hard to address as they do not leave social media companies the luxury of time. Consider the election day scenario presented above. If a social media company waits several hours before concluding that the reports of closed polling places are false before they take the action of shutting down the attackers' accounts, the damage will have already been done. By contrast, taking action within minutes could preempt the damage, but that would require a confidence level and a knowledge of the accounts behind the attack that could be nearly impossible to obtain over that short a time scale. Even if confidence in the falsity of the information could be quickly obtained, there would still be the question of which accounts to block. This is especially true if, as in the election day scenario, the attack is constructed in a manner to cause legitimate accounts to unwittingly contribute to propagating the disinformation.

For unsophisticated disinformation campaigns, such as those involving a flood of copy-and-paste postings from recently created accounts with few or no followers, it is a straightforward matter for detection and mitigation software to respond rapidly. But for sophisticated attacks like the one described above, legitimate accounts and accounts created and controlled by the attackers can act very similarly. The time necessary to disentangle what is true and what is not true, and to identify which accounts are acting in good faith and which are not, is far longer than that needed for the disinformation to inflict its most significant damage.

Fortunately, the need to combat online disinformation has received increasing attention among academic researchers, civil society groups, and in the commercial sector among both startups and established technology companies. This has led to a growing number of paid products and free online resources to track disinformation. Part of the solution involves bot detection, as bots are often used to spread disinformation. (The overlap is not complete: bots are used for many other purposes as well, some nefarious and some innocuous, and not all disinformation campaigns involve bots.) One simple and easily accessible illustration is the set of tools provided by the Observatory on Social Media at Indiana University.[2] One of the tools, Botometer, "checks the activity of a Twitter account and gives it a score. Higher scores mean more bot-like activity".[3] There are also a growing number of commercial products aimed at detecting and managing bots.[4]

Bots alone are only part of the problem, as not all disinformation campaigns that use bots will be picked up by bot detection software. It is therefore also important to have tools that can look at how suspect content is impacting the broader ecosystem. Another of the tools from Indiana University's Observatory on Social Media, Hoaxy, can be used to "observe how unverified stories and the fact checking of those stories spread on public social media".[5] Hoaxy tracks online activity relating to stories and their fact checking by third parties. As useful as Hoaxy is, it does not attempt or purport to draw its *own* conclusions about the accuracy of a story. Rather, it simply gathers information about what *other sources* have said about the accuracy of a story, without exploring the extent to which those sources may themselves be accurate. The upstream

---

[2] "Misinformation Tools", Indiana University Observatory on Social Media.

[3] https://botometer.osome.iu.edu.

[4] See, e.g., "Top 10 Bot Management Solutions", EM360 Tech, 19 December 2019.

[5] HOAXYbeta. Visualize the spread of claims and fact checking, https://hoaxy.iuni.iu.edu/faq.php.

problem – and the one that is ultimately far more difficult to resolve – is to establish whether an online claim is true or false.

The Challenge of Data Labels

Responding in sufficiently short time scales to rapid disinformation attacks will require AI. But AI is not magic; for it to be effective in addressing disinformation, it needs access to data as well as to information enabling it to evaluate data accuracy. To explore this further, it is helpful to first consider how AI-based approaches can be used to detect disinformation in the absence of any time pressure, and then to address the additional complexities that arise with the need for rapid detection.

Disinformation is easiest to detect when there are large sets of "training data" that have been accurately labeled. Training data is used to enable an AI system to learn, so that it when it sees new data that was not in the training set, it knows how to classify it. Consider a drug that has been scientifically proven to be ineffective for curing Covid-19, but that many social media users and some news sites nonetheless continue to claim is a cure. A training data set can be constructed by 1) compiling and labeling as false a large number of social media posts that incorrectly assert that the drug cures Covid-19, and 2) compiling and labeling as true a large number of social media posts and news stories that correctly assert that the drug does not cure the illness. A machine learning algorithm can then learn using this training set. This corresponds to "supervised" learning, i.e. learning using a data set that has already been labeled regarding the attribute of interest. Once the training process has been completed, the algorithm will be highly effective at rapidly classifying new social media posts or news stories regarding this drug as either inaccurate or accurate.

There is also a class of machine learning techniques based on "unsupervised" learning, in which the algorithm must learn to identify categories of interest in the data without the benefit

of pre-existing labels. An example of unsupervised learning in the context of disinformation can be found in a 2019 paper titled "Unsupervised Fake News Detection on Social Media: A Generative Approach", published by researchers at Shanghai Jiao Tong University, Penn State University, and Arizona State University.[6] The authors mathematically analyze "users' engagements on social media to identify their opinions towards the authenticity of news" and use that as a basis to infer "users' credibility without any labelled data".[7]

One hurdle facing any learning algorithm, whether supervised or unsupervised, is access to a sufficiently large set of training data. Information suitable for use as training data regarding a particular issue or question can take significant time to accumulate on social media. To be useful in AI systems for detecting disinformation, the data would in many (though not all) instances require at least some degree of manual coding at the outset. Such an approach works if the topic concerned is one – like false claims regarding medical cures – for which the length of time the disinformation needs to be combated is much longer than the time needed to build and use a large set of training data. But it is far less effective for situations when disinformation defenses need to be deployed very quickly, and in which there will typically be a smaller amount of data that can be used as a basis for the algorithm to learn.

Another issue, both in rapid and in less time-constrained attempts to identify disinformation, is the accuracy of the data labels on which an AI algorithm is relying during the learning process. "Noisy" data (i.e. data in which the labels are not necessarily accurate) is a well-known problem in machine learning. To take a simple example unrelated to disinformation attacks, consider a machine learning algorithm that is attempting to learn to automatically distinguish images of cars from images

---

[6] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu, "Unsupervised Fake News Detection on Social Media: A Generative Approach", Proceedings of the AAAI Conference on Artificial Intelligence, July 2019.
[7] Ibid.

of bicycles. To do this, the algorithm might scour the internet and find millions of images that are labeled "car" and millions of other images labeled "bicycle". In most cases, those labels will be correct. But in some instances, the labels will be incorrect; an image labeled "car" might show a truck, a bicycle, or content completely unrelated to vehicles. The higher the fraction of incorrect labels, the more difficult and slower it will be for the algorithm to learn to accurately distinguish between cars and bicycles. Working with noisy data is an active area of research, and there are emerging approaches that can help mitigate if not completely eliminate the loss in accuracy that results when a machine learning algorithm learns from data in which there a is substantive fraction of labeling inaccuracies.[8]

Attempts to use AI to identify disinformation are particularly likely to encounter noisy data for the simple reason that intentional deception is involved. Most people who post an image of a car to the internet would not choose to label it "bicycle" just to throw off machine learning algorithms. But disinformation attacks will by definition be associated with a set of conflicting claims about whether online statements are true. Returning to the election day example from above, in response to a tweet falsely stating that a polling location is closed, someone who has actually just voted at that location might reply with a tweet stating that the initial tweet is false and that the polling location was in fact open. That reply is in effect a label. An account controlled by the attackers might also reply to the initial tweet by asserting that it is true. That reply is also a label, though one that directly contradicts the reply from the real voter. Over short time scales, it would be exceedingly difficult for an algorithm or a human to know which label to trust. Responding quickly to disinformation thus requires addressing the twin hurdles of limited data and unreliable and in some cases intentionally wrong labels of that data.

---

[8] See, e.g., Junnan Li, Yongkang Wong, Qi Zhao, and Mohan Kankanhalli, "Learning to Learn from Noisy Labeled Data", *arXiv preprint*, 12 April 2019.

Researchers have recognized these issues and are developing new approaches that do not rely on a large set of pre-existing training data. In April 2020, a team of researchers from Microsoft and Arizona State University posted a pre-publication version of a paper[9] describing new results on techniques for quickly detecting fake news.[10] In the paper the authors note that traditional approaches to detecting fake news "rely on large amounts of labeled instances to train supervised models. Such large labeled training data is difficult to obtain in the early phase of fake news detection".[11] To address this, the authors introduce a method that requires only a "small amount of manually-annotated clean data"[12] which can be used to rapidly and automatically label a larger set of data based on posts and comments on news articles by social media users. User credibility is one of the factors considered in forming the labels. According to the authors, this approach "outperforms state-of-the-art baselines for early detection of fake news".[13] Frameworks like this not only help solve the problem of limited data, but could potentially also help mitigate labeling accuracy issues.

As the above examples show, one common theme in research addressing disinformation is the importance of measuring the credibility of online sources. Approaches to establish and then leverage credibility will be critical to quickly identifying truth in the presence of a well-constructed rapid disinformation attack. For instance, in the election day scenario discussed earlier, it would be advantageous to give high credibility weight to the social media accounts of local television and news stations and the reporters who work at those stations. That way, as soon as those stations are able to identify that the claims of closed polling

---

[9] Kai Shu, et al., "Leveraging Multi-Source Weak Social Supervision for Early Detection of Fake News", *arXiv preprint*, 3 April 2020.

[10] As the Shu et al. paper uses the term "fake news" rather than "disinformation", in discussing that paper herein the term "fake news" will be used.

[11] Ibid., at 2.

[12] Ibid., at 3.

[13] Ibid., at 1.

places are false and disseminate that fact on social media, the AI system can calibrate truth and falsity and move to the next step of addressing the posts known to contain disinformation.[14]

It is also important to recognize the limits of what AI can be expected to accomplish. Earlier this year, Samuel Woolley of UT Austin published an excerpt of his book *The Reality Game* in *MIT Technology Review*. In the article, which was titled "We're fighting fake news AI bots by using more AI. That's a mistake", Woolley noted that "There simply is no easy fix to the problem of computational propaganda on social media".[15] It would be unreasonable to expect that any AI solution likely to be available in the near future would be able to quickly and unambiguously identify a rapid disinformation attack. However, AI will certainly be able to provide insight into the dynamics of emerging disinformation attacks, pinpoint at least some of the social media accounts at the source, and compute confidence levels regarding the likely truth or falsity of a claim making the rounds on social media. After that, the response will need to be overseen by humans making decisions based on a combination of the AI outputs and guidance from policy frameworks.

Policy Considerations

Public policy will play a central role in both the human and technological aspects of the response to rapid disinformation attacks. At the technology level, policies will need to be embedded into the algorithms to cover questions such as: What confidence level that a rapid disinformation attack is occurring should trigger notification to human managers that an activity

---

[14] Of course, the approach of identifying accounts that can serve as "trust anchors" is not foolproof. A very sophisticated attacker might be able to hack into those accounts and use them to spread disinformation.

[15] S. Woolley, "We're fighting fake news AI bots by using more AI. That's a mistake", *MIT Technology Review*, January 2020.

of concern has been identified? Over what time scales should the AI system make that evaluation, and should that time scale depend on the nature and/or extent of the disinformation? For example, suspected disinformation regarding violence should clearly receive a higher priority for immediate resolution than disinformation associated with conflicting online characterizations of what a politician said at a recent campaign speech. Other questions that can drive policies to be embedded in AI disinformation detection systems include: Under what circumstances should an AI system preemptively shut down accounts suspected of originating a rapid disinformation attack? What types of autonomous actions, if any, should be taken to address posts from legitimate accounts that unwittingly contribute to propagating disinformation?

Policy considerations will be an important driver for the human response as well. When an AI system identifies a potential rapid disinformation attack, managers at social media companies will need a set of guidelines for how to proceed. Policies can also guide the extent to which people at social media companies should arrange in advance to be "on call" to watch for rapid disinformation attacks. It is clear that for short duration, high stakes events like a national election, social media companies will need to have people standing by ready to step in and address disinformation. For events in that category, the question is not whether disinformation will be present, but rather how much of it there will be, and how sophisticated the attacks will be.

For most topics and events, there simply will not be the resources to supply staffing dedicated to individually monitoring each of the essentially limitless list of situations in which disinformation might arise. This is especially true given that companies such as Facebook and Twitter operate globally; there are literally billions of accounts in nearly two hundred countries that could potentially be employed to disseminate disinformation. As a result, for the vast majority of instances of disinformation, human intervention at the social media

companies will of necessity occur only after a problem is flagged either algorithmically or through manual reporting channels.

There will also need to be policies for handling situations in which an AI system makes exactly the wrong decision. Because of the limited data available in the early stages of a rapid disinformation attack, the need to quickly make a determination might lead an algorithm to invert truth and falsity and conclude that the disinformation is accurate and that the attempts to debunk it are themselves a disinformation attack. This is a less far-fetched outcome than it might initially appear to be. Algorithms, like the people who design them, can be influenced by a confirmation bias effect, leading to a boost in confidence in a wrong conclusion by selectively giving greater weight to inputs bolstering that conclusion. Particularly given the short time scales of rapid disinformation attacks, this could lead an algorithm to quickly converge on an incorrect conclusion that would need human intervention to identify and invert.

In short, the combination of a growing social media ecosystem and the availability of increasingly powerful AI tools for content dissemination means that rapid disinformation attacks will be a recurring feature of the online landscape. Addressing these attacks will require further advances in AI, particularly in relation to methods that can quickly assess the reliability of online sources despite the presence of very limited data. It will also require attention within social media companies to ensure that the policies and resources are in place to leverage the capabilities of disinformation detection technology, to complement that with human intervention, and to maximize the likelihood that their platforms will be used to promote, rather than undermine, access to factually accurate information.

5. AI Revolution: Building Responsible Behavior

Darrell M. West

Artificial intelligence is the transformative technology of our time. It is being deployed in healthcare, education, transportation, e-commerce and national defense, and transforming many different sectors. In a number of cases, it is relieving humans of dirty, dangerous, or boring tasks, while in others it is promoting greater efficiency and performance.

Despite its potential benefits, though, there are concerns regarding AI's fairness, bias, safety, and human control. A number of individuals worry that emerging technologies will erode privacy, raise inequality, and reduce human autonomy. Already there are fears that technology is increasing monopoly power, damaging governance, disrupting social relations, and generating a host of challenges.[1]

To avoid technology problems, we need to take a number of steps which will build trustworthy and responsible AI. They include reforms such as establishing ethical principles, strengthening oversight through AI impact assessments, creating government advisory boards, defining corporate culpability, improving digital access, reducing AI biases through third-party audits, and maintaining human control. Implementation of these steps would move the world closer to the goal of AI for the public good.

---

[1] Portions of this chapter are drawn from D.M. West and J.R. Allen, *Turning Point: Policymaking in the Era of Artificial Intelligence*, Brookings Institution Press, 2020.

Establish Ethical Principles

As AI is deployed in many areas, it is vital to have guiding ethical principles that establish values, objectives, and criteria for its development. In many respects, large technology firms have become digital sovereigns that make policy through software code and online platforms. Their decisions affect billions of users and shape much of modern life, from communications and commerce to domestic policy and international relations.

Given this situation, it is crucial to make sure corporate decisions reflect broader societal considerations. The impact of technology innovation is too pervasive to leave to coders. There are human values and social concerns that should guide innovation. Unfettered innovation without consideration of the larger impact can lead to digital creations that are unfair, biased, or unsafe.

This era requires a "whole-of-society" and "whole-of-government" approach that considers how to maximize possible benefits and minimize costs. Having broad principles and concrete guardrails allow a society to identify criteria by which to judge AI applications, think about which deployments to encourage, and evaluate the impact on humanity in general. In addition, these principles are useful in considering what policies, laws, and regulations are needed for a future that likely will be dominated by AI and emerging technologies.[2]

In looking at AI deployments, there are eight principles that are vital for preserving human autonomy and societal values in the face of the coming revolution:[3]

- *Accountability*: AI should be deployed with meaningful oversight and accountability mechanisms.
- *Fairness*: New technologies should be developed and deployed in ways that reduce rather than accentuate societal biases.

---

[2] T. Wheeler and D. Simpson, Why 5G Requires New Approaches to Cybersecurity, Brookings, 3 September 2019.

[3] Drawn from J.R. Allen, *Draft Principles on A.I. and Emerging Technology,* Brookings, 12 July 2019.

- *Human control*: Humans should maintain meaningful control over AI and emerging technologies.
- *Human rights*: AI should be designed and deployed in ways that uphold basic human rights and dignity.
- *Inclusiveness*: AI should empower a wide a set of people and communities.
- *Privacy*: AI should respect the right to individual privacy on sensitive or confidential information.
- *Safety*: AI should protect human safety and not jeopardize current protections. Before digital solutions are widely deployed, they should be evaluated for their societal impacts and risks.
- *Transparency*: AI should be developed and deployed as openly as possible.

Of course, it is not easy to put broad ideas into practice or operationalize them for policy, legal, or regulatory purposes. Individuals may differ on how to define a notion such as fairness or what happens when the principle of transparency conflicts with accountability. It is difficult to implement abstract concepts, given the range of innovations taking place around the world, but it is vital to make progress in addressing ethical concerns.

In response to these challenges, some organizations are setting up "responsible AI" offices that assist product and service designers in the operationalization of ethical principles. A broad range of experts there help coders and developers ensure that new services integrate ethics into the planning process and address important concerns. Rather than waiting for adverse consequences or bias complaints to arise, they work from the very beginning of the design process to create products that are fair, unbiased, and transparent so as to avoid problems after deployment.

To help in this regard, the Organization for Economic Cooperation and Development has established an "AI Observatory", which shares best ethical practices across member

nations. It takes an evidence-based approach to AI, monitors the manner in which AI is being deployed and affecting governance, society, and the economy, and develops ways to deal with deleterious issues.

The European Commission, meanwhile, has published "Ethics Guidelines for Trustworthy AI" and a follow-up white paper on AI policy. Its guidelines are based on a high-level expert group and an open consultation process with more than 500 contributors. Among its central tenets are: 1) applying the principles of human autonomy, harm prevention, fairness, and explicability to AI; 2) requiring AI systems to respect human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, nondiscrimination and fairness, environmental and societal well-being, and accountability; and 3) adapting trustworthy AI principles to the specific use case being deployed and "continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders in this".[4]

In the follow-up white paper, the EC moved toward a "risk-based approach" in which the regulatory framework is proportional to AI risk levels. It defined "high risk" applications based on the scope of the sector and the intended use of AI. Transportation, healthcare, energy, and parts of the government sector were identified as high-risk areas warranting close oversight and regulation, while other areas were deemed to be of lesser risk and therefore able to rely on self-regulation, voluntary labeling, and voluntary disclosure. In the latter domain, organizations would not face mandatory requirements but rather could rely on their own judgments regarding AI development and deployment.[5]

---

[4] European Commission, "Ethics Guidelines for Trustworthy AI", April 2019.

[5] European Commission, *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*, 19 February 2020, pp. 17-24.

Strengthen Oversight Through
AI Impact Assessments

For much of the past several decades, private technology companies in many countries have been given broad leeway to determine what products were developed, and how and when they were deployed. In 1996, for example, American internet companies were exempted from legal liability regarding what happened on their platforms. If someone used the network for nefarious purposes, that digital platform could not be sued.

The rationale for this approach was that technology innovation was seen as generally good for society. Lawmakers gave industry leaders considerable freedom to test new products and see what could be accomplished through new products. As a result, private companies developed innovative services, and the public generally expressed high confidence in the technology.

Policymakers bolstered this vote of confidence by having few regulations and generous tax policies that enabled leading tech companies to pay a low rate of corporate tax. There was some oversight through federal and state agencies, but most of these efforts took place under the rubric of what is called "permissionless innovation."[6] That means as long as there are few discernible problems, companies are given broad flexibility to experiment and innovate as they wished.

Today, however, individuals are not as convinced about this broad delegation of authority. Many people worry about the societal and ethical aspects of technology, and fear digital solutions that might promote inequality, encourage extremism, and threaten privacy. When asked whether technology will improve life over the next fifty years, only 22% say yes. This is a sharp drop from the 42% who felt technology had improved life over the last fifty years.[7] People are concerned about the

---

[6] A. Thierer, *Permissionless Innovation: The Continuing Case for Comprehensive Technological Freedom*, George Mason University Mercatus Center, 2014.

[7] M. Strauss, Four-in-ten Americans credit technology with improving life most in the past 50 years, FactTank, Pew Research Center, 12 October 2017.

detrimental consequences of digital technologies and robots going rogue on humans.

To deal with these issues, agencies should consider "AI impact assessments" for publicly-financed projects, in which companies with gross annual revenues over $50 million assess the impact of their AI innovations on employee conditions and human safety. The hope is these analyses, which are designed to be similar to environmental impact statements, will stimulate efforts to mitigate deleterious side effects early in the planning process.[8] This requirement would put companies and government agencies on record in terms of how they propose to deal with AI's societal ramifications.

As with its environmental counterpart, an AI impact statement would outline the positive and negative features of an AI deployment financed by the government and consider how to mitigate possible problems. Small-scale projects would be exempted from the requirement, as would AI initiatives not financed by the government. Mitigation plans in the United States would be filed with the relevant national agency (such as the U.S. Federal Trade Commission, Food and Drug Administration, Consumer Financial Protection Bureau, or Department of Transportation) and be open for public inspection.

Create Government Advisory Boards Comprising Relevant Stakeholders

With the increase in public concern about technology companies, administration officials also need to think systematically about how they deal with artificial intelligence and emerging technologies. There are many issues, ranging from the need for improved data access to addressing issues of bias and discrimination. As illustrated below, it is vital that these and other concerns be considered in an inclusive, multi-stakeholder

---

[8] J. Garcia and M. Janis, "How to Keep Robots from Taking U.S. Jobs", *Politico*, 1 May 2019, p. 17.

manner by federal agencies so that people do not end up with biased or unfair technologies.

To advance national policies, several members of the U.S. Congress have introduced the "Future of Artificial Intelligence Act" designed to establish broad policy and legal principles for AI. It proposes the secretary of commerce create a federal advisory committee on the development and implementation of artificial intelligence, comprising a diverse set of experts from government, business, and academia. The legislation provides a mechanism for the federal government to get advice on ways to promote a

> climate of investment and innovation to ensure the global competitiveness of the United States, … optimize the development of artificial intelligence to address the potential growth, restructuring, or other changes in the United States workforce, … support the unbiased development and application of artificial intelligence, … [and] protect the privacy rights of individuals.[9]

Among the specific areas the committee would be asked to address are competitiveness, workforce impact, education, ethics training, data sharing, international cooperation, accountability, machine learning bias, rural impact, government efficiency, investment climate, job impact, bias, and consumer impact. The committee is directed to submit a report to Congress and the administration 540 days after enactment regarding any legislative or administrative action needed on AI.

In the national security area, an AI commission already has been established and produced a report. Authorized by the 2019 National Defense Authorization Act, this commission seeks to determine the risks of AI technology on the battlefield, international advances, and ways to promote better usage in the United States. It has representation from academia, government, and business, and has recommended a number of

---

[9] Congress.gov, "H.R. 4625 - FUTURE of Artificial Intelligence Act of 2017", 12 December 2017.

steps designed to advance national security through AI, ML, and data analytics.[10]

In the international arena, French President Emmanuel Macron has suggested an "international experts council" that could offer guidance on possible reforms. He argued "sometimes we will need to set limits to innovation and give rules so that community can be preserved".[11] In doing so, he noted the risk of instability arising from technology innovation and the need for companies, governments, and experts to work together to avoid major problems.

Define Corporate Culpability

Others have suggested reforms designed to increase culpability on the part of digital platforms. In 2018, American internet sites were made legally liable for human trafficking that takes place on their sites. The Stop Enabling Sex Traffickers Act holds firms liable for sex crimes that occur using their platforms.[12] This landmark shift arose out of policymaker concern that, free of lawsuit risk, tech companies simply looked the other way when unethical behavior occurred.

This is in keeping with the recommendation of legal scholars Danielle Citron and Benjamin Wittes, who argue it is time to rethink the current legal immunity accorded to digital platforms. In looking at the dramatic proliferation of online porn, human trafficking, and sexual abuses of many sorts, they argue policymakers should require websites to take "reasonable steps to prevent or address unlawful third-party content that

---

[10] J. Doubleday, "Chairman Pallone Taps Former FCC Cmr. Mignon Clyburn for National Security Commission on Artificial Intelligence", *Inside Defense*, 11 January 2019.

[11] E. Sugiura, "French President Calls for Council to Help Officials Regulate AI", *Nikkei Asian Review*, 27 June 2019.

[12] T. Jackman, "Trump Signs 'FOSTA' Bill Targeting Online Sex Trafficking, Enables States and Victims to Pursue Websites", *The Washington Post*, 11 April 2018.

it knows about."[13] In their view, such a step would increase accountability and help to root out abusive web behavior.

Similarly, some legislators have suggested it is time to extend legal liability to other issues as well. For example, Senator Joe Manchin (D-W.Va.) has argued that internet platforms should be accountable for problematic drug sales that take place on their sites. He points particularly to the opioid crisis as one that has been enabled by lax enforcement of laws regarding painkillers such as oxycodone and fentanyl, which have been widely abused. Complaining at a committee hearing to Facebook and Twitter executives, he argued, "A lot of people have been affected and a lot of people have died receiving information about how to obtain drugs through y'alls platforms".[14]

State attorneys general have called for greater enforcement authority over internet platforms. In a 2019, letter, the members of the National Association of [State] Attorneys General wrote to leading national legislators requesting amendments to the 1996 Communications Decency Act (CDA) that would allow them "to take appropriate action against criminal actors." They claim courts have interpreted CDA Section 230 legal exemptions too broadly and thereby made it impossible to hold digital businesses accountable for bad behavior. The attorneys general asked for the ability to investigate "online black market opioid sales, ID theft, deep fakes, election meddling, and foreign intrusion".[15]

As digital abuses have proliferated, many government regulators have started to step up their enforcement actions using existing laws. As a result, a number of technology companies have incurred millions of dollars in fines over privacy invasion, anti-competitive practices, or discriminatory behavior. Using

---

[13] D. Citron and B. Wittes, "The Problem Isn't Just Backpage: Revising Section 230 Immunity", *Georgetown Law Technology Review*, 2018, p. 456. Also see J. Nicas, "Sex Trafficking via Facebook Sets Off a Lawyer's Novel Crusade", *The New York Times*, 3 December 2019.

[14] C. Lima, "Tech Takes on Opioids", *Morning Tech*, *Politico*, 24 October 2018.

[15] Letter from the National Association of Attorneys General, 23 May 2019.

the powers of the Federal Communications Commission, Federal Trade Commission, and Department of Justice, as well as state attorneys general and the European Union, government officials are holding firms responsible for bad behavior. The old model of trusting technology corporations to deploy digital tools for the common good is giving way to greater skepticism and enforcement action.[16]

Through these and other actions, decisionmakers are using government policies and regulations to restrict anti-competitive business practices and improve public oversight of technology firms. In so doing, officials are starting to treat the tech sector like other business enterprises. They are encouraging actions that promote innovation and enacting rules to discourage practices that lead to unfairness, bias, inequity, or poor safety. Their goal is to put tech innovation on a productive course that balances the need for innovation with policies that promote best practices.

Improve Digit Al Access

During an era of advancing technologies, it is crucial to improve digital access. All people need to be in a position to participate in the digital economy, learn through online education, take advantage of developments in health information technology, and be able to communicate with other people.

These benefits are particularly valuable during times of crisis, such as with the recent coronavirus pandemic. It is dangerous to restrict internet access during a pandemic because it robs people of access to needed information. This has happened in countries such as India (especially in the Kashmir area), Ethiopia, Myanmar, and Bangladesh, among other places. Governments there either shutdown the internet or restricted mobile networks as a way to control local populations or keep political adversaries from mobilizing opposition forces.[17]

---

[16] C. Henrickson and W. Galston, Big Tech Threats: Making Sense of the Backlash Against Online Platforms, Brookings, 28 May 2019.

[17] A. Kumar, Mitigate Risks of Covid-19 for Jammu and Kashmir By Immediately

It is easier to maintain good health and social order if people do not panic during pandemics and are able to learn from the experience of other individuals. The internet allows ordinary folks to see how hospitals, doctors, and patients are dealing with the crisis. It gives students and teachers access to the latest educational materials. And as more commerce moves online, digital platforms become prevalent as places for business, trade, and employment. Governments need to recognize that internet crackdowns and slow broadband speeds can make the pandemic worse by denying sick people the opportunity to find needed information, see how care providers are dealing with the disease, and enable schools to educate young people.

Reduce AI Biases Through Third-Party Audits

Ethics, bias, and discrimination are serious issues for AI. There already have been a number of cases of unfair treatment linked to historic data, and steps need to be undertaken to make sure bias does not become prevalent in artificial intelligence. Existing statutes governing discrimination in the physical economy need to be extended to digital platforms. This will help to protect consumers and build confidence in these systems as a whole.

As an illustration, the U.S. Equal Employment Opportunity Commission (EEOC) uses an "80% rule" which "prescribes that if the probability of a positive outcome for members of a protected class is less than or equal to 80% than for a non-protected class, there's a disparate impact".[18] The idea is algorithms should not have a detrimental impact on protected individuals that is more than 20% at variance from non-protected individuals.

Restoring Full Access to Internet Services, Amnesty International, 19 March 2020.

[18] C. Bousquet, Algorithmic Fairness: Tackling Bias in City Algorithms, Ash Institute, Harvard University, 31 August 2018.

Using this and other kinds of standards, Nicol Turner-Lee, Paul Resnick, and Genie Barton argue in favor of third-party audits that scrutinize algorithmic data and outcomes. The features warranting assessment include "the algorithm's purpose, process and production".[19] According to them, audits should be undertaken by independent groups and triggered when there are apparent violations of the EEOC's 80% rule on disparate impacts. Focusing on unfairness in what comes out of algorithms increases the odds of more equitable AI applications.

Experts at the Future of Life Institute propose a series of AI technical standards designed to promote fairness and safety. In a letter submitted to the National Institute for Standards and Technology, the experts said the world needs "safe and trustworthy research, development, deployment, and use of AI technologies across sectors".[20] To move in that direction, they argue, there should be technical standards that detail AI explainability, safety, and trustworthiness.

Improve Data Access

Countries need to develop a data strategy that enables fair and unbiased AI. Having data that enable evidence-based algorithms is crucial for future development. Right now, there are no uniform standards in terms of data access, data sharing, or data protection, and this creates particular challenges for AI.[21] Much of the data that go into algorithms are proprietary in nature and not shared very broadly with the research community, thereby limiting innovation and system design. Artificial intelligence requires large data sets to test and improve its learning

---

[19] N. Turner-Lee, P. Resnick, and G. Barton, Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms, Brookings, 22 May 2019.

[20] Future of Life Institute, NIST letter on "Developing a Federal AI Standards Engagement Plan", 6 June 2019.

[21] F. Castanedo, *Understanding Data Governance: Practices and Frameworks for Regulatory Compliance and Security*, O'Reilly Media, 2018.

capacities.[22] Without access to structured and unstructured data sets, it will be difficult to gain the full benefits of AI.

In general, the research community requires better access to government and business data in all areas, both with AI and standard statistical modeling.[23] There are a variety of ways researchers could gain data access. One is through voluntary agreements with companies holding proprietary data. Facebook, for example, announced a partnership with Stanford University economist Raj Chetty to use its social media data to explore inequality.[24] As part of the arrangement, researchers were required to undergo background checks and could only access data from secured sites to protect user privacy and security.

For a long time, Google has made search results available in aggregated form for researchers and the general public. Through the Google Trends site, scholars can analyze topics such as interest in Donald Trump, views about democracy, and perspectives on the overall economy.[25] That helps researchers track movements in public interest and identify topics that galvanize the general public.

Twitter makes most of the tweets posted to its platform available to researchers through application programming interfaces, commonly referred to as APIs. These tools help people outside the company build application software and make use of data from its social media platform. They can study patterns of social media communications and see how people are commenting on or reacting to current events.

---

[22] Executive Office of the President, Artificial Intelligence, Automation, and the Economy, December 2016; and Executive Office of the President, National Science and Technology Council, Committee on Technology, Preparing for the Future of Artificial Intelligence, October 2016.

[23] B. Gansky, M. Martin, and G. Sitaraman, "Artificial Intelligence Is Too Important to Leave to Google and Facebook Alone", *The New York Times*, 20 November 2019.

[24] N. Scolar, "Facebook's Next Project: American Inequality", *Politico*, 19 February 2018.

[25] D.M. West, What Internet Search Data Reveals about Donald Trump's First Year in Office, Brookings, 17 January 2018.

In some sectors where there is a discernible benefit from data openness, governments can facilitate collaboration by building infrastructure that shares data. For example, the U.S. National Cancer Institute has pioneered a data sharing protocol where certified researchers can query health data using de-identified information drawn from clinical data, insurance claims information, and drug therapies. It enables researchers to evaluate efficacy and effectiveness, and make recommendations regarding the best medical approaches, without compromising the privacy of individual patients.

The Center for Data Innovation has proposed that federal agencies should develop "shared pools of high quality, application-specific training and validation data in key areas of public interest, such as agriculture, education, healthcare, public safety and law enforcement, and transportation." That type of information would facilitate pilot testing and help researchers refine AI models. The center cites other countries, such as France, which require "the private sector to share certain data sets in select circumstances, when it does not threaten a firm's business and relates to key public interests such as health and safety".[26]

In 2019, the U.S. federal government took major strides in making public data available with the enactment of the OPEN Government Data Act. It mandated that federal agencies publish public sector data in machine-readable and nonproprietary formats, and that every agency hire a chief data officer.[27] The goal of this legislation is to improve data accessibility and have data be in a form that is easily usable by researchers, businesses, journalists, and the general public.

A combination of these approaches would improve data access for researchers, the government, and the business community without impinging on personal privacy. As noted by Ian Buck,

---

[26] J. New, *Why the United States Needs a National Artificial Intelligence Strategy and What It Should Look Like*, Center for Data Innovation, December 4, 2018, p. 4.

[27] D. Coldewey, "Transparency-Seeking OPEN Government Data Act Signed into Law", *Tech Crunch*, 18 January 2019.

the vice president of Nvidian, "Data is the fuel that drives the AI engine. The federal government has access to vast sources of information. Opening access to that data will help us get insights that will transform the U.S. economy".[28]

Improve Mechanisms for Human Control

Many experts have argued there needs to be avenues for humans to exercise oversight and control of AI systems.[29] There has to be mechanisms to avoid unfairness, bias, and discrimination. In addition, companies and government agencies must understand there are audit risks, third-party assessments, and costs for noncompliance with existing laws and cherished human values.

For these reasons, the chief executive officer of the Allen Institute for Artificial Intelligence, Oren Etzioni, argues there should be rules for regulating algorithmic systems. First, he says, AI must be governed by all the laws that already have been developed for human behavior, including regulations concerning "cyberbullying, stock manipulation or terrorist threats," as well as "entrap[ping] people into committing crimes". Second, he believes that these systems should disclose they are automated systems and not human beings. Third, he states that, "An AI system cannot retain or disclose confidential information without explicit approval from the source of that information".[30] His rationale is that these tools store so much data that people have to be cognizant of the privacy risks posed by AI.

In the same vein, the IEEE Global Initiative has established ethical guidelines for AI and autonomous systems. Its experts suggest that models be programmed with consideration for

---

[28] I. Buck, Testimony before the House Committee on Oversight and Government Reform Subcommittee on Information Technology, 14 February 2018.

[29] A. Dahlmann and M. Dickow, "Preventive Regulation of Autonomous Weapon Systems", Berlin, Stiftung Wissenschaft und Politik, March 2019.

[30] O. Etzioni, "How to Regulate Artificial Intelligence", *The New York Times*, 1 September 2017.

widely accepted human norms and rules for behavior. Algorithms for AI need to take into account the importance of these norms, how norm conflict can be resolved, and ways these systems can be transparent about norm resolution. Software designs should be programmed for "nondeception" and "honesty," according to ethics experts. When failures occur, there must be mitigation mechanisms to deal with the consequences. In particular, AI must be sensitive to problems such as bias, discrimination, and fairness.[31]

The world is on the cusp of revolutionizing many sectors through artificial intelligence, machine learning, and data analytics. There already are significant deployments in finance, national security, healthcare, criminal justice, transportation, and energy management, and smart cities that have altered decision-making, business models, and system performance. AI is being utilized in virtually every sector and transforming the way people communicate, buy goods and services, undertake transactions, and learn from one another.

It will be impossible to answer major questions regarding human safety, individual freedom, national security, and societal well-being without evaluating the specific manner in which AI is being implemented. If leaders make appropriate policy and operational decisions of the sort recommended in this book, we are quite optimistic about our AI future. In many areas, advanced technologies will improve medical care and education, help seniors and the disabled gain mobility, promote social and economic opportunity, and safeguard national defense. However, if leaders fail to make the right policy choices, the world could disintegrate into stark inequality, a lack of personal privacy, widespread unfairness, and political authoritarianism. Leaders must take meaningful steps toward addressing these problems and make sure that we protect fundamental human values.

---

[31] *Ethical Considerations in Artificial Intelligence and Autonomous Systems*, unpublished paper, IEEE Global Initiative, 2018.

6. Pacem in Cyberspace, Auspicia Algoretichs

Paolo Benanti

Proper ethical debate should take into account all those criteria that favor or guide technological innovations toward the common good.

It is essential that we realize the need to create bodies or institutions that guarantee the governance of artificial intelligence (AI) technologies. The objective search for what is good can only take place if institutional venues exist for such ethical dialog and for regulating such technologies. In a paradigm of a sincere, objective searching for the truth, the only way we can truly face and manage the complexity of the technological world, with the various related problems, is if reflections and debate on ethical discernment are supported by a political structure that truly has the power to manage artificial intelligence technologies. The alternative, at best, is to draft proposals or evaluations that are mere words without a corresponding objective reality, thus lacking historical effectiveness.

Fear of the Uncertain

On one level, ethical reflection looks at the person as an individual who might interact with AI or be part of its knowledge. The different positions revolve around the unpredictability of the effects of some technologies, the ability of humans to control technology and the effects these technologies can have on an

individual. Such reflections can be grouped under the general definition of *fear of the uncertain.*

Ethical assessments grouped under this definition can be further classified according to a twofold understanding of the term *uncertainty.* Some understand the term *uncertainty* as referring to technological development and giving rise to what can be defined as *technological uncertainty.* Others understand the term *uncertainty* as referring to the fate of human evolution, leading to *evolutionary uncertainty.*

Equality and the Pursuit of Happiness

A second area of ethical reflection is about the type of social relations that would be created between subjects endowed with *technological synthetic* cognition and other members of society. The right of individuals to pursue their own happiness is stressed, on the one hand, and the right to equality among members of the same group or nation, on the other. Here, reflections based on the tension between individuals and society are grouped under the general definition of *equality and the pursuit of happiness.* While the previous arguments considered the person as an individual and tried to protect his/her fundamental rights and not expose him/her to excessive risks, this type of assessment looks more closely at the relationship between the individual and society, wondering what kind of relationships could be created between improved humans and the rest. Many of these arguments are based on two rights considered inviolable: the right to equality and the right to seek one's own happiness. In such an examination of the topic, AI is legitimate to the extent it allows humans to remain equal and favors the inalienable desire of each person to seek his/her own happiness. In this perspective, topics such as security, justice, informed consent, psychological aspects and respect for the autonomy of individuals are addressed, but these problems are ultimately attributed to arguments that show how AI technologies ensure or deny equality between individuals or the pursuit of happiness.

Policies

The management of technology[1] and its development in the near future consequently requires political and economic management. Such management is usually referred to as governance,[2] a term that refers to the existence of a new way of organizing and administering territories and populations.[3]

A two-way link exists between governance and development. On the one hand, using the term development alongside the term governance means placing the focus of social living, as a purpose, on the person. At the same time, stating development requires governance means taking on the ethical dimension not as a juxtaposed element in managing and addressing *technological innovation*, but rather as recognizing this brings a series of questions about meaning that are at the core of all authentic development.

So, the authentic governance of technology will not be based on moral considerations that are placed

> on the margins of development and ... [take concrete form] through the development of corrective tools, both at an individual (or at least private) and an institutional level… [but will seek] to become effective, including for production, through action that

---

[1] In this chapter the use of the word technology is understood in the broad sense and covers technique as well.

[2] The Anglo-Saxon term governance derives from ancient French and, lacking a corresponding term in Italian, the original language of this text, in the last twenty years has become popular in political and academic debate and tends to replace the use of the term *government* (Cf. Organisation for Economic Co-Operation and Development (OECD), *OECD Economic Glossary. English-France*, Paris 2006, p. 236).

[3] The same definition of the notion of governance has undergone changes and additions, although in general it can be argued that economists, political scientists and experts in international relations have used it, first of all, to highlight a distinction and a juxtaposition with *government* understood as an institution, apparatus and organization (cf. ibid, p. 236 and Commission of the European Communities, "*European* Governance, A White Paper", *Official Journal of the European Union* COM(2001) 428 final, pp. 1-29).

involves both individuals and groups, without limiting itself to a specific industry, and focuses on the person as a whole.[4]

Within the various meanings of the term, the governance of development becomes how correct governance is adopted and practiced in the light of those ethical assessments of the world of technology rooted in the *Social Doctrine of the Church,* which animates ecclesial reflection on a believer's actions in the material world. Governance is the space where anthropological and ethical considerations, through a process of mutual exchange and dialog, must become effective forces for shaping and guiding *technological innovation*, making it an authentic source of human development. This space for political and economic action, which constitutes the governance of technology, becomes a compulsory call to conscience and so must be translated into a commitment to the governance of technology.

The very nature of *technological innovation* makes is clear that governance will only be effective if it is conceived as a moment of dialog and comparison among the different skills provided by the empirical sciences, philosophy, moral-theological analyses and every other form of human knowledge that touches on this topic.[5] The different stakeholders must also interact in a constructive, coordinated manner. For example, institutions, academia and technology companies should reflect jointly on how AI governance can be implemented to ensure it is effective and makes it possible to fully take advantage of the opportunities provided.

Following the logic of these considerations, the role of ethical reflection in this process of governance is not primarily about directly identifying technical solutions to the various problems,

---

[4] P. Lacorte, G. Scarafile, and R. Balduzzi (eds.), *La* governance *dello sviluppo: etica, economia, politica, scienza*, Editrice AVE, 2004, p. 43.

[5] Cf. A. Rigobello, "Dinamiche interne ad un'etica coinvolta nella governabilità dello sviluppo", in ibid., pp. 43-48 and S. Latouche, "Altri mondi sono possibili, non un'altra mondializzazione", in ibid.

but – in debate – about clearly highlighting the critical question as to what is the sense of a human being mediated by *technological innovation* and the means that can guarantee authentic human development.

This level must act as the trigger so stakeholders ask a series of question so as to develop a legal framework and other tools that keep AI and related technologies within limits that express the foundational principles and values of democratic society.

The questions that follow try to express some vital issues on which these parties must focus in defining their legislative, administrative, judicial or regulatory action. To some extent, these questions might already form part of policymakers' operational and decision-making practices. The questions are presented with a focus on resilience. They are not designed as a "checklist" for policymakers and regulators, but rather as a trigger mechanism to cause the self-questioning of societies and institutions in a manner than can modify social practices and dynamics. They will also enable policymakers to consider the broader consequences of digital innovation.

1. Is AI necessary, legitimate, transparent and proportional in its use and implementation? How are these judgments made? Are there alternative practices that entail less problematic solutions?

2. How has the decision to use AI weighed up costs, benefits and risks, including the consequences on human rights, freedoms and democracy? Is the decision-making process publicly documented?

3. Given the transformative power of AI, how has a system of verification and protection of human rights been implemented?

4. How can we promote the dissemination of relevant information to expose official and corporate conduct to public scrutiny and implement other ways to increase responsiveness to public concerns?

5. What decisions have been made regarding the need and proportionality of intrusion into the private lives of

individuals through the use of personal data collected anonymously but able to influence the lives of those involved in the implementation? Is the decision-making process transparent?

6. How have the opinions of various stakeholders, in particular the public, been taken into account?

7. Since policymakers have identified potential damage, who is damaged and who benefits from these technologies? What are the potential knock-on effects? What are the consequences for society? After trying to identify all the consequences, have policymakers thought of reasonable solutions to deal with such damage?

8. On what areas do political decision-makers focus and what auditing practices are periodically put in place to review existing legislation and practices to ensure adequate responsiveness to changing circumstances and to implement and enforce current legislation?

9. What systems are in place to ensure adequate supervision, review and monitoring of implemented AI technologies?

10. How does the protection of privacy balance with public security and the interests of companies?

11. How do the implemented technologies help to ensure adequate cybersecurity?

12. Have the subjects of these technologies (which can generally be understood as the public) been informed of the existence and general purpose of the AI system? How can they learn more about the system's scope? How can they seek personal compensation for any damage suffered? How can they question or substantially challenge the implemented system?

13. How is accountability controlled, managed and implemented?

14. How can the political and decision-making process best manage the proliferation of AI?

15. How can society attain the required skills to face the challenges and changes related to AI?

16. How should the effects of using AI be continuously assessed or monitored?
17. How can international regulatory standardization and cooperation best meet the challenge of the global flow of personal information?
18. How can policymakers and regulators cooperate to promote digital innovation, maximizing good practice at the international level?

Organizational Charter - Behavior of Developers and Companies

Organizations are key players in contemporary economics, where the term organization means an entity – such as a company, an institution, or an association – comprising one or more people and having a particular purpose.

An organization that is established as a means for achieving defined objectives has been referred to as a formal organization. Its design specifies how goals are subdivided and reflected in subdivisions of the organization. Divisions, departments, sections, positions, jobs, and tasks make up this work structure. Thus, the formal organization is expected to behave impersonally in regard to relationships with clients or with its members. According to Weber's definition, entry and subsequent advancement is by merit or seniority. Each employee receives a salary and enjoys a degree of tenure that safeguards him from the arbitrary influence of superiors or of powerful clients. The higher his position in the hierarchy, the greater his presumed expertise in adjudicating problems that may arise in the course of the work carried out at lower levels of the organization. It is this bureaucratic structure that forms the basis for the appointment of heads or chiefs of administrative subdivisions in the organization and endows them with the authority attached to their position.[6]

_____

[6] J. Morin, *Leadership and Change Management*, October 2018.

Organizational culture is defined as the underlying beliefs, assumptions, values and ways of interacting that contribute to the unique social and psychological environment of an organization.

Organizational culture includes an organization's expectations, experiences, philosophy, as well as the values that guide member behavior, and is expressed in member self-image, inner workings, interactions with the outside world, and future expectations. Culture is based on shared attitudes, beliefs, customs, and written and unwritten rules that have been developed over time and are considered valid.

How can the elements that emerge from policy questions become an integral part of an organization?

The philosophy of technology, mainly thanks to the *empirical turn*, helps us to look at technology as a multidimensional reality. In such an approach, we can identify three different levels on which ethical reflection must focus.[7]

On the first level, technology can be seen as a means or an activity aimed at a purpose. At this level, we analyze technological artifacts as extensions of human capabilities and technique is a way to transform the surrounding real world. In such a conception, a technological artifact has no ethical dimension, as the artifact's only purpose is efficiency, with human will determining its correct use.[8]

---

[7] These three levels are not to be understood as direct extensions of those three stages that characterize the philosophy of technology in our analysis. They are an expression of the understanding of culture and technique-technology, which, thanks to those analyses, has begun to elaborate a new vision of society. See, for example, the reflections by E.H. Schein, who represents the mature expression of this transformation. Schein, with his vision of culture as a three-dimensional reality, is the author most influencing the reorganization of large U.S. corporations. The corporations seem to search in Schein's work for tools to take on an ethical profile that avoids the recurrence of the financial crisis started in 2008 (cf. E.H. Schein, *Organizational Culture and Leadership*, Jossey-Bass, San Francisco (CA), 2004, in particular pages 30-44).

[8] Cf. B. Mondin, *Manuale di filosofia sistematica. Antropologia filosofica*, Edizioni Studio Domenicano, 2006, pp. 194-195.

On the second level, techniques can be seen through their ability to transform areas of human life. At this level, technological artifacts are like containers that intrinsically *transmit* a certain *technological intentionality*, a way of understanding reality and responding to it that is able to influence the user toward a certain purpose.[9] This level of understanding technology has two ethically relevant consequences. First, technology has no morally neutral use so the *technological intentionality* that every artifact possesses requires free and conscious responsibility, and continuous ethical discernment.[10] Secondly, technology possesses a degree of ambiguity, which Ihde calls *multistability*. Since, as already seen, technology is closely connected to its being situated in a culture, every technological artifact, at this level, does not have an essence of its own, but acquires meaning in its use in a given context.[11] Ethics contributes greatly to the exercise of free and conscious responsibility, and to the exercise of ethical discernment capable of allowing the *technological intentionality* of the technique-technology to emerge. In particular, the style of *ethical dialog* helps believers and non-believers in the search for the good that is desired and understood.

Finally, on the third level, technology expresses a basic attitude of human beings toward the world. Technological artifacts represent the way in which a culture expresses itself and organizes itself in a given time and place. Buildings, technologies, artistic achievements and all the other products of human activity constitute the visible level of a culture. This level exists

---

[9] Cf. P.-P. Verbeek, "Don Ihde: The Technological Lifeworld", in H. Achterhuis (ed.), *American Philosophy of Technology. The Empirical Turn*, Bloomington and Indianapolis, Indiana University Press, 1997, p. 136.

[10] This is particularly evident if we approach border issues such as those raised by neuroscience or neuroethics (cf. P. Benanti, "Neuroenhancement in Young People: Cultural Need or Medical Technology?", *American Journal of Bioethics Neuroscience*, vol. 1, no. 1, 2010, pp. 27-29 and P. Benanti, "From Neuroskepticism to Neuroethics: Role of Morality in Neuroscience That Becomes Neurotechnology", *American Journal of Bioethics Neuroscience*, vol. 1, no. 2, 2010, pp. 39-40.

[11] Cf. P.-P. Verbeek (1997).

and is made real by virtue of a series of value judgments that are shared by a community. Each structured social group has a vision of the *common good* defined by the specific purposes that determine and govern the group itself.[12] This vision of a specific common good is the basis for structuring internal relationships within individual spheres of sociality. Such common good is also the basis for assessing the roles of individuals, and mutual expectations are formed.[13] The ethical question goes through the whole process of formation, progressive structuring and eventual modification that shapes the various levels and concrete forms of sociality; it questions the correctness, in terms of human authenticity, of the aims pursued and the ways of pursuing them.[14] This shared regulatory level constitutes the *ethos* of a population: it is an expression of the morality of individuals in the structuring of history.[15]

On this third level, technique tells of the hierarchies of a culture's values and an expression of the lived morality – the *ethos* – of society[16]:

> though the essence of a group's culture is its pattern of shared, basic taken-for-granted assumptions, the culture will manifest

---

[12] Cf. S. Bastianel and G. Parnofiello, *Moralità personale nella storia*, Roma, PUG, 2005, pp. 303-323.

[13] Cf. D. Abignente and S. Bastianel, *Le vie del bene. Oggettività, storicità, intersoggettività*, Il pozzo di Giacobbe, 2009, pp. 51-95; S. Bastianel and G. Parnofiello (2005).

[14] Cf. S. Bastianel and G. Parnofiello (2005).

[15] Cf. D. Abignente and S. Bastianel (2009); S. Bastianel and G. Parnofiello (2005).

[16] A verification of this type of analysis comes from the history of the Wes. Think of the *Mounts of piety* of the Middle Ages, which were an expression of a solidarity lived within a municipal context, or the *lagers*, the extermination camps that the Nazis established for the systematic elimination of all those who were deemed unworthy of living or in conflict with the national-socialist regime. Both of these artifacts can be read through the three levels we have described: as tools to achieve a goal with greater efficiency; as carriers of a technological intentionality that challenges the exercise of personal morality, the exercise of free and conscious responsibility in its role as mediation towards reality and, finally, as an expression of the values understood and experienced within a culture.

itself at the level of observable artifacts and shared espoused beliefs and values. In analyzing cultures, it is important to recognize that artifacts are easy to observe but difficult to decipher and that espoused beliefs and values may only reflect rationalizations or aspirations. To understand a group's culture, one must attempt to get at shared basic assumptions and one must understand the learning process by which such basic assumptions come to be.[17]

It is clear that technology expresses a basic attitude of human beings toward the world: technological artifacts both represent the way in which a culture expresses and organizes itself in a given time and place, and the vision humanity has of itself and the world, and how it tries to make its deepest desires come true. In other words, technology is always, in a certain way, linked to a reference anthropology. As Schein pointed out, the anthropological vision that underlies the development of technique-technology is the most difficult to decode by analyzing only technological artifacts, but this does not mean that it is less effective in shaping culture and society.[18]

This awareness triggers the creation of a corporate organizational culture. The organizational culture is, in Edward Schein's vision, the coherent set of fundamental assumptions that a certain group invented, discovered or developed while learning to deal with the problems related to its external adaptation or its internal integration, and which worked so as to be considered valid and therefore worthy of being taught to new members as the correct way to perceive, think and feel in relation to these problems. Organizational culture is a concept that bears multiple meanings, but it is the way in which a model can intersect with sociological, anthropological and ethical perspectives. Culture refers to the definition of strategy as a perspective: shared vision of the world, ideology, identity, paradigm, and business theory. Culture can be seen as

---

[17] E.H. Schein (2006), p. 36.
[18] Cf. E.H. Schein, *Organizational Culture and Leadership*, San Francisco (CA), Jossey-Bass, 2004, p. 36.

an important aspect of an organization's strategy and as a source of superior performance.

Since organizational culture can be seen as the glue that holds an organization together through the sharing of patterns of meaning, implementing ethically meaningful behavior in an organization requires the creation of an adequate organizational culture, such as, creating patterns of meaning and modes of behavior that confirm those instances that are fundamental to obtain what is sought. In fact, it is the organizational culture that:

1. Creates a sense of identity
2. Facilitates collective commitment
3. Promotes stability of the social system
4. Defines interpretative schemes
5. Acts as a control mechanism

Only if these issues become part of the organizational culture is it possible to take what the management identifies as strategic actions and identify the company's homogeneous, functional action. This turns the fundamental ethical level toward the developers' behavior and the features of the technological products that have been developed. The management's understanding and intention become what gives shape to the technological artifacts produced. Hence, the ethical assumptions will be able to become a set of values but also the genetic structure of the artifacts produced.

It is important to be clear there is no non-culture. Psychology has shown one cannot avoid communication and even those who do not want to communicate actually do communicate. Similarly, an organizational culture is created even without any explicit will to do so by the top leaders at a company. This is why we always need to know, understand and, if necessary, modify, strengthen, and support such culture. In the specific case of AI, the execution of systems with such a great social impact and the technological nature of this transformation require specific attention. The absence of a corporate culture would mean the

proliferation of micro-cultures in the various departments and sectors of management. The execution of a complex system, like an AI, requires an alignment of intentions and shared values in the different parts of a company involved in its creation. Think of the case of the risk assessment algorithms for the judicial world. Here, the absence of a series of shared values has led to the creation of algorithms that are technically correct and work, but which do not actually express or achieve the basic assumptions and values of justice they justify, and to the service of which their own development is intended.

Thus, the absence of an organizational culture does not result in a vacuum but in a series of cultures that express, mediate and give strength to different, multiple or even conflicting values and assumptions. Should this happen for technological development, we would be faced with unpredictable and uncontrollable system-level results. In the scenario of the pervasive introduction of AI for facial recognition, the unpredictable outcome, up to the most dreadful dystopian scenarios, is simply the proof that, as mentioned earlier, a technological artifact is a difficult level of decoding that can be the expression of extremely different values and assumptions. The only way to guarantee the presence of some fundamental values is to make them alive and present in an organizational culture. Of course, the presence of an organizational culture alone does not guarantee the desired result, but its absence makes it impossible to achieve.

The newness of the technological field in question adds another key element. Producing technological solutions with AI requires business decisions that must follow the development of the technological sector, social sensitivities and actual opportunities. If such decisions are not part of a culture - that is, a living organism that adapts and responds to internal and external changes - they risk becoming guidelines that, like words carved into a rock, cannot adapt to changing conditions. Solutions that are rigid and immutable, and solutions that do not take shape in an organized culture are in danger of producing

the opposite effects in fast changing scenarios. The AI market is ready for companies to enter it, knowing how to adapt to its constant transformations. Doing this as an organization requires overcoming an atomized model to acquire an organic model. Only an organizational culture can produce an organic and symphonic response.

Furthermore, the creation of high-tech products requires the work of different parts and components of the company entrusted with implementing the various means necessary to achieve the goal. Organizational culture offers the security of never confusing means with purposes as the development of each component of an AI solution is a means that is oriented toward the organization's unique, univocal purpose. If an adequate culture is not created, individual means risk becoming purposes. If this were to happen, the AI system would no longer guarantee the ethical and security requirements set out or desired by the organization that develops it. In AI, the system being "intelligent" is the result of the interaction of its different components. The result is not merely something we can reduce to the sum of its parts. This makes it necessary for every single part to look at the whole picture, understood not only as the produced artifact, but also the global purpose that it intends to achieve. This dynamic is the operational expression of a culture in action. Then, in the AI world, an ethically satisfactory result is only conceivable as the desired result of an organizational culture, since this is the tensor that holds together and directs the work of the parts. Many AI solution, today, have a degree of opacity precisely because of this absence of an organizational culture that generates them. They are great technical prodigies that, though, could cause unwanted or problematic scenarios, because the purpose they can pursue is not an explicit expression of the group that implemented and made them. Only the presence of an explicit organizational culture binds the means to the desired end. Only the dissemination of this organizational culture allows institutional partners and the company to clearly see the aim pursued and to conceive the

means – the technological artifacts – as tools and guarantees of value and corporate ethos.

Cummings and Worley[19] provide some guidelines for achieving a cultural change and these are described here within the meaning of what has been demonstrated:

1. *Formulate a clear strategic vision.* To make cultural change effective, a clear vision of the new company strategy, shared values and behaviors is needed. This vision provides the intention and direction for cultural change. The aspects that emerged in the previous sections must give shape to a company strategy that can properly express the values and ethical assumptions underlying the company's operations and the functioning of its products. At this level, it is a matter of making the ethical DNA visible and operative, which will then guide corporate development and identity.

2. *Showing top management's commitment.* It is very important to keep in mind that cultural change must be managed by the organization's leaders, since the management's will to change is an important indicator. The organization's leaders should be very supportive of cultural change by becoming involved in the corporate mission to effectively implement the change in the rest of the organization. In this sense, the control and diffusion of ethical drivers in the form of corporate culture is one of top management's key tasks. This entails appropriate training, regular consultation and the existence of specific business locations for these topics (ethics training seminars, case studies on AI and ethics, and debates on ethically sensitive topics for AIs, together with specific working groups are just a few tools that can be thought of as role and in-role training for this key company level).

---

[19] Cf. T. G. Cummings and C.G. Worley, *Organization Development and Change*, South-Western College Pub, 2004, pp. 491-492.

3. *Modeling cultural change at the highest level.* In order to prove the management team is in favor of change, the change must be noticeable at first at this level. Management behavior must symbolize the types of values and behaviors that are required in the rest of the company. This process may also include the creation of committees, task forces for employees, value managers or similar. Change agents are key in this process and the main communicators of the new values. They should be brave, flexible, have excellent interpersonal skills, company knowledge and patience. It is a matter of finding corporate staff who act as catalysts (evangelists) and not dictators. In fact, change requires being mediated in training and must be verified. In this regard, one can conceive specific roles, such as an *AI* Ethics Manager whose task is to clarify the goals at all levels involved in the development of AI solutions and ensure the transparency of ethical drivers, knowing how to explain them using tools suitable for every company level and degree of expertise. AI ethics content must form part of internal personnel training in two forms: the dimension of values, theoretical and abstract training; and ethical skills creation, that is, the ability to know how to include and incorporate corporate ethical *values* and ethical *drivers*, expressed in the strategic vision outlined in point 1. These staff must make ethical guidelines the specific aim of their work, knowing how to communicate what has been done at the different company levels. Likewise, an *AI Ethics Manager* would possess the skills this technological innovation necessarily brings with it. This refers not only to technical knowledge, such as information technology or data sciences, but also to the need to develop and integrate humanities. A fundamental theme in this area is the ethical paradigms involved in development so the managers responsible for this must be able to say where and to what extent

operations followed the principles of consequentialism (Bentham/Mill), the ethics of duty (Kant), the ethics of virtues tempered by a concept of natural moral law (Aristotle-Thomas), the ethics of recognition of others (Ricoeur/Levinas) and so on. In developing AI systems capable of ethical learning, it will be the job of these corporate *assets* to define and communicate how the system will learn all this and how it will choose between the various paradigms. Moreover, it will be up to such figures to manage the situations in which an individual programmer can decide independently (e.g., in the event of an accident must a self-driving car make its decision based on a utilitarian calculation - save the greatest number of lives -, according to a paradigm of duty - to follow an a priori rule that protects the driver - etc.?), and when it is the company's task to define the ethical reference paradigm. It will also be the task of these resources to be able to respond to and resolve any conflicts that might arise between individuals and the company. For example, if such implementations cause ethical conflicts to arise, who must the employee obey? His/her conscience? The tradition of his/her faith community? The position accepted and proposed by a certain political idea? The corporate guidelines? Or the desire or commission of the system's client? And if the client were to choose to "customize" the ethical paradigm of the AI system or to ask third parties, including potentially a hacker, to modify it, what would be the ethical implications of such "customization"?

4. *Changing the organization to support organizational change*. This includes identifying which systems, policies, procedures and current rules need to be modified in order to align with the new values and desired culture. This may include a change in the systems for accountability, pay, benefits and reward structures and recruitment and retention programs to better align with

the new values and send a clear message to employees that the old system and culture belong to the past. One way to implement a culture is to link it to the sense of belonging to the organization. Encouraging employee motivation and loyalty to the company is essential and will also lead to a healthy culture. The company and change managers should be able to articulate the connections between the desired behavior, and the way it will affect and improve the company's success, to further encourage buy-in in the change process. Training should be provided to help all employees understand the new processes, expectations and systems. An effect of this type is conceivable only through dedicated roles in the company's organizational chart. *AI Ethical Managers* must be recognized company figures operating at different levels, people with knowledge of the company's assets and tools who contribute to defining the company's best operational practices. These figures are not assigned to oversee only the product and the process by which this is achieved, but also the quality of the ethical culture that develops and takes shape in the company's "becoming" process.

5. *Developing ethical and legal awareness.* Changes in culture can lead to tensions between organizational and individual interests, which can create ethical and legal problems for professionals. This is particularly relevant for changes to employee integrity, control, fair treatment and job security. It is also useful, as part of the change process, to include a periodic assessment process to monitor the progress of change and identify areas that need further development. This step will also identify barriers to change and recognize and reward employee improvement, which will encourage continued change and development. It may also be helpful and necessary to incorporate new change managers to refresh the process.

Cultural change in an organization is very important and inevitable. As such, the emerging ethical issues can only survive or be effective if they are incarnated in an organizational culture that makes them alive and operational. Cultural innovation is bound to be more difficult than cultural maintenance because it involves the introduction of something new and substantially different from what prevails in existing cultures. People often resist change, so it is management's duty to convince their collaborators that the likely gain will outweigh the losses. In addition to standardization, deification is another process that tends to occur in highly developed organizational cultures. The organization can be considered precious in itself, a source of pride and, in a certain sense, unique. The organization's members begin to feel a strong bond with it that transcends material returns and they begin to identify with it.

Conclusion

The core issue in the management and development of artificial intelligence is a broad area of ethical discernment that must take into account the potentially disruptive effect of these technologies because of their potential for *technological innovation*. By its nature, this innovation process has an intrinsic capacity for social *transformation*. In light of the ideas emerging in research and the resultant ethical-social issues, it has become evident the innovative process of artificial intelligence can only be positively evaluated if it is characterized as a justified, progress-oriented, human-friendly tool that takes the form of a true and sincere moral commitment by individuals and institutions in the search for the common good. Therefore, the management of the machine sapiens and its development in the near future requires political and economic management, making it necessary to establish international governance for the development of these technologies.

   In particular, since people are so intricately tied to this, the governance of artificial intelligences becomes the instrument

that can guarantee that the synthetic cognition made possible by *technological innovation* cannot take on dehumanizing forms. Governance is the space in which anthropological and ethical considerations must become effective forces and develop organizational cultures to shape and guide *technological innovation*, making it an authentic source of human development. This space for political and economic action, namely the governance of *technologies*, thus becomes a compulsory ethical call: *innovating and developing new technologies* must be translated into a commitment to the governance of *artificial intelligence technologies and widespread corporate responsibility*.

About the Authors

Paolo Benanti has been a Franciscan of the Third Order Regular since 1999. His studies have included engineering (La Sapienza), as well as philosophy and theology (Lateran, Gregorian), with special interest in ethics and moral theology; his doctorate was in the area of bioethics. Professor of Moral Theology, Bioethics and Neuroethics at the Gregorian. He is a member of the Task Force on Artificial Intelligence of the Agenzia per l'Italia Digitale and the Prime Minister's Office. The author of many academic and more popular articles, he has recently published the books *Homo Faber: The Techno-Human condition* (2018); *Postumano, troppo postumano. Neurotecnologie e human enhancement* (2017), and *Oracoli. Tra algoretica e algocrazia* (2018).

Thomas A. Campbell is Founder & CEO of FutureGrasp, which advises organizations on trends and implications of emerging technologies. From 2015 to 2017, Tom was the first National Intelligence Officer for Technology with the National Intelligence Council in the US Office of the Director of National Intelligence. He has informed senior policymakers, enabled millions of dollars of funding, broken ground in new research areas, and kept diverse groups abreast of the rapid pace and implications of technology change. Tom is a recipient of an Alexander von Humboldt Research Fellowship for post-doctoral research in Germany. He holds a Ph.D. in Aerospace Engineering Sciences from the University of Colorado at

Boulder, and a B.E. in Mechanical Engineering from Vanderbilt University.

Caitlin Chin is a research analyst at the Brookings Institution's Center for Technology Innovation. At Brookings, she contributes to research on privacy, artificial intelligence, antitrust, and broadband policy. She is a co-author of "Bridging the gaps: A path forward to federal privacy legislation", a comprehensive set of recommendations to reach middle-ground solutions on U.S. federal privacy legislation. Prior to joining Brookings, she completed the Google Public Policy Fellowship and Atlantic Media Fellowship programs and interned with Verizon's U.S. public policy team.

Samuele Dominioni, Ph.D., is a research fellow at the ISPI Centre on Cybersecurity. His main research interests revolve around cyberspace governance, nondemocratic regimes, and electoral processes. Before joining ISPI he was post-doctoral researcher at the Forum Internationale Wissenschaft (University of Bonn). He also worked as consultant and researcher for European Union funded projects, for the Venice Commission (Council of Europe), and the Office of Democracy and Human Right at the Organisation for Security and Cooperation in Europe (OSCE/ODIHR).

Mishaela Robison is a research intern at the Brookings Institution's Center for Technology Innovation. Her research interests focus on the overlap between psychology and social media manipulation, particularly fake news and extremism. At Brookings, she contributes to research on political polarization, artificial intelligence, social media, and the digital divide. Prior to joining Brookings, she worked with several research labs in the Psychology and Communication departments at Stanford University.

Counselor **Fabio Rugge** is Head of ISPI's Centre on Cybersecurity, in partnership with Leonardo. He is a diplomat currently working as Head of the Office in charge for NATO and Security and Politico-Military Issues, Directorate General for Political Affairs and Security, Ministry of Foreign Affairs and International Cooperation. From 2012 to 2016 he worked at the Italian Prime Minister's Office and prior to that he was Counselor at the Italian Delegation to the North Atlantic Council in Brussels and Consul General of Italy in Mumbai (India). He graduated cum laude at Catholic University of Milan, holds a Master at ISPI in International Relations and a European Studies Diploma from the "Université des Sciences Sociales de Toulouse" (France). Fabio Rugge is Adjunct Professor of Cyber Diplomacy at LUMSA University. He held courses and lectures in several Italian universities on Cybersecurity and International Relations.

John Villasenor is a nonresident senior fellow in Governance Studies and the Center for Technology Innovation at Brookings. He is also a professor of electrical engineering, law, and management at UCLA, as well as a member of the Council on Foreign Relations. Villasenor's work considers the technology, policy, and legal issues arising from key technology trends including the growth of artificial intelligence, the increasing complexity and interdependence of today's networks and systems, and continued advances in computing and communications. He has written for the *Atlantic*, *Billboard*, the *Chronicle of Higher Education*, *Fast Company*, *Forbes*, *Los Angeles Times*, *New York Times, Scientific American*, *Slate*, and *the Washington Post*, and for many academic journals. Prior to joining the faculty at UCLA, Villasenor was with the NASA Jet Propulsion Laboratory, where he developed methods of imaging the earth from space.

Darrell M. West is the Vice President of Governance Studies and a Senior Fellow in the Center for Technology Innovation at the Brookings Institution. He holds the Douglas Dillon Chair in Governance Studies. His current research focuses on American politics, technology policy, and artificial intelligence. He is the author of a number of books including *Turning Point: Policymaking in the Era of Artificial Intelligence* (2020; with John Allen), *Divided Politics, Divided Nation: Hyperconflict in the Trump Era* (2019), *The Future of Work: Robots, AI, and Automation* (2018); among others. He is the winner of the American Political Science Association's Don K. Price award for best book on technology and the American Political Science Association's Doris Graber award for best book on political communications. His books have been translated into Chinese, Japanese, and Korean, and he has delivered lectures in many countries around the world.