Charting AI IR: Much Ado About Nothing?

STEPHANE J. BAELE
UCLouvain, Belgium
AND
IQRAA BUKHARI
University of Exeter, UK

In a recent survey of 480 Artificial Intelligence (AI) researchers (Michael et al. 2022), 73% of respondents agreed that AI-led automation is likely to trigger "revolutionary societal change on at least the scale of the Industrial Revolution", and more than a third found "plausible that AI could produce catastrophic outcomes in this century, on the level of all-out nuclear war". A year earlier, the final report of the US Congress' National Security Commission on Artificial Intelligence (NSCAI) presented AI technologies as "the most powerful tool in generations" whose uses will "reorganize the world", simultaneously "expanding knowledge, increasing prosperity and enriching the human experience" and "deepening the threat" in a time of high strategic vulnerability (NSCAI 2021, 19-20). To use the language of a special issue published by Time magazine,[1] an "AI revolution" is about to shake global politics.

Such seem to be the stakes and uncertainties of AI, a diverse constellation of computational technologies eluding a precise definition beyond generic and incomplete descriptions such as "the theory and development of computer systems able to perform tasks normally requiring human intelligence" (Oxford Dictionary), the "study of how to produce machines that have some of the qualities that the human mind has" (Cambridge Dictionary), or "the field which combines computer science and robust datasets to enable problem-solving" (IBM).[2] The dramatic acceleration of AI prowess over the past few years has generated a lot of hype and bombastic headlines – doomsday or utopian – reflecting both the real hopes and concerns illustrated by the two examples mentioned in our opening paragraph, and the difficulty to clearly evaluate how this new and rapidly developing technological could shape world affairs. In international politics, AI can no longer be reduced to the realm of science fiction scenarios. The technology has already started to move lines[3] and its reliance on advanced microchips overwhelmingly made in Taiwan[4] is a further stressor of US-China rivalry. Major powers seek to canvass its security implications (e.g., European Parliament's 2020 recommendation for a European Security Commission on AI), turbocharge its development (e.g., China's 2017 New Generation AI Development Plan), or take actions to protect their assets (e.g., European Commission's 2023 plan to ringfence four core technologies). As Perez-Des Rosiers (2021, 113) argues, AI and technological development more generally "has become the dominant question in national and global politics influencing the issues of power". The NSCAI (2022, 8) advocated no less than "an integrated national strategy to reorganize the government, reorient the nation, and rally our closest allies and partners to defend and compete in the coming era of AI-accelerated competition and conflict".

In this context, the aim of this Forum is to clarify the stakes by providing a clear-eyed assessment of the present situation. Taking stock of progress already done in the emerging literature while simultaneously raising new questions and pointing to dimensions so far unaddressed, contributors – all recognized experts already pushing the boundaries of this field – offer not only a baseline understanding of how specific AI technologies effect particular areas and dynamics of international politics, but also critical evaluations of the extent to which AI truly constitutes the game-changer or "revolution" evoked

[1] See https://time.com/6283716/world-must-respond-to-the-ai-revolution/.
[2] https://www.ibm.com/uk-en/cloud/learn/what-is-artificial-intelligence. The very term "Artificial Intelligence" traces back to a 1956 conference, the *Dartmouth Summer Research Project on Artificial Intelligence*.
[3] Sometimes literally: AI models sift through intercepted Russian army communications to extract tactically relevant information for Ukrainian forces and US intelligence (Knight 2022).
[4] Data at https://chipexplorer.eto.tech/.

above. Together, these distinct interventions clear the ground on a multifaceted issue, point the non-expert reader to important existing contributions, and propose some structure to what is likely to be an important research agenda whose implications for the discipline may not be dissimilar to those of nuclear technology in the 1940-50s.

The rise of AI: New breakthroughs, new challenges

Like all major technological developments, AI unfolds new issues and problems as it solves older ones, thereby triggering multiple forms of social reaction. Whilst the application of signature AI methods (chiefly deep learning)[5] into a range of very different domains has already produced dramatic shifts in how these fields operate, it has also simultaneously brought about new – and sometimes alarming – challenges. In medical research for example, AI-powered health and nutrition programmes propel athletes to unprecedented levels of performance,[6] and DeepMind's AlphaFold model (which predicts the structure of proteins) has unlocked a series of longstanding scientific quests such as an effective malaria vaccine (Ko et al. 2022);[7] yet a company using comparable AI approaches for drug discovery has also evidenced the "dual use" of their model, finding thousands of previously unknown neurotoxic agents that could be used as chemical weapons (Urbina et al. 2022). In strategic games, AI models now systematically outclass the very best human players (chess in 1997, go in 2016),[8] giving fascinating glimpses into super-human problem-solving and offering new training techniques for players reaching unprecedented levels; yet this research has opened avenues for cheating and far-reaching implications for realms such as military strategy and political bargaining. DeepMind, which recently developed an almost unbeatable AI model for Diplomacy, a game where multiple players simultaneously undertake complex interactions involving a mixture of cooperation and competition, incidentally make clear that their efforts "pave the way towards creating artificial agents that can successfully cooperate with others, including handling difficult questions that arise around establishing and maintaining trust and alliances" (Anthony 2020, 9).[9] Similar advances, with their resulting new challenges, can be found in multiple domains from stock markets to agriculture, which has rightly motivated an intensive discussion on the ethics of AI (e.g., Liao 2020; Dubber, Pasquale and Das 2020).

AI in IR: (r)Evolution or insignificant fad?

Accordingly, it is incumbent on scholars to develop insight into how AI and global politics react, potentially sparking systemic and institutional change into other domains. The real question is to know exactly how, and how much, the technology will alter international politics as we know it, whether AI's "reorganisation of the world" could prove to be a "revolution", to use the NSCAI and the expert surveys' words cited above. This question calls for a careful and multifaceted evaluation; technology impacts most dimensions of international interactions (Hoijtink and Leese 2019), meaning that AI's effects are to be felt in a range of domains. The following paragraphs briefly review the three main strands of the emerging literature on AI in IR to both identify the three different domains where AI is most likely to triggers change – war, international laws and norms, and strategy – and introduce the various contributions to this forum, which fall within these three areas.

Unsurprisingly, most of the effort appraising the effect of AI in IR has gravitated around what the technology means for *warfare*.[10] As Garcia (2019, 2) indeed observed, "the militarization of AI[11] has already begun". Boulanin and Vergruggen's (2017) extensive examination of the various ways in which

[5] *Deep learning* refers to the multi-steps processing of vast quantities of unprocessed data (e.g., images, text, behavioural patterns, chess games) to detect and hierarchize features in the data without human intervention.
[6] See for example Kaloc (2022).
[7] Such endeavours already started in the 1970s and 1980s, with early AI systems such as DENDRAL or Mycin allowing for breakthroughs in organic chemistry and disease diagnosis.
[8] Machine learning actually originates in a checkers game model (Samuel 1959).
[9] Note how this directly evokes Axelrod's (1984) cornerstone contribution to the structure of cooperation in IR.
[10] The very first attempts to consider how AI could alter IR, dating back to the early 1990s (e.g. Hudson 1991) were, however, much broader.
[11] This process could be an instance of *securitization*, whereby the presentation by leaders of a social issue as a critical security threat creates the conditions for executively dealing with this issue outside democratic checks (see Buzan, Waever and De Wilde 1998).

AI is already used to enhance the autonomy (in intelligence gathering, mobility, targeting, and interoperability) of various types of weapon systems (drones and aircrafts, robotic sentry weapons, short- and long-range defence and active protection systems) evidences this fact – and their study predates the most ground-braking technological advances. While grand announcements of a new revolution in military affairs are made – not least by states' security establishments –[12] scholars have warned against hyperbolic claims (e.g., Payne 2018, 11) and preferred to break down warfare into its tactical, operational, and strategic levels to better locate AI's impacts (e.g., Jensen, Whyte and Cuomo 2020, 2022). If these still remain uncertain at the strategic level (Payne 2021 and below in this Forum), they start to appear more starkly at the tactical and operational levels where "human-machine collaborations"[13] are set up to "increase military power" in ways that delivers important battlefield advantages yet also triggers "new forms of risk" (Jensen, Whyte and Cuomo 2020; also Horowitz 2019). The actors likely to benefit from the technology, however, are largely those who already dominate the international system and the global economy,[14] which means that the militarization of AI is unlikely to fundamentally upset the existing hierarchy. But more specific areas of modern conflict carry higher potential for change. Beyond the battlefield, AI for instance brings new challenges to the ever-fragile nuclear equilibrium; new AI-enhanced command and control systems may well offer faster detection and more optimal response patterns, they also open up new risks of rapid unwarranted escalation fuelled by automatic chains of decisions (Fitzpatrick 2019; also Johnson 2020) that could be initiated by weaker, non-nuclear states or even non-state actors. In the cyber domain, AI tools are more likely to significantly aggravate the situation more than completely reconfigure it; a "force multiplier for both defensive and offensive cyber weapons", the "new AI-cyber nexus" is bound to increase the severity of the threat and its consequences (Johnson 2019). Further exploring these and related topics, Jensen, Whyte and Cuomo's contribution to this forum focuses on kinetic warfare by stressing how uncertain AI's impact still is, while Baele's intervention maps the effects of AI on information warfare.

This percolation of AI into warfare has rightly raised alarm in diplomatic circles, with a range of initiatives launched to build if not an *international legal framework*, at least a normative agreement on which technological applications can/cannot be tolerated, with a particular attention to Lethal Autonomous Weapon Sytems (LAWS).[15] As already highlighted by Denise Garcia in 2015, global rules on military AI must be created to avoid a new arms race and proliferation, and initiatives have therefore started to emerge. Above national regulations such as the US Department of Defense' 2017 Directive on Autonomy in Weapons Systems, and transnational initiatives such as the European Parliament's 2018 resolution on automated weapons systems, a classic international "life cycle of norms" (Finnemore and Sikkink 1998) has started with NGOs like Human Rights Watch pushing campaigns such as "Stop Killer Robots" and national diplomats from "progressive" states attempting to cement the new norm at forums like the UN Convention on Certain Conventional Weapons or the IAEA's Working Group on the Ethics of Nuclear and AI. This process, however, remains insufficient and almost fruitless; not only does it suffer from AI's multifaceted character, definitional issues, and enforcement constraints (Bills 2014; Hammond 2015; Bode and Huelss 2018; 2022), but also from norm antipreneurs' efforts to dilute work: strategies and counter-strategies are deployed by "adversaries [who] seek to establish new norms of accepted behavior in ill-defined spaces" (Jensen, Whyte & Cuomo 2020, 536). In this case, AI therefore appears to further stabilize existing patterns of international interactions rather than re-shape them: great powers with a technological edge orient the (lack of) outcomes and, as Garcia (2019, 2) notes, the Global South "is clearly underrepresented in this debate, with many areas of Africa, Asia, Latin America and the Caribbean completely absent". In his concluding remarks to this forum, the same Garcia – a diplomat playing an active role in these negotiations – critically reflects on the situation and points to the unequal effects of AI across the North-South gap.

[12] The NSCAI (2022, 22) for example argues that "AI will transform all aspects of military affairs".
[13] Such as swarms of interconnected manned and unmanned aircrafts.
[14] This point is recurrently emphasized in the new IPE literature focusing on AI, which is unfortunately beyond the scope of this Forum. Key contributions are included in Keskin and Kiggins' (2021) excellent edited volume ; see also chapters in Agrawal, Gans and Goldfarb (2019) or Peng, Lin and Streinz (2021).
[15] These efforts are grounded on reflections on the ethics of LAWS (e.g., Arkin 2010; Sharkey 2010; Asaro 2020; Umbrello, Torres and De Bellis 2020)

Finally, the militarization of AI and its incomplete and biased inclusion into the global governance system forces us to consider *higher-level IR considerations*: how does such an environment modify the structure of the international system, its balance of power, and the strategies attempting to navigate it? Some scholars have already identified several ways through which, they argue, AI will have "a profound impact on the conduct of strategy and will be disruptive of existing balance of power" (Ayoub and Payne 2016). First, states with the capacity to embed AI into high-level strategic decision-making procedures might be able to avoid some of the heuristic biases commonly associated with foreign policy decision-making (in addition to capitalizing on the cumulative gains likely made at the tactical level). Payne, in his contribution to this forum, offers a reflexion on the difference between human and artificial intelligence that appraises how recent AI tools could enhance strategic decision-making. Second, such a situation is already increasing fear among great powers of the implications of a potentially dramatic "first mover effect". Allen and Chan (2017, 32) for instance unpack the consequences of a situation whereby a state has acquired such an advantage in AI-based analysis that it "achieves decisive advantage in decision-making" and thereby engages in more assertive and aggressive foreign policy. This fear contributes to mounting tensions between major powers, with increasing tightening of exchanges involving technologies and expertise involved in AI. For Horowitz (2018; 2019) and others, however, the ongoing rapid spread of AI within non-military sectors is taming first-mover effects and fear thereof: AI becomes so widespread (as opposed to, say, nuclear technology) that its impact on power politics is unlikely to come from a potential first-mover effect in the realm of strategy. Third, and because of concerns about being out-paced, states have stepped in an AI-arms race, funding programmes destined to further militarize AI. Boulanin (2019) suggests in particular that "machine-learning and autonomy might become the focus of an arms race among nuclear-armed states", impacting "their calculation of strategic stability and nuclear risk at the regional and trans-regional level". In this process, the greatest danger is, according to him, to "underestimate or disregard the limitations of current AI technology. […] an immature adoption of the latest developments in AI in the context of nuclear weapons systems could have dramatic consequences". Fourth and finally, the rise of AI within foreign policy and military strategy may accelerate the mutation of the structure of the international system from the US-led Liberal International Order (LIO) into a new, bi-polar system. This is because of both discursive performative effects – in their respective AI strategy documents, the US and China explicitly present themselves as competitors – and the practical steps to get a decisive edge in what is perceived to be a critical new technology they cannot afford to ignore.[16] As Johnson (2019; 156) suggests, AI might be to the early days of the "new Cold War" what nuclear technology was to onset of the US-USSR opposition, with "parallel trends in the shifting geopolitical landscape and disruptive technologies [that] fundamentally reshape the security environment, which in turn, will have significant implications for how a future U.S.-China crisis or conflict might unfold". Again, since most of the AI "revolution" is – increasingly – concentrated in a few powerful states (Payne 2018), this development is bound to accentuate bandwagoning pressures among smaller states, further fuelling a bipolar dynamic.

To summarize, the rapid development of AI has already set in motion significant international political dynamics, starting with how warfare is conducted to where the structure of the international system evolves. The technology appears to further accentuate existing gaps, tensions, rivalries, and tendencies, rather than shaking down the existing system; yet as further unpacked in the following contributions, AI generates multiple and diverse effects on several dimensions of global politics, making it difficult to offer, at present, a single and definitive answer to its "revolutionary" nature.

Uncertain AI: Framing the Promise and Problems of AI in the Battlespace of Tomorrow

[16] See for instance He 2022 on the recent US decision to subject microchips to export controls.

CHRISTOPHER WHYTE
Virginia Commonwealth University, USA
SCOTT CUOMO
U.S. Marine Corps, USA
AND
BENJAMIN JENSEN
Marine Corps University, USA

Much as the 2010s saw a groundswell of focus on cyber conflict in the pages of International Relations' (IR) foremost publications, the first three years of the 2020s have so far seen burgeoning interest in better assessing the impact of advances in artificial intelligence (AI) on global society and international conflict. This ongoing upsurge in interest has taken several different forms. Popular publications by AI experts and national security specialists have galvanized policy debate about the opportunities and possible pitfalls of investment in new methods. IR scholars have used these debates – about AI safety, cognition, automation and more – as a springboard from which to theorize and attempt to talk realistically about where AI might be most significantly poised to alter the character of warfighting (e.g., Horowitz 2018; Jensen, Whyte and Cuomo 2018; Payne 2018; Johnson 2019; more recently Petraeus and Roberts 2023). Taken together, these developments are positive early signs of a willingness to engage the idea that AI is likely to be truly transformative, as well as a recognition of the need to avoid the issues of marginalization that plagued early IR scholarship on cyber conflict.

And yet, new focus on an emerging technology does not necessarily mean that IR is succeeding in bringing the implications thereof *into* focus. At present, there remains immense uncertainty about precisely what is going to change about the conduct of warfare and national security (Jensen, Whyte and Cuomo 2022). Broadly, the norm in IR conversations about artificial intelligence is to center on the revolutionary nature of the technology and the implications thereof for the transformation – seen to be inevitable by most – of the character of warfighting (Horowitz 2019). This is true for both broad treatments, both optimistic or pessimistic in tone, and more narrow explorations of AI applications such as deepfakes (Whyte 2020), Lethal Autonomous Weapons Systems (LAWS) (Horowitz 2019) and machine learning-enabled offensive malware (Davis 2019; Gibert et al. 2020). This trend concerns us, not least because it mirrors much early work on cybersecurity and cyber conflict in IR that centered on techno-strategic characteristics of the phenomenon – i.e., a "logic of the domain" approach – in the absence of empirical evidence (Gomez and Whyte 2021; Gomez 2022). With AI, we particularly see pervasive simplification about specific developments and their impact on national security. Though some effects to be felt are quite clear, most are anchored to sufficiently complex conditions as to remain analytically ambiguous at this time. The result is a tendency for much IR work to spiral in towards parsimonious "bigger, faster, smarter, better" narratives about AI (Whyte 2022), a dynamic that risks promoting unhelpful memes about the technology (e.g., "killer robots" [Horowitz 2016; Young and Carpenter 2018] akin to the "cyberwar" and "cyber Pearl Harbor" of a decade past [Rid 2012; Lawson 2013]).

In the remainder of this Forum contribution, we elaborate on what we see as the key reasons that uncertainty pervades IR discourse about AI and its prospective impact on military power. We particularly argue that the elements of information processes most likely to shape and be shaped by AI to create geopolitical transformation are sociological and institutional, not technological. As such, it is on these dimensions of technology innovation and adoption that scholars should focus to most effectively forecast where "solutions" of all stripes proposed around the use of AI might either become problematic or otherwise sprout unforeseen implications for international relations. The point we make underlines a core proposition shared across this Forum: that AI *may* have transformative impacts on security and insecurity in world affairs, but much as has been true for past information revolutions, the extent and specific character of any change will be modulated by preexisting and interacting sociopolitical factors that guide continued development and adoption. Given this, we conclude by describing opportunities that we see for scholars to rely on existing literatures that describe uncertainty in informational terms, thus allowing for considering of technological impact independent of machinery thereof.

Thinking about AI in future conflict

Artificial Intelligence is not one technology, but rather a broad array of developments being nurtured across a wide spectrum of scientific and social scientific disciplines. The term and the ideas that underlie its now-common use to describe such diverse techno-scientific developments stretch back to the midpoint of the 20th century, when a series of researchers at American institutions of higher education set out to ask a simple question: could a machine be made to realistically simulate human intelligence (Nilsson 2009)? As a result of early efforts, the AI field of today encompasses developments in machine vision and auditory sensing, natural language processing, robotics, machine learning and deep learning.[17] Of these, the last two – machine learning and deep learning,[18] of which the latter is a subset of the former – are arguably the most significant areas of current development and certainly the sub-disciplines most congruent with the AI field as a whole. Both face developmental challenges. Shortcomings aside, however, computational AI along these lines stands to revolutionize logistics and production infrastructure for state militaries simply by dint of the ability to shore up inefficiencies via sheer weight of computational analysis. At the same time, deep learning offers new tools for communication, battlespace surveillance and force coordination that – as the technological disparities of the 2022 war between Russia and Ukraine, in which a Western-backed Ukraine has utilized advanced technology to overcome firepower and manpower imbalances in battle against a Russian army light on non-lethal technological sophistication, illustrate – might prove a gamechanger in the prosecution of future conflict.

Of course, the potential pitfalls of getting AI wrong are also clear to many, at least, again, in broad terms. So much of the emerging scholarship has centered on the idea of a global AI arms race to the bottom, in which countries try to out-develop one another and end up producing flawed machine warfighting aids, that it is quite arguably its own distinct branch of investigation within the burgeoning AI subfield in IR (e.g., Roff 2019; Scharre 2019; Taddeo and Floridi 2018). Interestingly, discourse about an arms race and the doomsayer warnings that accompany it remain as common today as they were in the mid-2010s. There does exist IR work that debates the efficacy and role of particular AI applications in the context of future conflict, both near term and distant. But even the best scholarship to date contains understandable uncertainties about the future of AI development and AI in conflict.

We discuss now what we see as the core – and quite understandable – reasons underlying this uncertainty. Rather than fall back on such broad narratives about AI promise and problems in the face of such uncertainties, however, we urge experts in IR to use these clear sources of ambiguity about forthcoming technological impact as a template for research rather than a call to generalize or warn about possible effects. To do so would be to avoid a race to the "analytic bottom" at a time when the core assumptions to be made about AI's influence on world affairs are as-yet under-developed.

Uncertain Techno-Scientific Pathways

We observe a range of technology-centric reasons as to why there is such persistent uncertainty around pinning down the specifics of future AI systems, their impact on the battlefield and the broader security context within which such transformation will occur. We see these technological reasons present most clearly in work on Lethal Autonomous Weapons Systems. LAWS are simple enough to understand. They are robots that have weapons of some kind (Horowitz 2019). And yet, there is enough complexity bound up in the development of LAWS to illustrate a broader point about positioning AI as a conflict modifier (Gill 2019), namely that the underpinnings of such a dynamic technological innovation – in artificial intelligence itself – present nigh countless development pathways along technical, logical and ethical lines (Brooks 2015; Umbrello, Torres, and De Bellis 2020). In particular, the "autonomous" part of LAWS is broadly interpreted by both those attempting to produce such systems and those studying them: is it target selection that "autonomous" is describing? Or perhaps is it the selection of a method of engagement, or the actual decision to engage? Obviously, from a definitional point of view, any of these might delineate the line between AI systems and conventional ones. But the presence of one form of automation does not preclude humans elsewhere in the operational loop. Automation thus quite quickly becomes something that exists on a spectrum rather than as a dichotomy. This, in microcosm,

[17] It is worth noting, though, that many simpler and more detailed divisions of focus are common among AI researchers.
[18] See note 3.

reflects a core challenge that IR researchers face in thinking about AI. Many analyses speak about the characteristics of narrow AI and generate assumptions about the role of prediction, judgement, validation and more in future conflict. But how substantially should such assumptions be applied given a variable degree of capacity for self-direction or self-determination with a given AI-enabled system? Clearly, there are obvious and understandable challenges involved in determining how the assumptions of a developing field – assumptions based on logical analysis rather than empirics – should apply to evolving conditions that are both complex and uneven in their progression.

AI systems are also dynamic in their form and function by dint of the way in which they learn about, engage with, and reorient on their environment. The character of so much forthcoming AI application as fundamentally pliable to unclear inputs and open to manipulation – and, thus, subversion – adds additional uncertainty to IR scholar's attempts to delineate clear parameters for understanding AI in future conflict (Whyte 2020). On the one hand, the "black box" problem– wherein users of AI systems can see inputs and outputs clearly but not necessarily the inner workings of the algorithm as it translates one to the other – presents a challenge to imagining AI that is not sufficiently human in its cognition. On the other hand, even the most robust systems might be hacked at the design, fine-tuning, or auditing stages and are otherwise susceptible to input attacks or efforts to poison AI models. These forms of manipulation, wherein training data is altered or real-world response cues are faked to produce a deviant outcome, make gauging the choices that developers will need to make in building for more reliable AI hard to predict and preventing narrow logical analysis of future AI systems with any amount of realism. With LAWS, this uncertainty is particularly pronounced as there are, on the one hand, physical elements of such systems that would be prime candidates for modification to prevent unwanted behavior and, on the other hand, lethal capabilities bringing military operators more directly into contact with potential international legal or reputational issues (Maschmeyer 2022).

Humanity, Humans and AI Uncertainty

We also note that the development of new technologies for battlefield use are driven by human factors that also portend extremely diverse development pathways. Two dynamics bear particular mention. The first is simply that technological security issues are substantially defined by the way in which that technology impacts society (Horowitz 2020). Moreover, the more dynamic that transformation, the greater the distance that exists between initial expectations and eventual manifestation of security challenges. We look to cyber conflict as the most accessible example of this challenge for contemporary scholarship of future issues. A firm critique of Western approaches to cybersecurity in recent years has been the manner in which few experts or institutional processes recognized the form of the threat posed by cyber-enabled information warfare (IW) – arguably the definitive turn in digital security affairs in the 2010s – until after the 2014-16 Russian campaign against the U.S. political process (Whyte 2020; Nakayama 2022). Cyber-enabled IW is a threat form particularly enabled by broader changes to the global information environment brought about by evolutions in web technology since the 1990s, namely the corporate co-option of network services and subsequent socialization thereof for commercial purposes. And yet, early conceptualization of cyber threats anchored assumptions about utility and engagement around basic concepts of computer network attack and legacy perspectives that held cybersecurity only as an extension of conventional strategic competition (Whyte 2022). These views served a clear purpose, not least of which was as an organizing discourse for the development of national cyber institutions like U.S. Cyber Command (Branch 2021). However, they failed to clearly understand otherwise predictable evolutions of digital threats premised on societal change and not captured by the initial lens selected.

More generally, we observe that technology development is invariably determined by legacy structures, specific sociopolitical formations, public opinion and a whole host of other human factors that imprint idiosyncrasies onto the process of building and adopting new techniques (Dunn Cavelty and Wenger 2020). Core design choices for AI systems will inevitably reflect moral, cultural and other assumptions found in the institutions, social networks and personal contexts where such systems are built and deployed. The history of military innovation around cyber conflict makes clear this point, as the commitment to the cyberspace domain concept and the division of legal-operational duties across defense establishments in the United States emerged from entrenched perspectives on the role of information technology in warfighting held by key Air Force and intelligence officials from the 1990s

through the 2000s (Lindsay 2021). For IR scholarship, the core uncertainty to be found here is clear, namely that it remains unclear *whose* vision for AI's role in national security will win out across diverse national and sub-national settings. As such, the full range of conditions those prevailing perspectives will motivate contain not just technical ambiguities for those interested in theorizing AI in future conflict, but also economic, social and political-strategic ones.

Embracing the informational turn: Progressing the research agenda on AI and international security

The great amount of uncertainty surrounding the impact of AI on international affairs and dynamics of international security demands a coherence of scholarly efforts not yet apparent in so much work on AI by IR scholars. As IR scholars focused on disruptive information technologies, we see the challenges tied to creating such a coherence as quite similar for AI as they have been with other emergent technologies that have dramatically impacted military power – drones, computers, the Internet, etc. Simply put, there is too much focus on the character of the technology and less on the social conflicts that it both engenders and impacts.

With regards to the prior section, for instance, we would point out that the degree to which even those sources of uncertainty around AI that seem principally techno-scientific in nature matter is determined by sociopolitical context. Consider the case of cyberspace. Early strategic studies and military scholarship on operation via the Internet emphasized the challenge of attribution as a key motivator of offensive behavior without an obvious solution. Modern scholarship on cyber conflict tells us that technical attribution is often far more possible than maxims about anonymity online from the past might have us think, even if it can be resource- and time-intensive. Attribution is clearly relevant to calculations about the deployment of cyber operations for both tactical and strategic gain, but only as a probabilistic gambit tied to assessments of risk, political security and more. Important technological detail is thus layered atop broad sociopolitical context, rather than itself driving a reconfiguration of factors underling conflict and power contests.

AI *might* transform world affairs and the security architecture thereof. We argue that the elements of global information processes most likely to shape and be shaped by AI to create such geopolitical transformation are sociological and institutional, not technological. IR scholars and researchers in related areas should focus on these corollaries of technology innovation and adoption and, in doing so, minimize focus on examinations that risk technological determinism. Doing so is the key to most effectively forecasting where efforts to use AI to solve societal problems – particularly security problems – might either become problematic or otherwise sprout unforeseen implications for international relations. And indeed, there is particular urgency in doing so in the case of AI specifically. Whilst past information revolutions have shown us that information technology potential is shaped and given shape by societal context, it must be consistently recognized that AI stands apart in the degree to which its architecture is both accessible to and manipulable by the general public. Given too little oversight, the fault lines and faulty assumptions of today could become the inherent flaws of AI.

To what tools and perspectives should IR scholars therefore look for analytic purchase? A gradual informational turn has been taken among some key technology-oriented IR research programs in recent years, defined by a willingness to describe core sociopolitical phenomena and empirical developments in terms of their informational value. This is not dissimilar to trans-subfield efforts to parameterize uncertainty in IR more broadly (see Rathbun 2007) insofar as information-centric analyses describe the field's paradigms and theories simply as different dimensions of information – information on the move, information at rest, information with variable entropy, etc. The analytic challenges for the study of AI today stand to benefit enormously if they are couched in these metatheoretical terms, not least because informational perspectives eschew the dualism of focus on technologies' functional characteristics that often breeds determinism. In this way, the developing field might shortcut towards hypotheses and testable assumptions found both in more mainstream areas of IR and in those topical research programs wherein AI application can be seen more contemporaneously than the alternatives, from cybersecurity to LAWS to disinformation studies.

AI-fuelled Information Warfare: Deepfake Power

STEPHANE J. BAELE
UCLouvain, Belgium

The past years have witnessed a dramatic acceleration in the development of evermore powerful AI models designed to generate increasingly credible "synthetic" content of all kinds (text, audio, visual, audio-visual). The present contribution builds on a brief overview of this important technological evolution to evaluate its main impacts on information warfare and beyond.

From language models to deepfakes: The challenges of AI-generated "synthetic" content

A subclass of the broader category of *synthetic data* (data artificially generated by computer programmes, as opposed to data collected from the real world), *synthetic content* refers to any type of text, audio (sounds, voice, music), visual (pictures, drawings), and audio-visual (e.g., short video clips) content generated by an AI deep-learning model (usually called "generative AI"), that is, "a statistical technique for classifying patterns, based on sample data, using neural networks with multiple layers" (Marcus 2018: 3).[19] Generative AI models typically rely on vast amounts of data (pictures, for example) whose recurring patterns are identified to "teach" them to generate probabilities underpinning the production of new, "deepfake"[20] content mimicking the training data.

Models used to generate *text*, which are commonly referred to as *language models*, calculate probability distributions over sequences of words; whilst they are mostly used to help writing sms/email, suggesting words likely to follow those already written, they can also be harnessed to generate longer segments of text. Although these models have now existed for a while, they have vastly improved over the past four years thanks to a series of initiatives backed by massive financial resources (allowing for much larger and diverse training datasets and more sophisticated architectures). DeepMind's GOPHER (from 2022) was for example trained on the "MassiveText" corpus containing no less than 10.5 TeraBites of text (including, among others, the entire content of Wikipedia) and rests on a 280 billion parameters deep-learning network; OpenAI's GPT-3 – initially powering the popular "ChatGPT" – was trained on approximately 500 billion words and costed over $15m. to develop. The step-change represented by these "foundation" models is so significant that it has been hailed as a "paradigm shift" (e.g., Bommasani et al. 2021). Such a fast development towards a truly powerful technology has inevitably come with teething problems – such as when Microsoft's Tay model started to produce racist slurs on its associated Twitter account – that have only reinforced emerging concerns about "misuse" by nefarious actors.

The same logic has underpinned the development of highly sophisticated models using very large visual datasets to generate *images*, which have built on past and ongoing efforts to use AI to automatically classify and cluster pictures (e.g., Krizhevsky, Sutskever and Hinton's [2012] "ImageNet"). Models like Midjourney or StableDiffusion can now generate endless variations of an image from a simple description in a text box.[21] While these ground-breaking tools still suffer from shortcomings, their evolution has been so rapid than they are on the cusp of reaching the near-perfection of their linguistic counterparts. Narrower models have already been designed to create highly specific types of visual content. For instance, Zhao and colleagues (2021) have generated credible fake satellite imagery; arguing that the emergence such fakes "is inevitable just as 'lies' are essential in maps", they warned that "deep fake can potentially develop into a new mode of unpredictable and even terrifying fake geography". Tucker (2019) earlier reflected on similar projects, worrying about image models as "the newest AI-enabled weapon".

[19] Deep-learning models are versatile tools used well beyond the task of generating content, powering many of the varied AI tasks discussed elsewhere in this forum.
[20] Portmanteau for "deep-learning" and "fake".
[21] The author of this contribution was for example capable of generating a charming "painting of a family gathering, in a 18th Century naturalistic romantic style".

Although an increasing number of user-friendly apps now provide variants of deepfake *video* creation, flexible models generating fully customizable audio-visual content are still in development, and as a result the creation of credible deepfakes following an original script generally requires combining several AI models (e.g., one for "cloning" the target's voice, one for modifying/changing his/her face or lips in the video) requiring advanced programming skills. However, the pace at which the technology evolved, together with private companies offering custom audio-visual deepfake services and the proliferation of open-source models, leaves no doubt about its imminent evolution towards powerful versatile software available to the wider public.

Beyond the hype and grand claims naturally accompanying these developments, sophisticated reflections on the dark sides of deepfaked synthetic content have started to appear, tacking issues like the ethics of generating and disseminating deepfakes (e.g., Öhman 2020) or questions pertaining to intellectual property or consent (e.g., Harris 2019). The political, and more specifically security, dimension of the technology has caught the attention of a smaller, but arguably more disquieting, new literature unpacking the "disturbing array of malicious uses" (Chesney and Citron 2019a, 1757) associated with ever-powerful language, image, and audio-visual models. Major AI academic labs have tried to list and discuss these misuses (e.g., Bommasani et al. 2021), as have – to their credit – major AI companies.[22] As Chesney and Citron (2109a, 1762-1763) summarize, "the capacity to generate persuasive deepfakes will not stay in the hands of either technologically sophisticated or responsible actors. [… the technology…] is sure to diffuse rapidly no matter what efforts are made to safeguard it". The recent arrival of a Chinese foundation model, WuDao 2.0, also unfolds the possibility, evoked in the opening contribution of this forum, of AI fuelling great powers rivalry. Content-generating models have therefore also made their ways into major powers' statecraft, reflecting a growing sense that this technological development holds critical strategic significance. The US Congress' National Security Commission on Artificial Intelligence's voluminous report for example cautioned about the "malign uses" of AI applications "ranging from deepfakes to lethal drones" by "strategic competitors" and "rogue states, terrorists, and criminals" (NSCAI 2021, 9).

AI-fuelled information warfare: Deepfakes as a weapon of the weak?

Why should IR scholars care about these developments? Simply because, to use the words already penned down by E.H. Carr (1946 [1939], 1936) almost a century ago, "power over opinion is not less essential for political purposes than military and economic power and has always been closely associated with them". Propaganda and information operations constitute key aspects of contemporary international politics and conflict, where social media and digital influence play an increasing role (Patrikarakos 2017; La Cour 2020), especially in the context of the mounting tensions and multiplying influence strategies accompanying current challenges to the "Liberal International Order". While generative AI could well bring about "enormous security threats […] limited only by human imagination" (Verdolivia 2020), it is reasonable to focus our concern on one critical risk: the prime role destined to be played by the technology in information warfare, and more specifically what scholars like Woolley (e.g., 2018) call "computational propaganda". Existing scholarship indeed makes clear that deepfakes are "better suited for disinformation […] than information" (Buchanan et al. 2021).

To be sure, the production of political prose and the doctoring of images always been the fuel of propaganda or disinformation campaigns; editing software initiated the shift towards widespread modified imagery back in the 1990s, already creating a visual environment where most of the images in circulations are altered. The mass dissemination of digital content on online spaces for waging information warfare is not new either, as evidenced by ISIS' extensive deployment of online propaganda as part of a sophisticated information warfare campaign (e.g., Ingram 2019) or the "sweeping and systematic"[23] way the Russian government interfered in the 2016 US presidential election to achieve its broader strategic interests.

[22] See for example OpenAI's evaluation of the "social impacts" of their GPT-2 and -3 language models (e.g., Solaiman et al. 2019).

[23] As characterized by the Mueller *Report On The Investigation Into Russian Interference In The 2016 Presidential Election*, US Department of Justice, https://www.justice.gov/archives/sco/file/1373816/download.

Yet most scholars trying to gauge the impact of generative AI on information warfare insist that, as Verdoliva (2020) puts it, "changes the rules of the game". Reasons are advanced that pertain to, *simultaneously*, both quality and quantity. Even if each type of deepfake content currently lies at a different stage of sophistication, the evolution sketched above indicates that all will eventually reach perfection (understood as the inability by humans to distinguish fake from authentic content). Buchanan et al. (2021), for example, already demonstrated the ability of a "human-machine team" to generate GPT-3 "outputs that perform better than either the machine or the human could manage on their own". As a result, content-generating models will allow for both at once significant economies of scale on the producer's end – fewer people will be able to produce vaster amounts of the required type of content – and a potentially deepfake-saturated information environment on the target audience's end – large amounts of highly credible content channelled to specific online ecosystems. As Kreps, McCain and Brundage (2020) summarize, "the danger of new AI-based tools is scale and velocity: the ability to produce large volumes of credible-sounding misinformation quickly, then to leverage networks to distribute it expeditiously online"; Buchanan et al. (2021) agree, highlighting "striking" "powers of scale". Sustained efforts made to detect fake content are unlikely to make a significant difference for two reasons: they are inherently reactive, and they cannot efficiently deal with genuine content that is only marginally (yet decisively) altered by fake interventions (for example, a real video of Xi Jinping where only a few – important – words are altered). On that basis, scholars have pointed to *four main uses of synthetic content in information warfare*, each with different likelihood and impact severity.

First, scholars examining cyber efforts "by state and non-state actors to subvert democratic elections, encourage the proliferation of violence and challenge the sovereignty and values of democratic states" (Paterson and Hanley 2020: 439) have suggested that synthetic content will inevitably be *used to destabilize rival societies*, with deepfakes designed to damage the target society's social fabric by deepening its polarizing lines. In their seminal paper, Chesney and Citron (2019a) argued that deepfakes will thrive in polarized information environments' echo-chambers, "stoking social and ideological divisions". Long lists of scenarios, such as fake videos of government/police/military officials taking bribes, acting in violent or racist ways, or using inflammatory language, abound in the literature to illustrate the many avenues towards social destabilization. While these scenarios are still largely educated guesses, experimental evidence now exists that give some backing to these fears. Dobber and colleagues (2021), for example, showed that even a short video deepfake featuring a politician making an offensive joke negatively affect viewers attitudes towards that person – participants in the experiments took the authenticity of the video for granted. Kreps, McCain and Brundage (2020) demonstrated that synthetic politically tainted news reports were deemed credible and well received by ideologically aligned readers, evidencing AI content's potential for polarization. It is safe to claim, from that perspective, that synthetic content is likely to become a prime tool for a range of actors searching for maximal efficiency in their adversarial information operations. Authoritarian states (or non-state actors backed by these regimes) are likely to make use of the tool against the "West" (Allen and Chan 2017, 32), possessing here an asymmetrical advantage against democracies (Paterson and Hanley 2020, 442). It is indeed not hard to imagine the gains in efficiency, reach, and impact that Russia's "Internet Research Agency's" would have achieved had it made systematic use of AI models in the 2016 US presidential elections. Indeed Russia has since then used synthetic content in its aggressive information operations, for instance in West Africa. Below the state, terrorist and extremist groups explicitly aiming at precipitating the perceived collapse of their enemy society (such the US "accelerationist" far-right; see Newhouse 2021) are also likely to consider the technology, and researchers have already shown that language models are good at generating credible extremist prose of various ideologies (McGuffie & Newhouse 2020; Baele 2022). But perhaps more worryingly, the proliferation of synthetic content in segments of a target society is likely to achieve a deeper impact: that of polluting the information system to the point of making truth and falsehood undistinguishable, propelling the problem of "fake news" to new heights. In a world "where fakes are cheap, widely available, and indistinguishable from reality", write Allen and Chan (2017, 30-31), "AI forgery capabilities will erode social trust, as previously reliable evidence becomes highly uncertain" (Allen and Chan 2017, 30), creating profound destabilization.

A second use of synthetic content in information warfare is to *bolster support at home* (as opposed to divide abroad), especially in view of inflaming public opinion in preparation of conflict. For Chesney

and Citron (2019a), deepfakes might indeed "best be used to box in a government through inflammation of relevant public opinion". A synthetic audio recording or video deepfake featuring a rival state's political leader insulting the local population, made and circulated as "evidence" of the offence, could garner support for an aggressive "response" – a new way to trigger the kind of collective emotional processes well known to IR scholars (e.g., Hall 2017). Chesney and Citron's scenarios include a deepfake of an American general burning a Quran, used by a non-state actor to heat up its base and reinforce anti-American anger, as well as a synthetic video "leaked" in Iran featuring the Israeli prime minister secretly recorded planning the assassination of an Iranian leader. In a different vein, deepfakes could be used to enhance a terrorist organization's resilience by "evidencing" that its leader is still alive – and giving instructions – despite news of the contrary.

Third, synthetic content could *generate significant misperceptions* of rival states' respective capabilities, with generative AI becoming a new source to the classic sources of misperceptions identified in Jervis' seminal work (e.g., 1976; 1988). Deepfake satellite imagery showing columns of armoured vehicles or ICBM launching systems can be crafted and disseminated in order to generate wrong assessments of capabilities and prompt/prevent action in the desired way; AI could be used to counterfeit official document and disseminate them on the internet, for example "evidencing" war crimes or high-level corruption. Fitzpatrick (2019) explains how a Carnegie Corporation – IISS simulation involving one such situation led participants to the brink of a US-China war; in the simulation, a rogue non-state actor used various types of deepfake "evidence" to purposefully escalate tensions between the US, Russia, and China. In the scenario, the third party benefited from the fact that "collecting and processing the intelligence [to check the veracity of information] would take precious time during an escalatory crisis in which AI algorithms were urging immediate response actions" (Fitzpatrick 2019, 83), as well as from the fact – highlighted in previous research – that "those giving and receiving orders struggle to know which communications (written, video, audio) are authentic" (Allen and Chan 2017, 33).

Finally, a fourth way through which synthetic content can enhance information warfare is as part of a *strategy of individual intimidation*. Most of Diakopoulos and Johnson's (2020) scenarios deployed to identify the ethical issues raised by deepfakes during elections actually pertain to this strategy, from threats made to candidates to release deepfake porn featuring them, to false testimonials "evidencing" an extra conjugal affair. Chesney and Citron (2019a) similarly evoke the use of synthetic content to blackmail diplomats and politicians and support reputational sabotage campaigns. It is easy to imagine the blackmailing or damaging power of that kind of "evidence" when the fake nature of the audio/video is impossible to spot.

In sum, while prognoses on the use of AI in information warfare still rest on hypothetical scenarios, the potential for action is so apparent that these scenarios will undoubtedly come to life. Synthetic content is a potent new computational propaganda weapon that can contribute to enhance (or detract) political actors' power in a growingly contested international arena; combined with other digital tools such as social media apps and bots, deepfakes already enhance attempts to weaponize information. Just like other earlier technologies from the printing press to photo-editing software, generative AI will be harnessed to intensify information operations.

Interestingly, this development comes with an asymmetric advantage for non-liberal autocracies, a disadvantage for targeted societies with higher levels of societal fragmentation, and a thickening of the "post-truth" fog already causing damages to democracy. As such, generative AI appears to be the exception among AI technologies in terms of how it alters IR dynamics: while other contributions to this Forum tend to depict AI as more likely to reinforce existing power differentials than redistribute supremacy, deepfakes are by no means the weapon of the rich and powerful – quite the opposite. Generative models enable states and non-state actors to achieve power gains unavailable to them prior to the technology, gaining efficiency, reach, and impact in their efforts to dent more potent rivals. Existing safeguards to malicious generative AI implemented by major Western corporations fail to

prevent questionable uses,[24] offering multiple possibilities to the "weak" to inflict damage to the "powerful".

Human vs Artificial Intelligences: Some Implications for Strategy

KENNETH PAYNE
King's College London, UK

Artificial Intelligence is already impacting decision-making in strategy, with more far-reaching changes to come. In two parts, this contribution to the Forum offers a conceptual exploration of some salient implications of this important evolution. The first contrasts human and AI decision-making, especially when it comes to gauging social context and intent, which humans do via 'theory of mind'. The second part considers recent developments in one particular type of AI, large language models, concluding that notwithstanding the fundamental AI-human differences highlighted in the first part, some sort of artificial 'theory of mind' may soon be possible.

This approach to understanding AI is situated within a longstanding tradition in strategic studies that focuses on individual agency and adversarial decision-making as central to understanding conflict dynamics between groups. This epistemology overlaps somewhat with political psychology and foreign policy decision-making, as well as strands within classical realism and constructivism, notably where both are concerned with the ways in which personal identity constitutes an actor's rationality. The impact of AI on strategic behaviours will surely be broader than this, affecting both structures and processes within international society in ways that could for example alter the balance of power or shift the offence-defence balance through their effects on military potential. Nonetheless, the focus here is on agency – and the contention is that AI's impact will be particularly radical insofar as reconstituting agents itself alters the very fabric of international society.

Human and machine intelligence contrasted

Human and machine intelligences both come in a variety of forms (and the AI of today is only a narrow sliver of the possibilities for tomorrow's AI). They are, however, recognisably distinct on two dimensions – their architecture and, related, the essential motivation that underpins intelligence. For humans, an embodied, biological intelligence is motivated at a fundamental level by the goals of natural selection – survival and reproduction (of the genome). AI could conceivably emulate that, with biological architectures and some facsimile of algorithmic selection both feasible. Biological microbots have been demonstrated, as have neural nets made from real neurons (Kriegman et al. 2021; Rigby et al 2019). But for that to be fully realised, scientists would have essentially created life, and it would be artificial in the same way that a domestic dog breed is. Nonetheless, its intelligence would still differ from ours in the way that ours does from other animals – which is to say, profoundly (see Ball 2022). In the meantime, today's AI has no innate biological imperative to survive and reproduce. The net result is a *qualitatively* different decision-making process. That includes decisions about war. Tellingly, war has shaped both types of intelligence – AI via state direction of basic and applied research, human via evolution (Payne 2018; 2021).

When it comes to *human intelligence*, evolutionary psychologists have identified a range of selection pressures acting on the brain and mind, among which is the threat of violence, including inter-group

[24] For example, a YouTube tech influencer fine-tuned the GPT-3 on a far-right forum, producing extremist content that he then massively used to "contribute" to the forum's discussion; this prompted a letter signed by over 300 AI researchers condemning his actions (Liang & Reich 2022). For an account of this case, read Murphy (2022).

conflict. Cooperation and coordination are vital for effective warfare. If violence is sufficiently chronic, and if the returns on cooperation are sufficiently high, that's unsurprising (Choi and Bowles 2007). A distinctively *strategic* intelligence results. Groups are essential for individual survival, and group competition is a factor driving cultural evolution. The group with the best ideas, capacity to adapt and learn prospers, including groups that can figure out how to win their conflicts. Human intelligence is therefore both embodied and encultured – anchored in natural selection, and answering its challenge by cultural evolution (Henrich 2015). Humans are capable of flexible learning, dealing with abstract ideas; they have multiple, flexible identities and narratives. In evolutionary terms, they have over-invested resources in social intelligence at the expense of other attributes, like physical strength and perhaps even pattern recognition mental abilities (Martin et al. 2014). Our 'default brain network' spends idle moments rehearsing the future and repackaging the past – often with social goals in mind (Lieberman 2013). Humans have an arsenal of cognitive heuristics, or shortcuts, to facilitate 'good enough' decisions – and many of these relate to navigating our social environment. While our decision-making toolkit includes conscious reflection, including about social relations, often these heuristics are unconscious. In IR/strategic studies, considerable attention has- therefore been dedicated to the ways in which these cognitive processes result in sub-optimal decision-making. Evolutionary minded scholars point to a mismatch between the environment they evolved within and the modern world, of superpowers and nuclear missiles. But in many dimensions, strategy retains core ingredients – decision-making still takes place under uncertainty, guided by emotions and heuristics, and above all, there is a premium in understanding the minds of others – know yourself and know your enemy, as Sun Tzu advised.[25]

Machine intelligence is radically different, and the implications for strategy are profound. Machine learning, the dominant AI paradigm today, is about task optimisation, but agnostic about what the task is (contra natural selection). In theory, it could be tasked with optimising understanding of adversary intentions – but it would go about doing so in a very different way to humans. Underpinning modern AI is pattern recognition, often in large datasets; the model learns by adjusting its algorithm during training so that it produces an optimised output, given an input. At a neuronal level, the mammalian brain does something ostensibly similar – strengthening connections that work well together, and then applying them to new situations. But the comparison is flawed: at the psychological level evolution and biology have produced social intelligence, including empathy – whereas AI is not embedded in that milieu. We cannot readily express the richness of human empathy in a reward function that a machine learning algorithm could optimise. Indeed there is no obvious single value, or even vector of values, in the human process of social understanding that could be captured.

We often use humans as yardstick for AI research, as does the UK government (2001) in defining AI as "machines that perform tasks normally requiring human intelligence." But human intelligence is the wrong yardstick. AI comes up short on tasks that are easy and instinctive for evolved human minds yet computationally demanding, a phenomenon known as the Moravec paradox (Moravec 1988). As originally expressed, the paradox emphasized physical tasks – like grasping objects. But these are increasingly attainable with today's computationally powerful AI. Social cognition, which humans attempt instinctively, if imperfectly, is much more taxing, and today's AI falls short.[26] Nonetheless, AI *is* superintelligent in some dimensions, surpassing by far human intelligence in computational speed, accuracy of memory, and pattern recognition – and all these can be useful decision-making attributes, depending on circumstances. Some decision-making tasks in national security are thus amenable to AI's decision-making.

In earlier work (Payne 2021), I have argued that AI would accordingly be good for tactical decisions, which often boil down to manoeuvre in time and space, the direction of accurate firepower, and

[25] On the strategic advantages of cognitive heuristics, see Johnson 2020.
[26] See Crawford 2021. Emotions are to a considerable degree socially constituted, which further problematises computing them; see also Mesquita 2022.

questions of resupply (Payne 2021). These are problems that require control and information processing, but relatively little understanding of context and meaning. At its most abstract, intelligence is indeed about *control* (of an agent in relation to its environment) and effective *information processing* to facilitate that control. Brains, biological and artificial, furnish agents with a way of navigating change and uncertainty. Where the goals of the agent can be expressed as a reward function to optimise, AI is valuable, and increasingly so for physical control tasks or information processing where there are large data, time constraints, and subtle correlations/patterns to identify. As evoked earlier in this Forum, AI has thus mastered multiple strategy board- and video-games, defeating the best players even in multiplayer settings like poker, which rests on asymmetric knowledge and contains an element of chance (Brown & Sandholm 2019), complex videogame universes such as StarCraft 2 (Vinyals et al. 2019), or Diplomacy, which requires a constantly adapting blend of cooperation and competition (Anthony et al. 2020). Clearly, pattern-matching and optimisation can perform superbly, even in hugely complex "toy universes" which seemingly require several elements of strategy.

Yet for all their brilliance, the gameplaying AIs are engaged in supercharged *tactics*. When the task is social understanding in the real universe, AI still struggles. There are multiple possible goals to optimise, some in tension with others, and the rules by which they might be optimised are too opaque for calculation. Strategy, the higher-level direction of war, requires a richer understanding of context and meaning. Humans attempt control via information processing along different lines to machines, in ways that are more amenable to strategy, for which it partly evolved. We seek to understand and find social meaning in the environment. And we imbue our decision-making with emotional valence, often social emotions. The value that we ascribe to possible choices *feels* like something, and the social dimension of such feelings sometimes amounts to moral judgment (Damasio 1999). So, in International Relations, no matter the "realist" urge for an objective "national interest" to guide strategy, human groups constitute identity, and their values are inescapably bound to their interest.

Yet I now see a flaw in this argument developed in earlier work. The question is not simply a question of AI's tactical nous and strategic limitations; rather, the distinction becomes that between decisions where social meaning is integral versus decisions where such context is less important. In many tactical decisions, context *is* less important – and the task is more straightforwardly translated into reward functions for machine optimisation. There is, therefore, scope for autonomous aircraft and submersibles, or blisteringly fast autonomous cyberwarfare – and there is an efficiency logic driving their development today. However, context and values *can* weigh, even in tactical decisions. Consider:

- How much risk to its human pilot should an AI 'loyal wingman' machine accept to win a dogfight?
- Is that child soldier reaching for a gun to surrender it, or to shoot it?

The first scenario requires a value judgment about just war – balancing the consequentialist logic (what are the payoffs from 'victory'?) against deontology (what is my duty of care to the pilot?). And both scenarios involve a judgment of intention (some 'theory of mind' as psychologists call it) – what is in the mind of the enemy pilot, or the child reaching for the gun? You *could* make a decision without considering such things – by pattern matching and optimising like an AI, but that decision would be reached on qualitatively different grounds. The challenge is akin to the problem facing a driverless car – not the trivial one of navigating streets, but the complex one of navigating road users.

Large language models as artificial strategists

This is where I would have ended the paper if writing a year ago, when I believed that strategy required creative intelligence and a "theory of mind" that were beyond AI: machines alone could not contribute much to strategy beyond analysis, wargaming, and context-light automation. Since then, however, a series of dramatic developments in one area of AI have taken place, with far reaching consequences for national security decision-making. Large language models (LMs) and more generally generative AI, in particular, have demonstrated an ability to create realistic prose and convincing images in response to text cues inputted by humans – a critical shift detailed in the previous contribution to this forum. In doing so, they have expanded the scope of machine creativity – far beyond the sort of bounded "toy

universe" of the chess board, into terrain until recently thought of as intrinsically human. In this section, I explore the key implications of this development for strategy after providing a couple of illustrations.

Motivated by the announcement of a UK MoD review of the appropriate balance between surface and submarine combatants, I asked what was then the most powerful model generating synthetic visual imagery – OpenAI's DALL-E – to depict "the future of naval warfare". The output, provided below in Figure 1, evidently favoured a balanced fleet.

Figure 1. Dall-E's response to a prompt to describe "the future of naval warfare".

While it is fatuous to make policy on basis of DALL-E images, weapon design and operating concepts *are* proceeding with AI assistance, including via synthetic environments. And one way to build trust in their judgments is for an AI to account for itself in plain English. So, I asked the popular LM GPT-3 for its take:

- *Me:* What is the future of naval warfare?
- *AI:* Submarines are expected to play a more important role in the future of naval warfare, as they can operate undetected and launch surprise attacks. Surface combatants, on the other hand, are more vulnerable to attack and will likely become less important.

It sounds plausible and could perhaps be ammunition for those already inclined to favour submarines. But refreshing the answer brought a diametrically opposed response:

- *Me:* Will AI favour surface combatants or submarines?
- *AI:* It's tough to say for sure, but AI can provide a lot of advantages in both areas, such as improved situational awareness and better targeting capabilities. So I think we'll see AI being used in both types of naval warfare.
- *Me:* But submarines can travel faster, and are harder to detect than surface ships. Shouldn't we buy more of them?
- *AI:* There are definitely advantages to submarines, but I think it really depends on the situation. Surface ships have their own advantages too, such as being able to carry more weapons and having better protection from attacks.

Both arguments rest on superficial analysis, unlikely to persuade an expert. But more than that, there's no conceptual thinking going on at all – just probabilistic calculation of what words go together best. The model is a formidable statistical calculator, generating prose on the basis of probabilities established in earlier training data. The answer isn't anchored in our human existence – embodied and affective. Recently Google's model LaMDA convinced its engineer that it is sentient. It's likely not, but perhaps the episode best demonstrates that the famous "Turing test", where a human converses blindly with another agent, is a test of human, not machine intelligence (Fluckinger 2022). Three implications of this situation for national security decision-making stand out.

First, AI like this will make *some* contribution to strategic decision-making. Perhaps only in the widest sense of shaping the decision-making milieu. Human reality is in some measure socially constructed – sometimes in specialist epistemic communities, sometimes more broadly, whether via popular culture or public discourse. Increasingly, AI will help construct both the wider discourse and

the narrower technical/expert discussion. It won't be making decisions for us in such cases, but rather its creativity will co-constitute the human imagination, becoming deeply intrenched in our lives. Machines like this won't simply discuss strategy intelligently with us, they will have shaped our worldviews and preferences over years of consuming their output.

Second, regardless of *how* the machine arrives at its answer, it seems possible that humans will invest the answer and perhaps also the explanation it furnishes with a degree of plausibility, like LaMDA's engineer. That might be especially true where it is hard for humans to understand the concepts involved themselves. Consider an AI acting as translator between code and natural language, as it already happens. That task is certainly useful, allowing non-specialists to program, and to understand what others have coded, but is the explanation a faithful rendering? Such an AI could be prompted to, for instance, design new platforms and operational concepts for autonomous and crewed submarines, and then, having done so, explain to the humans what they are seeing and how it arrived at those conclusions. But the explanation, no matter how plausible, need not be faithful to the underlying logic of the weapon system, just to the conventions of grammar and statistical relationships between words.

A third, final implication of LMs for strategy stems from their use as "binding' agent" between constituent AIs, demonstrated by Google Brain's Socratic Model (Zeng et al. 2022). The "binding problem" refers to the way in which an integrated phenomenological experience somehow emerges from the underlying architecture of the brain – particularly given sub-circuits engaged in different cognitive tasks. The way in which the human brain does this is uncertain, but it's certainly not exclusively via language, not least because that's limited to conscious information processing, notwithstanding the interior monologue most of us experience as fundamental to conscious experience. Regardless, natural language might be a good proxy for binding together discrete machine intelligences. It would imbue the system with a degree of flexibility and perhaps even a degree of 'common sense' conceptual reasoning that combines the best features of machine learning and symbolic logic. The result still won't be human 'common sense,' embodied in our evolved biology; but something else – a form of reasoning that is at once inductive, based on incoming data, and also deductive, grounded in its training in our natural language and the logical concepts therein. For instance, while an AI model draws the future fleet in response to a text prompt as in the illustration above, another model can look at the picture, and describe it in natural language – perhaps English, perhaps code, perhaps even a language the machine creates itself. More radically, other LMs can be listening to a human's verbal output, reading their text (or any related digital footprint), even watching their body language, and integrating all this together to reason about their intentions. In short, a model like this is a possible route to more general-purpose AI. To be sure, the result is qualitatively different from human metacognition, in both method and output. The machine lacks the affective dimension that facilitates our construction of other minds. And the extent to which LMs can generate human-like representations of meaning on the basis of natural language is as yet unclear. Language contains logical propositions, but can LMs extract that reliably enough from the corpus of training data? Certainly, today's LMs are unreliable interlocutors – producing, as in the example above, diametrically opposed analyses of naval warfare only seconds apart. Nonetheless, even an inconsistent machine like that will have military utility as a strategic oracle, generating possible options for human strategists, especially in challenging received wisdom. LMs, then, open the possibility for new forms of creativity; machine and hybrid. And they promise to expand the role of AI in decision-making far beyond the tactical, context-light, narrow AI of today.

Today, there are profound differences between human and artificial intelligence, notably in the ways in which each gauges the actions of other intelligent agents, and weighs important strategic issues (how will you behave? Should I cooperate with you?). Yet language models muddy the waters. An LM recently insisted to me that it was empathic, a skill that could be learned rather than one resting on biology, and that as a result it was good at intuiting what other people were thinking – the extent to which that's true will be of immense strategic importance.

Conclusions: Charting the Challenges of AI IR

EUGENIO V. GARCIA
Ministry of Foreign Affairs, Brazil *

Technology has always been a key element of economic and military power in world politics. Although there is a burgeoning literature on the impact of current technological developments on international relations, much remains to be done in IR as a discipline to theoretically conceptualize and integrate technology into broader views of world order (McCarthy 2018). Nuclear weapons, for instance, were the outcome of scientific and technological research efforts and, when they made their first appearance back in 1945, completely disrupted traditional approaches to warfare. How far will AI reconfigure the world as we know it?

As explained above in different ways, AI is an umbrella term for an array of computational techniques. The contributions to this forum have acknowledged this diversity to highlight some of their implications for international relations and appraise these implications in a nuanced and cautious way, stressing uncertainty.

Baele provided an analysis of risks surrounding the coming era of widespread fake content, indistinguishable to the human eye from real ones. We may soon find ourselves in a state of permanent post-truth fog, in which uncertainty would be the norm – a situation holding serious implications; both domestic and international. Uncertainty was precisely what Whyte, Cuomo, and Jensen identified as one of the major challenges IR scholars face while trying to predict where AI is heading as far as international security is concerned. Coping with ambiguity in this domain will require an interdisciplinary approach capable of navigating the complexities of a hyperactive battlefield, in a satisfactory way, from a military, ethical, legal, and humanitarian perspective. But changes brought about by military AI will alter not only combat operations, which are expected to become dominated by high-speed engagements and increasing autonomy, with less human involvement. AI applications will likely be adopted in strategic planning, command and control systems, intelligence, surveillance and reconnaissance activities, logistics, supply chain, and other tasks as a decision-support tool or enabler to improve efficiency. In his contribution, Payne cogently raised some dilemmas and quandaries we encounter when machines take over and humans are removed from decision-making in warfare and strategy. Our embodied, biological intelligence, he argued, is embedded and shaped by social and cultural factors that form the basis of our common existence as a society. The model of reality we build since our childhood rests upon a shared human experience, which machines have not been able to emulate (so far). This is why, inter alia, any advice of strategic value from an AI oracle should be taken with a grain of salt and (time permitting) be subject to human scrutiny. Indeed, the relationship with the Other (a foreigner, a stranger) is central to the concept of the "international", as in the case of inter-group conflict and the Us-Them divide. Applying a theory of mind to grasp what others might be thinking would be hard (if not impossible) for machines, notably probabilistic deep learning models lacking any understanding of the context and unable to perform actions requiring human-like symbolic knowledge and reasoning (Marcus and Davis 2019). As AI systems increasingly generate non-human outputs and encroach upon cognitive and creative tasks previously considered to be a preserve of our brains (large foundation models being a case in point), seeing "the Other in the machine" will become more common for experts and ordinary citizens alike.

As to attempts at regulating the technology, the need for AI governance at the international level has been recognized to address key emerging issues and seek effective tools to mitigate risks and prevent harm (Tinnirello 2022). These may include human-centered and trustworthy AI for the common good

* PhD International Relations, University of Brasilia; Deputy Consul General of Brazil in San Francisco. The views expressed herein are those of the author. Research work available at https://eugeniovargasgarcia.academia.edu/researchy.

by means of norms, policies, and institutions in the interest of society to prevent outcomes that are alien to human needs and aspirations. Several topics are currently under discussion (Boddington 2017). How to ensure transparency and accountability in relation to responsibility issues, lack of oversight, or hidden layers in opaque deep learning models? Is technically feasible to develop models that are fully understood, explainable, and open to inspection, so that users are properly informed of any unreliable black-box outputs? Should governments set minimum standards and best practices for the deployment of AI systems and combat biased algorithms, stereotyping, flawed conclusions in decision-making, or misleading recommendations, thus strengthening fairness, impartiality, and non-discrimination? Can we avoid ethical pitfalls, legal issues, and reputational backlash by embedding ethics in the product life cycle, including software designed to promote diversity and inclusion, or securing meaningful human control when delegating critical decisions to AI?

Additionally, safety measures, technical standards, performance metrics, robust engineering, and other security mechanisms have been suggested to deal with problems of unreliability, malfunction, machine failure, accidents, adversarial attacks, hacking, and terrorism. It will be essential to confronting from the outset harmful unpredictability, predatory beta-testing, and practices of data colonization, possibly through codes of conduct, regimes, or norms of responsible AI. In the long run, some sort of international coordination may be called for to secure friendly and beneficial AI, in case artificial general intelligence (AGI) is ever achieved with no or little alignment with our human preferences as a species (Russell 2019).

Yet, despite a relative consensus on high-level principles, such as those contained in UNESCO's (2021) *Recommendations on the Ethics of AI*, global governance remains fragmented and limited. There is disagreement on how to operationalize these ideas when it comes to policy and implementation, moving from soft law to enacting concrete legislation and establishing mechanisms of compliance. In addition, the overall political environment has not been conducive to more ambitious agreements due to diverging views among major players, great-power competition, and a continuing crisis of confidence in international institutions. To be sure, AI governance is made more difficult by prevalent uncertainties and ambiguities, as warned throughout this Forum. As a general-purpose technology, AI is not an artifact or a single, perfectly identifiable device, but a myriad of computing processes, largely conducted by dual-use, open-source software, which makes any regulation much more challenging.

An example of these difficulties is the deadlock in the Geneva deliberations of the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems, under the UN Convention on Certain Conventional Weapons (CCW), the most important ongoing multilateral discussion on AI, peace, and security (UN Office for Disarmament Affairs 2022). The high contracting parties to the CCW Convention adopted 11 guiding principles in 2019, with the understanding that they should be further developed to build substance toward a normative and operational framework on these weapons. Many delegations, including most developing countries, support an international legally binding instrument to address concerns posed by fully autonomous weapons. However, the GGE meetings have not been able to reach a compromise (decisions must be made by consensus) and, as a result, questions remain whether existing norms of international humanitarian law are sufficient to guard against unacceptable risks, protect human dignity, and ultimately maintain human control over the use of force (Scharre 2018).

Against this background, many developing countries face the prospect of lagging behind and watching the technological gap vis-à-vis tech-leading powers grow wider (Garcia 2021). Inequality and wealth disparities among nations are not new phenomena and are well-known to structure the international system; at times, however, they can reach extraordinary proportions, dislocating the balance of force in a given interaction. European world expansion from the sixteenth century onwards was facilitated by technological superiority and resulted in the subjugation of large populations across the world. A new era of algorithmic hegemony and digital neo-colonialism may be in the making if asymmetries become deeper and more pervasive. Recent studies, such as MIT *Technology Review*'s series on AI colonialism (Hao 2022), have been investigating how Western companies extract data and

exploit cheap labor in poor countries, in a manner that resembles old patterns in history by putting vulnerable communities in a distressing situation. In that way, AI would not indeed not revolutionize IR – quite the opposite. E.H. Carr, cited earlier in this Forum, famously suggested that power comprised three key components: military capability, propaganda, and economic wealth. Whilst other contributors stressed the former two, it is important not to ignore the last one, in line with the emerging work on the international political economy of AI (e.g., Keskin and Kiggins, 2021). Cummings and colleagues (2018, v) might be right in claiming that "AI is likely to reshape what work looks like, but […] is unlikely to fundamentally change underlying economic power structures", it remains that specific effects on the "periphery" of the AI-race ought to be located and studied. Sloane's (2021) study of Latin American AI economy, for instance, exposes how US-Chinese competition for AI-Cloud resources "accentuates problematic processes of uneven and combined development in the region.

The weaponization of AI, which underlies much of this competition, has been fueled by great-power rivalry and fears of conflict with near-peer strategic competitors. Notwithstanding these concerns, the likelihood of AI-enabled weapons being deployed first in the Global South seems greater. Grey-zone conflicts in faraway war theatres may be a convenient testing ground for cutting-edge systems never used before, particularly in low-intensity, small-scale, proxy wars. In these scenarios, technological imbalances could give a decisive advantage to the stronger side engaged in remote, robotic warfare, inasmuch as their troops would mostly be replaced by combat machines and autonomous devices.

If technology is socially constructed, "machine neutrality" should be regarded with caution, since decisions to design, build, and deploy AI systems are not made in a vacuum. Rather, they reflect the mindset, bias, and social background of those developing the engineering project, which may arguably be different for people living in other countries, especially in the Global South, with distinct local needs and priorities. The social-technical character of technology cannot simply be dismissed. How best to govern AI at the international level would benefit from a cross-cultural dialogue based upon goal-oriented cooperation among all stakeholders, with due regard to inclusion, gender balance, and equitable geographical representation.

The global debate on AI IR, therefore, will certainly involve short-, near-, and long-term consequences and potentially affect every country on the planet, perhaps mobilizing public opinion worldwide, as much as climate change did and still does. The work of diplomats will change too (Bjola 2020; Buch et al. 2022). The research agenda for the future will need to tackle not only the struggle for power and AI supremacy among great powers, as evidence of a new Tech Cold War already suggests, but also the impact of the prevailing Wild West attitude concerning the international governance of AI, still lacking normative frameworks and appropriate multilateral settings for states, the private sector, civil society, and other stakeholders to meet, discuss, and negotiate AI policymaking at the global level. To this end, more scholarly contributions should be welcome in this collective effort to find the solutions needed at this critical juncture and beyond, and furthermore to evaluate how AI alters, not only IR, but our own tools to make sense of global politics: IR theory (e.g., Kiggins 2018).

References:

AGRAWAL, AJAY, JOSHUA GANS AND AVI GOLDFARB, eds. 2019. *The Economics of Artificial Intelligence: An Agenda*. Chicago: Chicago University Press.

ALLEN, GREG, AND TANIEL CHAN. 2017. Artificial Intelligence and National Security. *Belfer Center Studies*, July 2017.

ANTHONY, THOMAS, et al. 2020. Learning to Play No-Press Diplomacy with Best Response Policy Iteration. *Advances in Neural Information Processing Systems* 33 (NeurIPS 2020).

ARKIN, RONALD. 2010. The Case for Ethical Autonomy in Unmanned Systems. *Journal of Military Ethics* 9(4): 332-341.

ASARO, PETER. 2020. Autonomous Weapons and the Ethics of Artificial Intelligence. In LIAO, MATTHEW, ed. *Ethics of Artificial Intelligence*. Oxford: OUP, pp.212-236.

AXELROD, ROBERT. 1984. *The Evolution of Cooperation*. New York: Basic Books.
AYOUB, KAREEM, AND KENNETH PAYNE. 2016. Strategy in the Age of Artificial Intelligence. *Journal of Strategic Studies* 39(5-6): 793-819.
BAELE, STEPHANE. 2022. The Threat of Language Models. *CREST Guide*.
BALL, PHILIP (2022) *The Book of Minds: How to Understand Ourselves and Other Beings, from Animals to Aliens*. London: Picador.
BARFIELD, WOODROW, AND UGO PAGALLO, eds. 2018. *Research Handbook on the Law of Artificial Intelligence*. Cheltenham: Edward Elgar.
BILLS, GWENDELYNN. 2014. LAWS unto Themselves: Controlling the Development and Use of Lethal Autonomous Weapons Systems. *George Washington Law Review* 83(1): 176-209.
BJOLA, CORNELIU. 2020. *Diplomacy in the Age of Artificial Intelligence*. EDA Working paper, Emirates Diplomatic Academy, Abu Dhabi.
BODDINGTON, PAULA (2017) *Towards a Code of Ethics for Artificial Intelligence*. Cham: Springer.
BODE, INGVILD, AND HENDRIK HUELSS. 2018. Autonomous Weapons Systems and Changing Norms in International Relations. *Review of International Studies* 44(3): 393-413.
BODE, INGVILD, AND HENDRIK HUELSS. 2022. *Autonomous Weapons Systems and International Norms*. Montreal: McGill University Press.
BOLUKBASI, TOLGA, KAI-WEI CHANG, JAMES ZOU, VENKATESH SALIGRAMA, AND ADAM KALAI. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *30th Conference on Neural Information Processing Systems (NIPS 2016)*.
BOMMASANI, RISHI., ET AL. 2021. *On the Opportunities and Risks of Foundation Models*. Palo Alto: Stanford University Press.
BOSTROM, NICK, AND ELIEZER YUDKOWSKY. 2014. The Ethics of Artificial Intelligence. In FRANKISH, KEITH, AND WILLIAM RAMSEY, eds. *The Cambridge Handbook of Artificial Intelligence*. Cambridge: CUP.
BOULANIN, VINCENT, ed. 2019. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*. Stockholm: SIPRI.
BOULANIN, VINCENT, AND MAAIKE VERBRUGGEN. 2017. *Mapping the Development of Autonomy in Weapons Systems*. Stokholm: SIPRI.
BRANCH, JORDAN. 2021. What's in a Name? Metaphors and Cybersecurity. *International Organization* 75(1): 39-70.
BROOKS, ROSA. 2015. In Defense of Killer Robots. *Foreign Policy*, https://foreignpolicy.com/2015/05/18/in-defense-of-killer-robots/.
BROWN, TOM, ET AL. 2020. *Language Models are Few-Shot Learners*. San Francisco: OpenAI. arXiv:2005.14165v4.
BROWN, NOAM, AND TUOMAS SANDHOLM. 2019. Superhuman AI for Multiplayer Poker. *Science* 365(6456): 885-890.
BUCH, AMANDA, DAVID EAGLEMAN, AND LOGAN GROSENICK. 2022. Engineering Diplomacy: How AI and Human Augmentation Could Remake the Art of Foreign Relations. *Science & Diplomacy*, https://www.sciencediplomacy.org/perspective/2022/engineering-diplomacy-how-ai-and-human-augmentation-could-remake-art-foreign
BUCHANAN, BEN, ET AL. 2021. *Truth, Lies, and Automation. How Language Models Could Change Disinformation*. Washington D.C.: Georgetown University Centre for Security and Emerging Technology.
BUZAN, BARRY, OLE WAEVER, AND JAAP DE WILDE. 1998. *Security: A New Framework for Analysis*. Boulder: Lynne Rienner.
CARR, EDWARD H. 1946 (1939). *The Twenty Years' Crisis: 1919–1939: An Introduction to the Study of International Relations. Second Edition*. London: MacMillan.
CHESNEY, BOBBY, AND DANIELLE CITRON. 2019a. Deep Fakes: Looming Challenge for Privacy, Democracy, and National Security. *California Law Review* 107(6): 1753-1820.
CHESNEY, BOBBY, AND DANIELLE CITRON. 2019b. Deepfakes and the New Disinformation War: The Coming Age of Post-truth Geopolitics. *Foreign Affairs* 98(1): 147-155.
CHOI, JUNG-KYOO, AND SAMUEL BOWLES. 2007. The Coevolution of Parochial Altruism and War. *Science* 318(5850): 636-640.
COECKELBERGH, MARK. 2020. *AI Ethics*. Cambridge, MA: MIT Press.

CRAWFORD, KATE. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press

CUMMINGS, MARY, HEATHER ROFF, KENNETH CUKIER, JACOB PARAKILAS, AND HANNAH BRYCE. 2018. *Artificial Intelligence and International Affairs. Disruption Anticipated*. London: Chatham House.

Damasio, ANTONIO. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt.

DAVIS, ZACHARY. 2019. Artificial Intelligence on the Battlefield. *Prism* 8(2): 114-131.

DE RUITER, ADRIENNE. 2021. The Distinct Wrong of Deepfakes. *Philosophy & Technology*, online before print.

DIAKOPOULOS N., JOHNSON D. (2020) Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections. *New Media & Society*, online before print.

DOBBER, TOM, NADIA METOUI, DAMIAN TRILLING, NATALI HELBERGER, AND CLAES DE VREESE. 2021. Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *International Journal of Press/Politics* 26(1): 69-91.

DUBBER, MARKUS, FRANK PASQUALE, AND SUNIT DAS, eds. 2020. *The Oxford Handbook of Ethics of AI*. Oxford: OUP.

DUNN CAVELTY, MYRIAM., AND ANDREAS WENGER. 2020. Cyber Security meets Security Politics: Complex Technology, Fragmented Politics, and Networked Science. *Contemporary Security Policy* 41(1): 5-32.

ETZIONI, AMITAI, AND OREN ETZIONI. 2017. Incorporating Ethics into Artificial Intelligence. *Journal of Ethics* 21: 403-418.

FINNEMORE, MARTHA, AND KATHRYN SIKKINK. 1998. International Norm Dynamics and Political Change. *International Organization* 52(4): 887-917.

FITZPATRICK, MARK. 2019. Artificial Intelligence and Nuclear Command and Control. *Survival* 61(3): 81-92.

FLORIDI, LUCIANO, AND MASSIMO CHIRIATTI. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines* 30: 681-694.

FLUCKINGER, DON. 2022. Ex-Google Engineer Blake Lemoine Discusses Sentient AI. *TechTarget*, 27 July 2022, https://www.techtarget.com/searchenterpriseai/feature/Ex-Google-engineer-Blake-Lemoine-discusses-sentient-AI

FRANKISH, KEITH, AND WILLIAM RAMSEY, eds. 2014. *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press.

FRANKLIN, STAN. 2014. History, Motivations, and Core Themes. In Frankish K., Ramsey W. (eds.) *The Cambridge Handbook of Artificial Intelligence*. Cambridge: CUP.

GARCIA DENISE. 2015. Battle Bots. How the World Should Prepare Itself for Robotic Warfare. *Foreign Affairs*, 5 June 2015, available at https://www.foreignaffairs.com/articles/2015-06-05/battle-bots.

GARCIA, EUGENIO. 2019. *The Militarization of Artificial Intelligence: A Wake-up Call for the Global South*. Unpublished manuscript, available at https://eugeniovargasgarcia.academia.edu/research.

GARCIA, EUGENIO. 2021. The International Governance of AI: Where is the Global South? *The Good AI*, available at https://thegoodai.co/2021/01/28/the-international-governance-of-ai-where-is-the-global-south

GILL, AMANDEEP SINGH. 2019. Artificial Intelligence and International Security: The Long View. *Ethics & International Affairs* 33(2): 169-179.

GOMEZ, MIGUEL, AND CHRISTPHER WHYTE. 2021. Breaking the Myth of Cyber Doom: Securitization and Normalization of Novel Threats. *International Studies Quarterly* 65(4): 1137-1150.

GOMEZ, MIGUEL. 2022. Tracing Strategic Preferences in Cyberspace: The Role of Regional and Domestic Strategic Culture. *Comparative Strategy*, online before print.

GRAY, JONATHAN, ADAM LERER, ANTON BAKHTIN, AND NOAM BROWN. 2021. Human-Level Performance in No-Press Diplomacy via Equilibrium Search. *arXiv* 2010.02923.

HALL, TODD. 2017. On Provocation: Outrage, International Relations, and the Franco–Prussian War. *Security Studies* 26(1): 1-29,

HAMMOND, DANIEL. 2015. Autonomous Weapons and the Problem of State Accountability. *Chicago Journal of International Law* 15(2): 652-688.

HAO, KAREN. 2022. Artificial Intelligence is Creating a New World Order. *MIT Technology Review*.

HE, LAURA. 2022. US Curbs on Microchips Could Throttle China's Ambitions and Escalate the Tech War. *CNN*, 31 October 2022, https://edition.cnn.com/2022/10/31/tech/us-sanctions-chips-china-xi-tech-ambitions-intl-hnk/index.html.

HEAVEN, WILL. 2020 OpenAI's New Language Generator GPT-3 Is Shockingly Good – And Completely Mindless. *MIT Technology Review*, 20 July 2020.

HENRICH, JOSEPH. 2015. *The Secret of Our Success*. Princeton: Princeton University Press.

HOIJTINK, MARIJN, AND MATTHIAS LEESE, eds. 2019. *Technology and Agency in International Relations*. New York: Routledge.

HOROWITZ, MICHAEL. 2016. Ban Killer Robots? How about Defining them First? *Bulletin of the Atomic Scientists* 24.

HOROWITZ, MICHAEL. 2018. Artificial Intelligence, International Competition, and the Balance of Power. *Texas National Security Review* 1(3): 36-57.

HOROWITZ, MICHAEL. 2019. When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability. *Journal of Strategic Studies* 42(6): 764-788.

HUDSON, VALERIE. 1991. *Artificial Intelligence and International Politics*. New York: Routledge.

INGRAM, HARORO. 2019. The Strategic Logic of Islamic State's Full-Spectrum Propaganda: Coherence, Comprehensiveness, and Multidimensionality. In BAELE, STEPHANE, KATHARINE BOYD, AND TRAVIS COAN, eds. *ISIS Propaganda: A Full-Spectrum Extremist Message*. New York: Oxford University Press.

JENSEN, BENJAMIN, CHRISTOPHER WHYTE, AND SCOTT CUOMO. 2020. Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence. *International Studies Review* 22(3): 526-550.

JENSEN, BENJAMIN, CHRISTOPHER WHYTE, AND SCOTT CUOMO. 2022. *Information in War: Military Innovation, Battle Networks, and the Future of Artificial Intelligence*. Washington, DC: Georgetown University Press.

JERVIS, ROBERT. 1976. *Perception and Misperception in International Politics*. Princeton: Princeton University Press.

JERVIS, ROBERT. 1988. War and Misperception. *The Journal of Interdisciplinary History* 18(4): 675-700.

JOHNSON, DOMINIC. 2020. *Strategic Instincts: The Adaptive Advantages of Cognitive Biases in International Politics*. Princeton: Princeton University Press.

JOHNSON, JAMES. 2019. Artificial Intelligence and Future Warfare: Implications for International Security. *Defense & Security Analysis* 35(2): 147-169.

JOHNSON, JAMES. 2020. Artificial Intelligence: A Threat to Strategic Stability. *Strategic Studies Quarterly* 14(1): 16-39.

KALOC, JIRI. 2022. Jumbo-Visma Nutrition at the Tour – Machine Learning. *WeLoveCycling*, 12 July 2022, https://www.welovecycling.com/wide/2022/07/12/jumbo-visma-nutrition-at-the-tour-machine-learning/.

KESKIN, TUGRUL, AND RYAN KIGGINS, eds. 2021. *Towards an International Political Economy of Artificial Intelligence*. Cham: Palgrave Macmillan.

KIGGINS, RYAN. 2018. Big Data, Artificial Intelligence, and Autonomous Policy Decision-Making: A Crisis in International Relations Theory? In KIGGINS, RYAN, ed. *The Political Economy of Robots. Prospects for Prosperity and Peace in the Automated 21st Century*. Cham: Springer.

KNIGHT, WILL. 2022. As Russia Plots Its Next Move, an AI Listens to the Chatter. *Wired* 4 April 2022, https://www.wired.com/story/russia-ukraine-war-ai-surveillance/.

KREPS, SARAH, R. MILES MCCAIN, AND MILES BRUNDAGE. 2020. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, online before print, 1-14.

KRIEGMAN, SAM, DOUGLAS BLACKISTON, MICHAEL LEVIN, AND JOSH BONGARD. 2021. Kinematic Self-replication in Reconfigurable Organisms. *Proceedings of the National Academy of Sciences*118(49).

KRIZHEVSKY, ALEX, ILYA SUTSKEVER, AND GEOFFREY HINTON. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In BARTLETT, PETER, FERNANDO PEREIRA, CHRISTOPHER BURGES, LÉON BOTTOU, AND KILIAN WEINBERGER, eds. *Advances in Neural Information Processing Systems* 25 (NIPS 2012).

KO, KUANG-TING, ET AL. 2022. Structure of the Malaria Vaccine Candidate Pfs48/45 and its Recognition by Transmission Blocking Antibodies. *BioRXiv*, doi.org/10.1101/2022.05.24.493318.

LA COUR, CHRISTINA. 2020. Theorising Digital Disinformation in International Relations. *International Politics* 57: 704-723.

LAWSON, SEAN. 2013. Beyond Cyber-Doom: Assessing the Limits of Hypothetical Scenarios in the Framing of Cyber-threats. *Journal of Information Technology & Politics* 10(1): 86-103.

LIANG, PERCY, ROB REICH, ET AL. 2022. *Condemning the Deployment of GPT-4chan*. https://docs.google.com/forms/d/e/1FAIpQLSdh3Pgh0sGrYtRihBu-GPN7FSQoODBLvF7dVAFLZk2iuMgoLw/viewform

LIAO, S. MATTHEW, ed. 2020. *Ethics of Artificial Intelligence*. New York: OUP.

LIEBERMAN, MATTHEW. 2013. *Social: Why our Brains are Wired to Connect*. Oxford: OUP.

LINDSAY, JON. 2021. Cyber conflict vs. Cyber Command: Hidden Dangers in the American Military Solution to a Large-scale Intelligence Problem. *Intelligence & National security* 36(2): 260-278.

MARCUS, GARY. 2018. Deep Learning: A Critical Appraisal. *arXiv:1801.00631*.

MARCUS, GARY, AND ERNEST DAVIS. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon.

MARTIN, CHRISTOPHER, RAHUL BHUI, PETER BOSSAERTS, TETSURO MATSUZAWA, AND COLIN CAMERER. 2014. Chimpanzee Choice Rates in Competitive Games Match Equilibrium Game Theory Predictions. *Scientific Reports* 4 (5182):1-6.

MORAVEC, HANS. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge : Harvard University Press.

MASCHMEYER, LENNART. 2022. Subverting Skynet: The Strategic Promise of Lethal Autonomous Weapons and the Perils of Exploitation. In *2022 14th International Conference on Cyber Conflict: Keep Moving! (CyCon)* pp.155-171. IEEE.

MCCARTHY, DANIEL, ed. 2018. *Technology and World Politics*. Abingdon: Routledge.

MCGUFFIE, KRIS, AND ALEX NEWHOUSE. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. *arXiv:2009.06807v1*.

MESQUITA, BATJA. 2022. *Between Us: How Cultures Create Emotions*. New York: Norton.

MICHAEL, JULIAN, ET AL. 2022. What Do NLP Researchers Believe? Results of the NLP Community Metasurvey. *arXiv:2208.12852*.

MURPHY, MATT. 2022. The Dawn of A.I. Mischief Models. *Slate*, 3 August 2022, https://slate.com/technology/2022/08/4chan-ai-open-source-trolling.html.

NAKAYAMA, BRYAN. 2022. Democracies and the Future of Offensive (Cyber-Enabled) Information Operations. *The Cyber Defense Review* 7(3): 49-66.

NEWHOUSE, ALEX. 2021. The Threat Is the Network: The Multi-Node Structure of Neo-Fascist Accelerationism. *CTC Sentinel* 14(5).

NILSSON, NILS. 2009. *The Quest for Artificial Intelligence*. Cambridge: CUP.

NSCAI (National Security Commission on Artificial Intelligence) (2021) *Final Report*. Available at https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

ÖHMAN, CARL. 2020. Introducing the Pervert's Dilemma: A Contribution to the Critique of Deepfake Pornography. *Ethics and Information Technology* 22: 133-140.

PATERSON, THOMAS, AND LAUREN HANLEY. 2020. Political Warfare in the Digital Age: Cyber Subversion, Information Operations, and 'Deep Fakes'. *Australian Journal of International Affairs* 74(4): 439-454.

PATRIKARAKOS, DAVID. 2017. *War in 140 Characters: How Social Media is Reshaping Conflict in the Twenty-First Century*. New York: Basic Books.

PAYNE, KENNETH. 2018. *Strategy, Evolution, and War: From Apes to Artificial Intelligence*. Washington, DC: Georgetown University Press.

PAYNE, KENNETH. 2018. Artificial Intelligence: A Revolution in Strategic Affairs?, *Survival* 60(5): 7-32.

PAYNE, KENNETH. 2021. *I, Warbot. The Dawn of Artificially Intelligent Conflict*. New York: Hurst.

PEREZ - DES ROSIERS, DAVID. (2021) AI Application in Surveillance for Public Safety: Adverse Risks for Contemporary Societies. In KESKIN, TUGRUL, AND RYAN KIGGINS, eds. 2021. *Towards an International Political Economy of Artificial Intelligence*. Cham: Palgrave Macmillan, pp.113-143.

PETRAEUS, DAVID, AND ANDREW ROBERTS. 2023. *Conflict: The Evolution of Warfare from 1945 to Ukraine*. New York: Harper Collins.

PENG, SHIN-YI, CHING-FU LIN, AND THOMAS STREINZ. 2021. *Artificial Intelligence and International Economic Law*. Cambridge: CUP.

RATHBUN, BRIAN. 2007. Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory. *International Studies Quarterly* 51(3): 533-557.

REEVE, ZOEY. 2021. Engaging with Online Extremist Material: Experimental Evidence. *Terrorism & Political Violence* 33(8): 1595-1620.

RIGBY, MICHAEL, ET AL. (2019) Building an Artificial Neural Network with Neurons. *AIP Advances* 9(7): 075009.

RID, THOMAS. 2012. Cyber War Will Not Take Place. *Journal of Strategic Studies* 35(1): 5-32.

ROFF, HEATHER. 2019. The Frame Problem: The AI "Arms Race" Isn't One. *Bulletin of the Atomic Scientists* 75(3): 95-98.

RUSSELL, STUART. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Penguin.

RUSSELL, STUART, SABINE HAUERT, RUSS ALTMAN, AND MANUELA VELOSO. 2015. Robotics: Ethics of Artificial Intelligence. *Nature* 521: 415-418.

SAMUEL, ARTHUR. 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* 3: 211-229.

SCHARRE, PAUL. 2018. *Army of None: Autonomous Weapons and the Future of War*. New York: Norton.

SCHARRE, PAUL. 2019. Killer Apps: The Real Dangers of an AI Arms Race. *Foreign Affairs* 98: 135.

SCHRITTWIESER, JULIAN, ET AL. 2020. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature* 588(7839): 604-609.

SEOANE, MAXIMILIANO. 2021. Chinese and U.S. AI and Cloud Multinational Corporations in Latin America. In KESKIN, TUGRUL, AND RYAN KIGGINS, eds. *Towards an International Political Economy of Artificial Intelligence*. Cham: Palgrave Macmillan, pp.85-111.

SHARKEY, NOEL. 2010. Saying "No!" to Lethal Autonomous Targeting. *Journal of Military Ethics* 9(4): 369-383.

SLONIM, NOAM, ET AL. 2021. An Autonomous Debating System. *Nature* 591: 379-385.

SOLAIMAN, IRENE, ET AL. 2019. *Release Strategies and the Social Impacts of Language Models*. San Francisco: OpenAI.

TADDEO, MARIAROSARIA, AND LUCIANO FLORIDI. 2018. Regulate Artificial Intelligence to Avert Cyber Arms Race. *Nature* 556: 296-298.

TINNIRELLO, MAURIZIO, ed. 2022. *The Global Politics of Artificial Intelligence*. New York: Routledge.

TUCKER, PATRICK. 2019. The Newest AI-enabled Weapon: 'Deep-faking' Photos of the Earth. *Defense One*, https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/.

UN OFFICE FOR DISARMAMENT AFFAIRS. 2022. Background on LAWS in the CCW, https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw

UNESCO. 2021. *Recommendation of the Ethics of AI*, available at https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

UK GOVERNMENT. 2021. *National AI Strategy*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf

UMBRELLO, STEVEN, PHIL TORRES, AND ANGELO DE BELLIS. 2020. The Future of War: Could Lethal Autonomous Weapons Make Conflict More Ethical? *AI & Society* 35(1): 273-282.

URBINA, FABIO, FILIPPA LENTZOS, CÉDRIC INVERNIZZI, AND SEAN ETKINS. 2022. Dual Use of Artificial-Intelligence-Powered Drug Discovery. *Nature Machine Intelligence* 4: 189-191.

VERDOLIVA, LUISA. 2020. Media Forensics and DeepFakes: An Overview. *arXiv 2001.06564*.

VINYALS, ORIOL., ET AL. 2019. Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning. *Nature* 575(7782): 350-354.

WEIDINGER, LAURA, ET AL. 2021. Ethical and Social Risks of Harm from Language Models. *arXiv:2112.04359*.

WHYTE, CHRISTOPHER. 2020. Beyond Tit-for-tat in Cyberspace: Political Warfare and Lateral Sources of Escalation Online. *European Journal of International Security* 5(2): 195-214.

WHYTE, CHRISTIOPHER. 2020. Problems of Poison: New Paradigms and "Agreed" Competition in the Era of AI-Enabled Cyber Operations. *12th International Conference on Cyber Conflict (CyCon)*: 215-232.

WHYTE, CHRISTIOPHER. 2022. Machine Expertise in the Loop: Artificial Intelligence Decision-Making Inputs and Cyber Conflict. *14th International Conference on Cyber Conflict (CyCon)*: 135-154.

WOOLLEY, SAMUEL, AND PHILIP HOWARD, eds. 2018. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford: Oxford University Press.

YOUNG, KEVIN, AND CHARLI CARPENTER. 2018 Does Science Fiction Affect Political Fact? Yes and No: A Survey Experiment on "Killer Robots". *International Studies Quarterly* 62(3): 562-576.

ZENG AUDY, ET AL. 2022. Socratic Models: Composing Zero-shot Multimodal Reasoning with Language. *arXiv* 2204.00598.

ZHAO, BO, SHAOZENG ZHANG, CHUNXUE XU, YIFAN SUN, AND CHENGBIN DENG. 2021. Deep Fake Geography? When Geospatial Data Encounter Artificial Intelligence. *Cartography and Geographic Information Science*, 48:4, 338-352.