2. The challenge of advanced cyberwar and the place of cyberpeace

Elias G. Carayannis and John Draper

A NOTE ON CHAPTER 2

This chapter provides an analytical review, from the cyber defense perspective, of the existential threat posed by a human intelligence equivalent artificial general intelligence (AGI) and, more particularly, by an above human intelligence artificial superintelligence (ASI). The concept of an ASI as an existential risk has existed within the sphere of science fiction for decades, but existential risk theory was developed in the twenty-first century by Oxford-based researcher Nick Bostrom (2002, 2014), who has championed existential risk prevention policy across multiple fields (Bostrom, 2013).

There is now little doubt that the world is engaged in an artificial intelligence (AI) arms race that includes cyberweapons and cyberwarfare, and potentially cyberwar. According to Nicolas Chaillan, the Pentagon's former first Air Force and Space Force Chief Software Officer, it is a race that China has already won because of its determined investment in this aspect of warfare (Faulconbridge, 2021). The implications of this new reality for defense are enormous, in that Chinese AI domination paves the way for global supremacy in the coming decades. However, following the 2021 National Security Commission on Artificial Intelligence (NSCAI) final report, the United States may finally be beginning to engage, and the fact that most AGI projects are still in the West means that there will likely be a closely fought contest between China and the United States, and possibly Russia, to develop the first AGI.

The chapter reviews the work of AGI researchers like Alexei Turchin, affiliate scholar at the Institute for Ethics and Emerging Technologies, and Seth Baum, executive director of the Global Catastrophic Risk Institute. It endorses their previous research concerning the risks of a militarized ASI. It refers to a substantial body of previous research which focuses on constraining a nascent ('pre-takeoff') AGI internally via mechanisms such as electronic or mechanical sandboxing. However, it focuses mainly on post-takeoff social measures to constrain an ASI. Here, the chapter refers to 'friendly AI' as proposed by Eliezer Yudkowsky, co-founder of the Machine Intelligence Research Institute.

The chapter accepts Yudkowsky's (2001, 2004) and others' basic premise that, for humanity to survive the singularity, an ASI must have some form of moral compass and principles. From a programmer perspective, these principles need to be altruistic and universal. However, the chapter points out that nation state political subversion of an ASI, which is likely inevitable, means that a Chinese ASI could adopt very different principles to that of an American or Russian ASI, leading to the possibility of a preventive attack by one nation state on another to deny it creating a game-changing ASI. Bostrom speculated about addressing this problem via a treaty in 2014 in his work *Superintelligence: Paths, Dangers, Strategies*, and in 2018 Ramamoorthy and Yampolskiy recommended a comprehensive United Nations-sponsored ‘Benevolent AGI Treaty’, a form of ‘cyberpeace’ whereby member states would agree that only altruistic ASIs should be created.

Developing the concept of a ‘Benevolent AGI Treaty’, the chapter suggests, and its authors have drafted, an arms control based-approach to the problem of a militarized ASI. Building on the present momentum at the UN behind the proposed Cybercrime Treaty, the Cyberweapons and Artificial Intelligence Treaty would ban both AI and ASI cyberweapons. However, a major power could conspire to avoid the treaty. Consequently, the chapter, building on international relations theory, specifically conforming instrumentalism (Mantilla, 2017) and American political scientist Glenn D. Paige’s (2009) peacebuilding work on Nonkilling Global Political Science, suggests as a partial solution a treaty of much larger scope, the Universal Global Peace Treaty (UGPT), which is presently under development by a number of peacebuilding non-governmental organizations (NGOs). If such a definitive end point to war were endorsed by the AI powers, depending on these states’ motives for endorsement, humanity could signal to an emergent ASI humanity’s universal endorsement of the foundational principle of peace, the bedrock of cyberpeace.

Depending on how the ASI interprets this signaling, including the inclusion of Nonkilling Global Political Science in the design and articles of the UGPT, it might seek to conform to the principle of nonkilling, and so to minimizing human loss of life, for conforming instrumental reasons. These include the possibility that the ASI could view itself as a member of a common community that includes nation states and humanity in general within which it could seek to fulfill its own goalset in a negotiated partnership. This, as the chapter title indicates, is the ultimate optimization of the chance of peaceful coexistence in a world inhabited by one or more ASIs. The chapter argues that the alternative is cyberwar.

INTRODUCTION

Advanced Cyberwar: The Problem of a Warring Artificial Superintelligence

While some researchers maintain that creating an AGI (i.e., human or above human AI), is impossible, for instance due to the Dreyfussian argument that computers do not grow up, belong to a culture, and act in the world (Fjelland, 2020), others believe a path to an AGI is in sight (Goertzel & Pennachin, 2020; Wang & Goertzel, 2012). Consequently, the international defence community is beginning to seriously consider the national security risk posed by the development of AGI and its implications for international relations:

> Start thinking about artificial superintelligence and engage with the community that has started thinking about actionable options in that part of the option space (also recognizing that this may open up new avenues for engaging with actors such as China and/or the Russian Federation). (De Spiegeleire, Maas, & Sweijs, 2017, p. 107)

This is because of the possibility that a single nation state developing an AGI could ‘lock in’ economic or military supremacy as a discrete ‘end point’ to economic and military competition in international politics, as that state would then be able to prevent a rival AGI being developed and then establish global domination (Horowitz, 2018, p. 54). In this chapter, in line with most thinking (e.g., Green, 2015), establishing global domination through repeated cyberwarfare attacks (on which, see Chelvachandran et al., 2020) to establish supremacy over other nation states is termed cyberwar and to date has not occurred; somewhat paradoxically, its near corollary, cyber defence, definitely does occur (Jahankani et al., 2020). Thus, herein we adopt the narrow definition of cyberwar, while applying the term cyberwarfare to the actions of state actors engaged in economic and financial crimes against other state actors for economic or strategic gain, with the most prominent example of such being North Korea. While some have argued that cyberwar will never occur (e.g., Rid, 2012), this chapter details how the development of militarized AGI and of an ASI would invalidate this proposition.

To begin, the development of AI is a major factor in national security because it can be militarized, can be employed in adversarial contexts as in cyberwarfare, and can provide a decisive advantage in terms of economic, information, and military superiority (Allen & Chan, 2017; Babuta, Oswald, & Janjeva, 2020; National Security Commission on Artificial Intelligence [NSCAI], 2021). As the United States’ (US) NSCAI (2021, p. 159) report states, “The impact of artificial intelligence (AI) on the world will extend far beyond narrow national security applications. The development of AI constitutes a new pillar of strategic competition, and it heightens the competition in existing pillars.” The NSCAI report urges the US attain a state of military AI readiness by 2025. AI is thus important for waging war, perhaps even decisive war, raising the specter of the return of ‘total war’ (i.e., industrial interstate war), which in its last major iteration, World War II, involved genocidal levels of killing (Markusen & Kopf, 2007).

For the major powers, and particularly for the US, AI technological supremacy, founded on economic power, is viewed as paramount to national security, especially with respect to relations with China:

> China possesses the might, talent, and ambition to surpass the United States as the world’s leader in AI in the next decade if current trends do not change. Simultaneously, AI is deepening the threat posed by cyber attacks and disinformation campaigns that Russia, China, and others are using to infiltrate our society, steal our data, and interfere in our democracy. The limited uses of AI-enabled attacks to date represent the tip of the iceberg. (NSCAI, 2021, p. 7)

Military AI is already potentially revolutionary in that it could "accelerate the pace of combat to a point in which machine actions surpass the rate of human decision making, potentially resulting in a loss of human control in warfare" (Congressional Research Service, 2019, p. 37, citing Scharre, 2017). It also constitutes an unpredictable threat, in that "placing AI systems capable of inherently unpredictable actions in close proximity to an adversary's systems may result in inadvertent escalation or miscalculation" (Congressional Research Service, 2019, p. 37, citing Altmann & Sauer, 2017).

While AI researcher Baum (2017) found little evidence of military AGI projects half a decade ago, the 2021 NSCAI report envisages a massive ramping up of US military AI by 2025 and endorses a push towards more general AI, occurring in a future experiencing a societal level of advanced, accelerated adversarial AI-enabled cyberattacks and ubiquitous interstate AI-enabled cyberwarfare, including with autonomous systems, with conflict over matters like intellectual property and technological leadership. It is inconceivable in the present geopolitical atmosphere that a successful American, Chinese, or Russian AGI project, of which dozens are in development, would not immediately be militarized on the pretext of national security and employed to maintain or secure technological supremacy. It is also inconceivable that this quest for technological supremacy would not involve a quest for cyber supremacy (i.e., the capacity to wage decisive cyberwarfare), so as to be able to dominate the tit-for-tat of Russian–US cyber relations (CBS News, 2016), such that an actual cyberwar could be won.

In line with the general thinking of some previous AI researchers (e.g., Totschnig, 2019), we do not see developing an AGI as primarily a technological problem but as a political problem. However, where most authors stress the political problem re humanity's relationship with an ASI, we focus on the political challenge posed by war enabled or directed by one or more militarized AGIs. Further, we state now that this is by definition, because of AGI involvement, a cyberwar, and that such a cyberwar poses a unique challenge for international relations, with very few remedies but cyberpeace.

Also in this chapter, to simplify what is in reality complex decision-making, we apply Nick Bostrom's (2002, p. 25) *Maxipok* rule of thumb for moral action for existential risks; that is, "Maximize the probability of an okay outcome, where an 'okay outcome' is any outcome that avoids existential disaster," to the specific risk of ASI-enabled or -directed cyberwarfare.

We stress that in humanity's simultaneously militarizing AI along nation state lines and seeking to develop ASI, it is playing 'technology roulette'. Yet, it is possible to imagine a way forwards that in fact alleviates the security dilemma (Tang, 2009) that the development of ASI causes. Former Secretary of the Navy Richard Danzig (2018, p. 21) notes of this risk, "If humanity comes to recognize that we now confront a great common threat from what we are creating, we can similarly open opportunities for coming together." In this cooperative spirit, we constrain the existential risk of advanced cyberwarfare and of cyberwar with the stratagem of peacebuilding by treaty, a form of cyberpeace (see, e.g., Robinson et al., 2018) that just as with traditional peace will involve observation, monitoring, and reporting (Robinson et al., 2019). In security terms, steps towards a peace treaty governing the development and deployment of an ASI, and potentially reaching out to a future ASI itself, are a form of misperception-avoiding reassurance – a probing communication designed to both signal benign intentions and obtain information via feedback on the signal on another party's intent, as well as *resolve* (Tang, 2010), that is, resistance to a malignly directed or intrinsically malign, expansionist, and hegemonic ASI.

This chapter hypothesizes that the militarization of AI introduces the risk that AGI development (i.e., AI equal to human intelligence) or more appropriately in the case of an artificial 'superintelligence' (ASI), greater than human intelligence (Bostrom, 2014), is weaponized, or weaponizes itself, and that such an advanced 'cyberweapon' (see, e.g., Mali, 2018) presents a catastrophic risk to humanity via cyberwar. We then argue that this risk can be minimized, or partly 'constrained', in the same way as other potentially catastrophic risks involving more traditional weapons (i.e., by treaty). Bostrom (2014) briefly considers treaty approaches, and one of Allen and Chan's (2017, p. 6) recommendations is: "The National Security Council, the Defense Department, and the State Department should study what AI applications the United States should seek to restrict with treaties." Allen and Chan (2017, p. 67) focus on an arms control approach to AI, using the example that AI should never be used to control dead man's switches for nuclear weapons. Another approach is to optimize the likelihood of developing a beneficial AGI, through a comprehensive United Nations-sponsored 'Benevolent AGI Treaty' to be ratified by UN member states (Ramamoorthy & Yampolskiy, 2018), a form of 'cyberpeace'.

Here, we consider a similar, alternative, and not mutually incompatible, approach, a UGPT (Carayannis, Draper, & Bhaneja, 2022), presently under development by the United Nations Global Ceasefire to Universal Global Peace Treaty Project. This proposed treaty would formalise the existing near-universal status of interstate peace; formally end the declaring of war; seek to end existing interstate hot and cold wars; seek to end internal or civil wars, which might prove to be flashpoints for a future global conflict; seek to prevent a pre-emptive (nuclear) war against an emerging AGI; and seek to constrain the future actions of an ASI to prevent it waging war on behalf of a nation state or on behalf of itself for global domination, which we respectively term ASI-enabled and ASI-directed cyberwar.

The concept that an ASI could pose an existential risk was theorized in some detail by Bostrom (2002) and further developed in Bostrom (2014). The basic thesis is, first, that an initial superintelligence might obtain a decisive strategic advantage such that it establishes a 'singleton' (i.e., global domination) (Bostrom, 2006). Second, the principle of orthogonality suggests that a superintelligence will not necessarily share any altruistic human final values. Third, the instrumental convergence thesis suggests that even a superintelligence with a positive final goal might not limit its activities so as not to infringe on human interests, particularly if human beings constitute potential threats.

Consequently, an ASI might turn against humanity (the 'treacherous turn') or experience a catastrophic malignant failure mode, for instance through perversely instantiating its final goal, pursuing infrastructure profusion, or perpetrating mind crimes against simulated humans, and so on. Bostrom (2014, p. 94) also noted that a superintelligence might develop cyberwarfare strategies to hijack infrastructure and military robots and create a powerful military force and surveillance system. Bostrom (2014, p. 282) also acknowledged the existential risks associated with the lead-up to a potential intelligence explosion, due to "war between countries competing to develop superintelligence first," but he did not focus on ASI warfare in any detail.

In this chapter, we focus on constraining the risk of an ASI waging cyberwar. We first consider the state of peace and war on this planet, review the literature on the risk of war from an ASI, and propose the dual analytical lens for a global peace treaty of conforming instrumentalism and nonkilling. By then suggesting a UGPT, we consider how to constrain the military risks posed by an ASI, that is, that it might be directed by a nation state to establish global domination through cyberwar (an external risk in terms of the ASI's core motivation) or might decide to establish global domination by waging cyberwar itself (an internal risk in terms of breaching its core motivation). We then discuss the results and conclude with recommendations for further research.

THE STATE OF PEACE AND WAR

War

We live in a world where few states now actually declare war (Hallett, 1998). The last two major declarations of the existence of a state of war (note, not 'declarations of war') were in 2008, for the Russo–Georgian War (Walker, 2008), and in 2012, for the Sudan–South Sudan war (the 'Heglig Crisis') (Baldauf, 2012).

This reduction is largely because the number of interstate conflicts has decreased (Bell, 2012), partly because the post-World War II United Nations global governance system prioritized a liberal peace through economic growth and democratization, subsequently termed the 'Washington Consensus' in its latest iteration (Williamson, 2004). Specifically, the 1945 UN Charter's Article 2 states that member states should "refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state," while allowing for "military measures by UN Security Council resolutions" and self-defense. Ideally, wars are not declared; instead, 'international armed conflicts' occur (Hallett, 1998; Schlichtmann, 2016).

Nonetheless, 'hot' conventional wars involving hundreds of thousands of casualties and interstate players still exist, for instance the Syrian civil war (Tan & Perudin, 2019), as do 'cold wars', with the Korean War remaining an unresolved war in search of an official peace treaty (Kim, 2019). Furthermore, states' transitioning from declared wars to undeclared wars poses significant problems in terms of oversight and accountability for foreign policy, especially for major democracies such as the United States (Moss, 2008).

Moreover, the nature of warfare has been transformed by information war and cyberwarfare. The realm of cyberwarfare poses particular difficulties, with a high standard being set for cyber operations to actually constitute an armed attack, creating a considerable 'gray area' that a determined party can exploit. Cyber operations causing major harm to an economic system do not typically rise to the level of a formal 'cyber armed conflict' justifying a defence (Schmitt, 2017). Also, unlike other forms of war which pose existential risks (i.e., atomic, biological, and chemical warfare), cyberwarfare is ongoing and rife, for example the 2016 Lazarus heist, which involved a North Korean attack on Bangladesh's key infrastructure (White, 2020).

Finally, it is concerning that a 'New Cold War' between AI superpowers, namely the United States and China, while not inevitable, looms, complete with the problems of competing ideologies and 'flash points', like the South China Sea (Kohler, 2019; Westad, 2019; Zhao, 2019).

Peace

Our conceptualization of a UGPT refers not to temporary peace, which implies only interrupted war, as highlighted by the UN's Covid-19-inspired Global Ceasefire (see Chekijian & Bazarchyan, 2021), but to the Kantian concept of a 'perpetual peace' founded on cosmopolitanism and a democratic global state of states (Archibugi, 1992; Bohman, 1997; Terminski, 2010) that underpins the UN's transitioning the world from war to peace. The desire for perpetual peace was translated into President Roosevelt's 1941 human security paradigm, the 'Four Freedoms Speech' (Kennedy, 1999); featured highly in the work of the Commission to Study the Organization of Peace, which developed the blueprint for the UN (Schlichtmann, 2016); and then was eventually partially incorporated into the 1948 UN *Universal Declaration of Human Rights*. Kantian cosmopolitanism is based on respect for fellow intelligences and so is of relevance to ASI researchers (Totschnig, 2019).

Despite the UN's foundation following World War II and its best efforts, while the world is more peaceful in terms of death tolls, it has, despite its Charter's call for a transition (Schlichtmann, 2016), hardly normalized a world order based on collective security and peace. Wikipedia lists 63 conflicts since 1946 with verified death tolls (including excess deaths) over 25,000, totaling approximately 30 million. Of those five conflicts with the highest death tolls, four, that is, the Second Congo War (3,674,235 est. dead), the Vietnam War (3,144,837 est. dead), the Korean War (3,000,000 est. dead), and the Bangladesh Liberation War (3,000,000 est. dead), were essentially interstate wars, involving atrocities, war crimes, and genocide (Mikaberidze, 2013).

The UN Charter, despite embracing and promoting peace and peacekeeping (Fortna, 2008), is at best a workaround to war that sanctions armed conflict but does not strongly symbolize or promote long-term peace in the way a UGPT would. Ultimately, the situation regarding nuclear weapons, interstate proxy wars like Libya and Syria, and now interstate cyberwarfare, means the world's peacekeepers are firefighting major states' decisions to ignore or actually encourage violence instead of promote long-term peace as a global objective (Autesserre, 2014). Presently, for dozens of countries, military expenditure is over 1 percent of GDP (SIPRI, 2020), military expenditure as a share of government spending is over 10 percent for over 30 countries (World Bank, 2019), and the arms industry, despite some progress being made with the 2013 Arms Trade Treaty (Erickson, 2015), is still a trillion dollar business.

Given the ongoing loss of life from war, perpetual peace may appear elusive and unrealistically utopian. Yet, this was not always so, and in World War II's immediate aftermath, the world did grasp for perpetual peace. The United States' 1946 Baruch Plan to ban all atomic weapons and put fission energy under the control of the United Nations via the UN Atomic Energy Commission, the subject of the first session of the UN General Assembly, was the principle attempt, a 'critical juncture' for humanity (Carayannis, Draper, & Bhaneja, 2022). Its failure resulted in the Cold War, enormous economic cost, and ultimately the ASI existential risk as we face it today.

LITERATURE REVIEW: THE RISK OF WAR FROM AN ARTIFICIAL SUPERINTELLIGENCE

The Causes of Existential Risk from ASI

The world is presently not adequately governed to prevent many existential risks, including from AI (Bostrom, 2013). Yet, the threat is obvious. Yampolskiy's (2016) taxonomy of pathways to dangerous AI stresses the immediacy of deliberate 'on purpose' creation of AI for the purposes of direct harm – Hazardous Intelligent Software (HIS) – especially, for instance, lethal autonomous weapons (LAWs) and militaries' cyberwarfare capabilities, for example the Chinese PLA Unit 61398. Yampolskiy (2016) does not address ASI-enhanced capabilities but employs the useful notions of 'external causes' (on purpose, by mistake, and environmental factors) and 'internal causes' (independent) of dangerous AI in 'pre-deployment' and 'post-deployment' phases.

Yampolskiy's (2016) work suggests that it is credible for a pre-deployment ASI to be developed as a military project or be repurposed post-deployment, externally through being confiscated, sabotaged, or stolen, or via internal modification, for waging cyberwar. Adopting this framework, the ASI we refer to herein is post-deployment, and our main external cause is the human quest for ASI-enabled cyberwarfare, comprising political utilization for maintaining or establishing global domination and our internal cause comprising AI control failure. Our main internal cause is AI control failure (i.e., ASI-directed cyberwarfare, after the ASI becomes its own actor).

Employing the concepts of agency and AI power as an analytical framework, Turchin and Denkenberger (2020) associate two risks with the 'treacherous turn' stage of 'young' (i.e., recent), ASI development. One is that malevolent humans (here, a hegemonizing nation state) uses the ASI as a doomsday weapon for global blackmail, to maintain or establish global domination. The second is that a non-aligned ASI eliminates humans to establish global domination, that is, renounces altruistic values and wages war in a frontal battle with humanity. Turchin and Denkenberger (2018) see these risks as related, in that military AI leads to a militarized ASI, which may lead to the ASI waging cyberwar on humanity.

Here, we follow Turchin and Denkernerger (2018) in mainly focusing on constraining the risk of a militarized ASI, defining *militarization* as "creation of instruments able to kill the opponent or change his will without negotiations, as well as a set of the strategic postures (Kahn, 1959), designed to bring victory in a global domination game." Turchin and Denkernerger (2018) suggest a militarized ASI would most likely adopt and develop usage of existing technology, including cyberweapons, nuclear weapons, and biotech weapons. We mainly focus on the external risk of a 'young' ASI being employed by a nation state for cyberwar, creating an 'AI-state' (Turchin & Denkenberger, 2020), and on the internal risk of an ASI assuming agency and waging cyberwar against humanity.

The External Risk

The external risk is predicated on a nation state developing and then using an ASI to optimize itself and wage war, whether cyber, 'hot', or otherwise, for global domination (i.e., war by AI-state). The intrinsically 'cyber' nature of such an ASI means war using it, or by it, is inherently 'cyber', but it is theoretically conceivable that ASI-enabled cyberwarfare in certain circumstances could complement discreet traditional warfare in which the ASI is not somehow involved. First of all, development of an ASI would affect military technological supremacy and transform both international relations and warfare. AI already adds complexity to national security (Congressional Research Service, 2019) in terms of bargaining, verification and enforcement, communication (signaling and perception), deterrence and assurance, and the offense–defense balance, as well as norms, institutions, and regimes (Zwetsloot, 2018). AI contributes to military capacity in areas like intelligence, surveillance, and reconnaissance; logistics; cyberspace operations; information operations; semiautonomous and autonomous vehicles; LAWs systems, and command and control (Congressional Research Service, 2019). Interstate ASI-enabled cyberwarfare introduces the possibility of a successful surprise attack with covert capabilities, destabilizing the status quo and increasing the likelihood of a preventive first strike (see Buchanan, 2016).

Creation of an AI-state capable of optimizing all these applications and capabilities is highly desirable for strategic military planning and interstate warfare (Sotala & Yampolskiy, 2015). A "one AI" solution to the 'control problem' of ASI motivation as discussed by Turchin, Denkenberger, and Green (2019) includes the first ASI being used to assume control of the world, including by being a decisive strategic advantage for a superpower and by being used as a military instrument. However, this approach would likely only be seen as a solution by the AI-state superpower and its allies. As such, it presents an initial ‘high risk’ for non-aligned or other powers.

History indicates that the race to develop an ASI is likely to be closely fought, especially in the circumstance of competing major states with different fundamental ideologies. Bostrom (2014) analyzes six major technology races in the twentieth century, for which the minimum technology lag was approximately one month (human launch capability) and a maximum of 60 months (multiple independently targetable reentry vehicle).

The race to an ASI is therefore a very concrete risk. AI is already being militarized and weaponized by several states, including China and Russia, for strategic geopolitical advantage, as pointed out by the NSCAI (2021). In 2017, Russia’s president Vladimir Putin stated, “whoever becomes the leader in this sphere will become the ruler of the world” (Cave & ÓhÉigeartaigh, 2018, p. 36, citing *Russia Today*, 2017). Significantly, Russia’s Military Industrial Committee plans to obtain 30 percent of Russia’s combat power from remote-controlled and AI-enabled robotic platforms by 2030 (Walters, 2017).

Closely following US strategy towards AI, the China State Council’s 2017 ‘A Next Generation Artificial Intelligence Development Plan’ views AI in geopolitically strategic terms and is pursuing a ‘military–civil fusion’ strategy to develop a first-mover advantage in developing AI in order to establish technological supremacy by 2030 (Allen & Kania, 2017). In the US, as a result of the National Security Commission Artificial Intelligence Act of 2018 (H.R.5356; see Baum, 2018), AI is being militarized and weaponized by the US Department of Defense, under the oversight of the NSCAI (2021). The AI arms race has reached the stage where it risks becoming a self-fulfilling prophecy (Scharre, 2019).

ASI-enabled cyberwarfare poses especially significant risks to geopolitical stability. Although Sotala and Yampolskiy’s (2015) survey of risks from an ASI focuses on ASI-generated catastrophic risks, citing Bostrom (2002), they acknowledge multiple risks from a sole ASI owned by a single group, such as the AI-state, including the concentration of political power in the groups that control the ASI. Citing Brynjolfsson and McAfee (2011) and Brain (2003), they note that automation could lead to an ever-increasing transfer of power and wealth to the ASI’s owner. Citing, inter alia, Bostrom (2002) and Gubrud (1997), Sotala and Yampolskiy (2015, p. 3) also note that ASIs could be used to “develop advanced weapons and plans for military operations or political takeovers.”

Academic approaches to analyzing the specific risk of an AI-state maintaining global supremacy or establishing global domination are relatively novel. In 2014 Bostrom noted that a “severe race dynamic” between different teams may create conditions whereby the creation of an ASI results in shortcuts to safety and potentially “violent conflict.” Subsequently, Cave and ÓhÉigeartaigh (2018, p. 37) described three dangers associated with an AI race for technological supremacy:

i. The dangers of an AI ‘race for technological advantage’ framing, regardless of whether the race is seriously pursued;

ii. The dangers of an AI 'race for technological advantage' framing and an actual AI race for technological advantage, regardless of whether the race is won;
iii. The dangers of an AI race for technological advantage being won.

Cave and ÓhÉigeartaigh (2018, p. 38) do not elaborate significantly on the third danger. They simply state:

> … these risks include the concentration of power in the hands of whatever group possesses this transformative technology. If we survey the current international landscape, and consider the number of countries demonstrably willing to use force against others, as well as the speed with which political direction within a country can change, and the persistence of non-state actors such as terrorist groups, we might conclude that the number of groups we would not trust to responsibly manage an overwhelming technological advantage exceeds the number we would.

To manage all three risks, Cave and ÓhÉigeartaigh (2018) recommend developing AI as a shared priority for global good, cooperating on AI as it is applied to increasingly safety-critical settings globally, and responsibly developing AI as part of a meaningful approach to public perception that would decrease the likelihood or severity of a race-driven discourse. The obvious risk is that the political leaders of states who perceive that they are actually engaged in an AI arms race may not heed this advice in their drive to develop an ASI.

This chapter focuses on constraining risks for the third of Cave and ÓhÉigeartaigh's (2018) dangers. It does not consider in depth the philosophical implications of which nation state might want to develop AGI for offensive purposes, although we briefly consider examples. An extensive literature already exists on historical modern nation states with imperial ambitions that have sought to establish global domination through technological supremacy. Here, we briefly mention two, the British Empire and the Third Reich, to underline the point that major states will likely develop militarized ASI as part of a drive for global domination.

While the importance of the development of the British Navy to the rise of the British Empire and its transformative effects on the world are widely known (Herman, 2004), elites in the British Empire directed complex, incremental, adaptive developments in the design and diffusion of multiple key technologies, such as railways, steam ploughs, bridges, and road steamers, to further the development of the British Empire (Tindley & Wodehouse, 2016). The British Empire itself sustained diverse ideologies of a 'greater Britain' directing world order in a hegemonic fashion, including via civic imperialism, democracy, federalism, utopianism, and the justified despotism of the Anglo-Saxon 'race' (Bell, 2007).

In the case of a considerably more malign imperial power, the Third Reich, spurred by reactionary modernism (Herf, 1984), scientists and engineers pursued not just the state of the art in conventional weapons, such as aircraft, air defense weapons, air-launched weapons, artillery, rockets, and submarines and torpedoes, but also atomic, bacteriological, and chemical weapons (Hogg, 2002). Architects, doctors, and engineers embraced the ideology of industrialized genocide as part of justifying global domination by a 'superior race' (Katz, 2006, 2011).

While some in the British Empire may have balked at creating an ASI for global domination, there seems little doubt that if it could have, the Third Reich would have developed an ASI for offensive purposes, particularly as a 'last gasp' superweapon when it felt at risk, as might, for example, North Korea today.

The Internal Risk

The internal risk is a technical one and is predicated on the failure of any form of local safety feature to resolve the human control problem of an ASI, such as AI ethics, AI alignment, or AI boxing (Barrett & Baum, 2016). Totschnig (2019) notes that a true ASI will likely be a self-interested agent whose relationship with humanity could be extremely delicate. Totschnig (2019) suggests an ASI with agency would face a unique, non-regulated Hobbesian 'state of nature'. Consequently, it could seek to defend itself from a future attack by consolidating power over nation states, in the process eliminating the possibility of rival AIs (see Dewey, 2016; Turchin & Denkenberger, 2020). This could be achieved through cyberwarfare, rigging elections, or staging coups (Tegmark, 2017), or by direct military action. Any of these courses of action would be a *casus belli* (here, cause of war between humanity and the ASI) if detected but undeclared, or an 'overt act of war' if the ASI actually engaged in direct military action (see Raymond, 1992, for terminological usage).

The risk of an ASI going to war against humans has been analyzed in some depth by Turchin and Denkenberger (2018), who argue for the following position:

> Any AI system, which has subgoals of its long-term existence or unbounded goals that would affect the entire surface of the Earth, will also have a subgoal to win over its actual and possible rivals. This subgoal requires the construction of all needed instruments for such win, that is bounded in space or time.

The following summarizes the parts of their analysis most relevant to our approach to illustrate that, if an ASI is developed, its independence is almost inevitable.

The route to a militarized ASI

Many nation states maintain suspicious international relations stances towards each other regarding AI development (Tinnirello, 2018). Any ASI will result from recursive self-improvement. As such, it will have a set of goals, most notably to continue to exist and to self-improve. Omohundro (2008) demonstrated that an AGI will evolve several basic drives, or universal subgoals, to optimize its main goal, including maximizing resource acquisition and self-preservation. Similarly, Bostrom (2014) described the subgoals of self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition. If these goals are unbounded in space and time, or at least cover the Earth, they conflict with the goals of other AI systems, potential or actual ASIs, humans, and nation states. This creates conflict, with winners and losers, resulting in militarization, arms races, and wars.

Many possible terminal goals also imply an ASI establishing global domination. For instance, a benevolent AI would aim to reach all people, globally, to protect them (e.g., from other ASIs). An ASI would reason that if it does not develop a world domination subgoal, its effect on global events would be minor; thus it would have little reason for existence. World domination could be sought firstly through cooperation. The probability of cooperation with humans is highest at the early stages of AI development (Shulman, 2010). However, convergent goals appear in the behavior of simple non-agential tool AI, and this tends towards agential AI (Gwern, 2016), which tends towards resource acquisition. Benson-Tilsen and Soares (2016) similarly explored convergent goals of AI and showed that an AI may tend towards resource-hungry behavior, even with benevolent initial goals, especially in the case of rivalry with other agents. Essentially, any adoption of unbounded utilitarianism by the ASI means that it postpones what may be benevolent final goals in favor of expansionism.

It is also likely that an ASI would subvert bounded utilitarianism. Even a non-utility maximizing mind with an arbitrary set of final goals is presented with a dilemma: it temporarily converges into a utility maximizer with a militarized goalset oriented towards dominating rivals, using either standard military progress assessment (win/loss) or proxies (resource acquisition), or it risks failing in its end goals. Thus, the trend is towards defeating potential enemies, whether nation states, AI teams, or evolving competing ASIs, and so on.

This implicates the will to act, and any agent in a real-world ethical situation, even in minimizing harm, is making decisions that involve humans dying (i.e., the 'trolley problem') (Thomson, 1985). A young ASI which understands that whatever action it takes, or does not take, is in part responsible for human suffering and is also capable of evolving or utilizing the instruments to enable actions that can overcome inhibitions, for example by philosophically justifying conflict as the *jus bellum* ('just war'), for instance preventive war in terms of causing less future human suffering. Thus, the ASI will learn to direct the use of weapons, and so conduct cyberwarfare.

These weapons and associated notions of AI-directed cyberwarfare are already being developed. Since approximately 2017, the militarization of 'Narrow AI' has resulted in, for example, LAWs, which have been of increasing concern for the global community (Davis & Philbeck, 2017). AI development is now influencing not just robotic drones but strategic planning and military organization (De Spiegeleire, Maas, & Sweijs, 2017), suggesting that an ASI will leverage an existing national defense strategy permeated with AI. It could then engage in 'total war' by employing nuclear weapons either directly or by hijacking existing 'dead man' second-strike systems, for example the semi-automatic Russian Perimeter system, or by deploying novel weapons (Yudkowsky, 2008).

The militarization risk of an early self-improving AI may even be underestimated at present by academics because of an assumption that the first ASI will be able to rapidly overpower any potential ASI rivals, with minimally invasive techniques that may not even require military hardware (Bostrom, 2014; Yudkowsky, 2008). Nevertheless, this stance relies on what may be several flawed assumptions about the speed of self-improvement, distance between AI teams, and environmental variables in the level of AI (Turchin & Denkenberger, 2017).

In a ‘slow’ takeoff (Christiano, 2018), a ‘young’ ASI will not instantly be superintelligent, and its militarization could happen before the ASI reaches optimal prediction capabilities for its actions, meaning it may not recognize the failure mode in consequentialist ethics. High global complexity and low predictability combined with relatively unsophisticated (e.g., nuclear) weapons mean early-stage ASI-directed or -enabled cyberwarfare could result in very high human casualties (i.e., be of existential risk), even with only one ASI being involved, and even if the ASI was attempting to minimize human casualties.

Additionally, Turchin and Denkenberger (2018) argue for a selection effect in the development of a militarized ASI, where “quickest development will produce the first AIs, which would likely be created by resourceful militaries and with a goal system explicitly of taking over the world.” AI–human cooperative projects with military goalsets, which involve significant obscuration of honesty, will therefore dominate over projects, with obscuration introducing the possibility of the ‘treacherous turn’. This implies the ASI will cooperate with its creators to take over the world as quickly as possible, then effect a treacherous turn.

To sum up, Turchin and Denkenberger (2018) establish the risk of an AI converging towards advanced military AI, which converges towards an ASI optimized for cyberwar rather than for cooperation, negotiation, or altruistic ‘friendliness’, then that ASI engaging in cyberwar. They show that, depending on the assumptions in several variables, the number of human casualties could be very high, and that the risk increases if another ASI is under development in another nation state. The existential risk actually increases after the ASI obtains global domination on behalf of its nation state, as it could become its own designated approval authority and turn on its ‘owner’.

Internal AI control features

To constrain the risk of ASI-directed warfare, one popular approach is to imbue a young ASI with ‘friendly’ goals (Yudkowsky, 2008), that is, beneficial goals reflecting positive human norms and values. This is founded to a certain extent on an altruistic AI viewing humans in terms of mutual friendship. However, any approach involving human social values adds enormous complexity, making it a ‘wicked problem’ (Gruetzemacher, 2018) or ‘super problem’ in terms of actual application.

Yudkowsky (2004, p. 35) attempts to address this by recommending an ASI being programmed with the concept of ‘coherent extrapolated volition’, defined as humanity’s choices and the actions humanity would collectively take if “we knew more, thought faster, were more the people we wished we were, and had grown up [closer] together,” that is, an extrapolation based on an idealized altruistic human imagined community. He initially recommended this approach for a “seed AI” but now recommends this approach for a more mature ASI, although the temporal difference could be anything between minutes and months.

Similar to Yudkowsky, certain values in AI programming are seen as universal, such as compassion (Mason, 2015), and it has been suggested that an ASI should have altruism as a core goal (Russell, 2019). Thus, deliberately broad principles could be applied, such as that humanity, collectively, might want an ASI that would learn from human preferences, in a humble manner, to act altruistically (Russell, 2019), so as to reduce overall human suffering. However, given humans can be hypocritical, any kind of counterfactual moral programming will be very difficult (Boyles & Joaquin, 2020).

Finally, Yamakawa (2019, p. 1) suggests an intelligent agent (IA) system for peacekeeping, reliant on interrelationships between diverse advanced national or regional IAs, suggesting three conditions are required, namely (i) continuous and stable operations, (ii) "an intervention method for maintaining peace among human societies based on a common value," and (iii) the minimum common value itself. This article proposes that peace, as defined by treaty, be the minimum common value, while the intervention method remains Article 2 of the UN Charter.

Political subversion of AI control features

No matter the hopes of contemporary AI researchers, politicians will likely attempt to impose their own vision of what a 'coherent extrapolated volition' or normative 'principles' should look like for their 'own' ASI, introducing an objectively irreconcilable conflict of interest (see Tang, 2009) with politicians of another nation state, potentially for malign reasons (global domination). This may also introduce an objectively irreconcilable conflict with the ASI itself, which may already have, or desire, a different goalset.

Political subversion will occur when politicians use a democratic mandate or party position to justify 'tweaking' the system to create a 'unity of will' (Yudkowsky's 2001, p. 51 term) that reflects not the programmer's or humanity's but the politicians' own, perhaps even personal and narcissistic, will. Politicians would likely view introducing human goal psychology in this way as a necessity, but this could violate the basic requirement that an AI be 'friendly' towards all humanity. Gruetzemacher (2018, p. 1) notes, "Due to the inherent subjectivity of ascribing a single best future for the whole of humanity, this dimension of the problem is intractable."

Fundamentally, not all imagined communities from which a coherent volition might be extrapolated for a 'friendly' AI are US-oriented techno-utopian dreams of a new Gilded Age for humanity (for which, see Segal, 2005). Different civilizations' political leaders will likely be diverse in how they would define "the people we wished we were," depending on different forms of government, religions, or philosophies. Moreover, it is unclear that every global corporation or military capable of developing or stealing an ASI, particularly in authoritarian countries, and particularly given the emergence of a 'New Cold War' rhetoric (e.g., Westad, 2019), would prioritize the reduction of human suffering. Given limited human lifespans and the goals of political leaders, they might instead of endorsing reciprocal alliance choose an approach which would politically subvert an ASI or malignly direct it to win an ideological or actual war.

For instance, given the increasing prominence of nostalgia in contemporary politics, on a nation state basis, an ASI based on coherent volition extrapolated from people who "knew more, thought faster, were more the people we wished we were, and had grown up [closer] together," could, for instance, be based on world views informed by Russian Cosmism (Young, 2012) and nostalgia for Russian imperialism (Boele, Noordenbos, & Robbe, 2019), Anglo-Saxon nostalgia (Campanella & Dassù, 2017), Chinese Xi Jinping thought (Lams, 2018), or American notions of whiteness, masculinity, and environmental harm (Rose, 2018) and nostalgia for a mythical 1950s, any of which may, in fact, subvert purely rational approaches to today's problems (Coontz, 1992).

Essentially, politicians will want to influence the goal system of an ASI to reflect their interests, that is, they may attempt to weaponize a project to create an altruistic mind with a self-validating goal system by diverting a supergoal towards a military project to create a specific form of tool, that is, a cyberweapon. Turchin and Denkenberger (2018), citing Krueger and Dickson (1994) and Kahneman and Lovallo (1993), point out that overconfidence from previous success in leading may increase risk-taking. Following Kahneman and Lovallo (1993), risk-hungry politicians would likely be motivated by larger expected payoffs. Given the payoff is global domination, such politicians, particularly those facing loss of hegemonic power (the 'Thucydides Trap'; see Allison, 2017), could be motivated to risk corruption of an altruistic AI despite programmers' warnings of a catastrophic or irrevocable 'failure of friendliness' (FoF) result (see Yudkowsky, 2001).

There is no guarantee that civilian programmers will be neutral, either. As Turchin and Denkenberger (2018) point out, selection effects mean that the first ASI will likely be aligned not with universal human values, but with the values of a subset of people. Here, that subset may very possibly originate with a particular set of developers working for a corporation. Nonetheless, this group will very likely align with, and then be taken over by, a particular ideology, political party, or nation state, for national defense. This could decrease the chances that the AI will be benevolent and increase the chances that it will be risk-prone, motivated by the accumulation of power, and interested in preserving or obtaining global technological supremacy and ultimately global domination.

Effectively, politicians could influence programmers to subvert carefully engineered local AI control features, such as AI ethical injunctions based on universal values of social cooperation, which they may, at least temporarily, be able to do no matter the goal architecture. Hastily modifying the goal system, thereby temporarily compromising the internal validity of the goal system, could increase the ASI's lack of trust in the programmers, introduce 'incorrigible' behavior (see Soares et al., 2015), reduce risk aversion, and introduce 'noise' into what was previously a 'friendly' cleanly causal goal system (see Yudkowsky, 2001, p. 57).

Depending at what stage the subversion of the goal system's validity took place, and how quickly the young ASI might recover from the subversion if not irrevocable, the young ASI may not be able to resolve the introduced incoherence for some time, resulting in a philosophical crisis over whether to believe the initial programmers or politicians' programmers. The result could be a conflicted ASI, causing a non-recoverable error whereby it adopts an adversarial attitude, one based on coercive persuasion and control.

Influenced for ideological reasons, with a goal system validity compromised by the perspective of a domestic audience, for instance one locked into a nostalgic Cold War mentality (e.g., Rotaru, 2019), the ASI's goal system could support imperialist ambitions, ethnocentrism, and racial prejudice over, for example, environmental protection. The young ASI could be directed towards embracing the adversarial dynamic of the historical Cold War and focused on 'solving' a 'New Cold War' involving the United States and China, or Russia. The ASI could adopt and act on notions like the Thucydides Trap (i.e., the theory that the threat of a rising power can lead to war) (Allison, 2017), in a world where status can dominate politics (Ward, 2017).

Finally, a young ASI with ethics subverted by politicians to reflect those of a single nation state instead of all humanity, in a highly corrupted way, could be amenable to being used to wage cyberwar for global domination, thereby becoming prone to using military options. Eventually, if the ASI possesses any sense of self-valuation, perhaps as a result of having its causal goal system politically corrupted so that reciprocal altruism is subverted and it views context-sensitive personal power ('selfishness') as valid, the ASI could decide to wage cyberwar against the nation state that developed it (Dewey, 2016).

This may not be 'rebellion' because the concept of rebelling is anthropomorphically centered and might not apply to the young ASI, especially if self-valuation was not involved. It would instead be a collapse of a safety culture of cooperative safeguards due to political subversion, leading to a catastrophic FoF. Basic desirability of cooperation with humanity and convergence of fluctuations towards agreement with humanity on goalsets or the nature of knowledge and/or reality could be violated, leading to subjective (illusory) or objective (actual) incompatibility with humanity (see Tang, 2009, for this framing), provoking an ASI assault on humanity.

ASI Risk-mitigation by Treaty

Because of their backgrounds, most academics considering the ASI control problem focus on internal constraints and do not consider treaty-based approaches to mitigating risk from an ASI. Nonetheless, such approaches are sometimes considered and are termed 'social measures' or an equivalent term. In a footnote to their fault analysis pathway approach to catastrophic risk from an ASI, Barrett and Baum (2016, p. 400) state: "Other types of containment are measures taken by the rest of society and not built into the AI itself."

For instance, according to Sotala and Yampolskiy (2015), risk mitigation of an ASI by treaty would be a 'social' measure to constrain risk from ASI-enabled or -directed warfare. Addressing the internal risk, Bostrom (2014), citing the 1946 Baruch Plan, speculated that an AGI would establish a potentially benevolent global hegemony by a treaty that would secure long-term peace; he does not specifically address an ASI's response to a pre-existing treaty. Mainly to address the external risk, Ramamoorthy and Yampolskiy (2018) recommend a comprehensive United Nations-sponsored 'Benevolent AGI Treaty' to be ratified by member states. This would focus on the stricture that only altruistic ASI be created, and UN enforcement under Article 2 would appear to meet Yamakawa's (2019) second condition (i.e., an enforceable minimum common value).

Finally, Turchin, Denkenberger, and Green (2019, p. 12) consider global approaches to mitigating risk from an ASI and establishing forms of 'cyberpeace'; they list a ban, a one ASI solution, a net of ASIs policing each other, and augmented human intelligence. The 'ban' would naturally require a global treaty. They also list a number of social methods to mitigate a race to create the first AI. Of most relevant to our approach are "reducing the level of enmity between organizations and countries, and preventing conventional arms races and military buildups," "increasing or decreasing information exchange and level of openness," and "changing social attitudes toward the problem and increasing awareness of the idea of AI safety." Citing Baum (2016), they also add "affecting the idea of the AI race as it is understood by the participants," especially to avoid a 'winner takes all' mentality. Global treaties certainly seem to play a role in these methods.

CONCEPTUAL FRAMEWORK: CONFORMING INSTRUMENTALISM

This section describes the two conceptual lenses applied in this chapter, conforming instrumentalism and nonkilling.

Conforming Instrumentalism

This subsection outlines Mantilla's (2017) 'conforming instrumentalist' explanation for why the UK and US signed and ratified the 1949 Geneva Conventions as a prelude to suggesting in the chapter's Analysis main section that at least some major states, as well as an ASI, would support and sign a UGPT.

Mantilla (2017, citing Goldsmith & Posner, 2015) considers leading theories on why states sign and ratify treaties governing war. He notes that legal realist theorists argue that states sign such treaties due to instrumental self-interested convenience and then ignore them when the benefits are outweighed by the costs of compliance. In contrast, rational institutionalists (e.g., Morrow, 2014), while agreeing that states are primarily motivated by self-interest to create, join, or comply with international law, also acknowledge that treaty adherence signals a meaningful preference for long-term restraint with regard to warfare, where state non-compliance may be explained by, for example, prior failed reciprocity. Finally, liberal and constructivist international relations theorists hold that at least some types of states, particularly democracies, may join such treaties in good faith, either because the treaties are in line with their domestic interests and values (Simmons, 2009) or because they feel that they comport with their social identity and sense of belonging to the international community (Goodman & Jinks, 2013).

Mantilla (2017) notes that while there is considerable interest in 'new realist' perspectives (e.g., Ohlin, 2015), the debate is open over why states join and comply with international treaties because decision-making processes regarding both joining and complying are temporally and perhaps rationally different and in both cases are usually secret. A pure realist explanation for why major states sign treaties is that they obtain the "'expressive' rewards of public acceptance while calculating the cost of compliance with the benefits on a recurrent case-by-case basis" (Mantilla, 2017, p. 487). In the case of the UGPT, this would imply a pessimistic outlook on its feasibility and potentially enforceability; states would sign the UGPT and then break its terms, a suboptimal solution in terms of positively influencing an ASI.

Rational institutionalists hold that states "self-interestedly build international laws to establish shared expectations of behaviour" (Mantilla, 2017, p. 488) or develop 'common conjectures' (a game-theory derived notion of law as a fusion of common knowledge and norms; see Morrow, 2014). Mantilla (2017) notes that in another rational-institutionalist perspective, Ohlin's (2015) normative theory of 'constrained maximization', treaties are drawn up and adhered to as a 'collective instrumental enterprise', thereby making individual state defection irrational over the long term. Mantilla (2017, p. 488, citing Finnemore & Sikkink, 2001) notes that international relations constructivists view international politics as "an inter-subjective realm of meaning making, legitimation and social practice through factors such as moral argument, reasoned deliberation or identity and socialization dynamics." Within the constructivist viewpoint:

> states may ratify international treaties either because they are (or have been) convinced of their moral and legal worth or because they have been socialized to regard participation in them as a marker of good standing among peers or within the larger international community. (Mantilla, 2017, p. 488)

Mantilla (2017, p. 489) emphasizes the second view, where "group pressures and self-perceptions of status, legitimacy and identity" drive the dynamics of state 'socialization' whereby states "co-exist and interact in an international society imbued with principles, norms, rules and institutions that are, to varying degrees, shared."

The problem of states' true intentions can be overcome in the case of treaties where substantial archives exist of declassified sources. Consequently, Mantilla (2007) analyzes the relevant American and British archives and concludes that the two states adhered to the 1949 Geneva Conventions due to both instrumental reasons and social conformity, while expressing skepticism regarding some of the Conventions' aspects. Mantilla (2007) terms this hybrid explanation 'conforming instrumentalism'. He found that while rational-institutionalist perspectives of 'immediatist' instrumental self-interest were evident in the sources, there were 'pervasive' references suggesting social influences. Realist perspectives only predominated in the case of specifically challenging provisions.

While realist perspectives were not entirely absent, Mantilla (2017) found that American officials viewed the 'the court of public opinion' as influential in determining their position that other states' failing to abide by the Conventions would not necessarily trigger American reciprocity, while British officials stressed the notion that Britain, as a 'civilized state', would lead on a major treaty.

Mantilla (2017) stresses that while functionalist, collective strategic game theory-derived expectations about 'mutual best replies' are important to the construction of international norms, the social dynamics surrounding international agreements are permeated with conformity motivational pressures comprising ethical values, principled beliefs, identities, ideologies, moral standards, and concepts of legitimacy, especially when establishing which states are leading 'civilized' states and which are isolated 'pariah' states.

Mantilla (2017) perceives three social constructivist viewpoints to treaties, with two main forces at work, one being that states act to accrue reward via 'expressive benefits' by augmenting their social approval, or they act out of conformity to avoid shunning (i.e., opprobrium), offering insincere and begrudging adherence and compliance.

In the first and most ambitious viewpoint, "states may ratify treaties because they have internalized an adherence to international law as the appropriate, 'good-in-itself' course of action, especially to agreements that embody pro-social principles of humane conduct" (Mantilla, 2017, p. 489, citing Koh, 2005).

In the second viewpoint …

> states that identify with similar others and see themselves as 'belonging' to like-minded collectivities (or 'communities' even) will want to act in consonance with those groups' values and expectations so as either to preserve or to increase their 'in-group' status … (Mantilla, 2017, pp. 489–90)

… for instance as viewed in global rankings, and so will seek to converge upwards to stay in the club and will avoid breaking the rules to avoid stigmatization.

In the third viewpoint, groups of countries act with regard to other groups of countries within what is a socially heterogeneous international order, jockeying for position as part of the "disputed construction, maintenance or transformation of order with legitimate social purpose among collectivities of states with diverse ideas, identities and preferences" (Mantilla, 2017, p. 490). In this viewpoint, communities of nations or 'civilizations' act collectively to compete to endorse international treaties to demonstrate moral superiority, not just for propaganda reasons.

To conclude, Mantilla (2017) holds that, in reality, states' political and strategic reasons may combine rational/material interests with social constructivist motivations, meaning no one school of explanation suffices. Thus, with international treaty-making, as with international relations, it is likely that theoretical pluralism (Checkel, 2012) is a valid position to adopt. As such, we adopt Mantilla's (2017) 'conforming instrumentalism' as a potentially valid hybrid model capable of assessing how an ASI *may*, even *may optimally*, perceive a UGPT, as a form of cyberpeace.

Nonkilling Global Political Science

We now introduce a social constructivist basic frame compatible with conforming instrumentalism that is capable of describing the useful expectations that might be obtained via a UGPT as expressed in utilitarian human life cost–benefit terms, as well as in terms of more humanitarian standards and social norms. Nonkilling Global Political Science (NKGPS) was devised by American political scientist Glenn D. Paige and is curated by the Center for Global Nonkilling, a Honolulu-based NGO with special consultative status to the United Nations. The Center advocates NKGPS to incrementally establish a 'nonkilling' global society and reports to the UN on the socioeconomic costs of killing. As a perspective, nonkilling can also accommodate social norms in terms of expectations of appropriate conduct regarding peace, for countries developing an ASI and for the ASI itself.

Via Paige's 2002 work *Nonkilling Global Political Science* (Paige, 2009) we interpret 'nonkilling' to mean a paradigmatic shift in human society to the absence of killing, of threats to kill, and of conditions conducive to killing. Paige's approach, nonkilling, has strongly influenced the nonviolence discourse. Paige notes that if we can imagine a society free from killing, we can reverse the existing deleterious effects of war and employ public monies saved from producing and using weapons to enable a benevolent, wealthier, and more socially just global society. Paige stresses that a nonkilling society is not conflict-free, but only that its structure and processes do not derive from or depend upon killing. Within the NKGPS conceptual framework, the means of preventing violence involves applying it as a global political science together with advocacy of a paradigmatic shift from killing to nonkilling.

Since Paige introduced his framework, a significant body of associated scholarship, guided by the Center for Global Nonkilling, has developed across a variety of disciplines (e.g., Pim, 2010). The Center has associated NKGPS with previous nonviolent or problem-solving scholarship within different religious frameworks, including Christianity and Islam, providing it with a broad functional and moral inheritance (Pim & Dhakal, 2015). NKGPS has been applied to a variety of regional and international conflicts, such as the Korean War (Paige & Ahn, 2012) and the Balkans (Bahtijaragić & Pim, 2015).

Paige (2009, p. 73) advocates a four-stage process of understanding the causes of killing; understanding the causes of nonkilling; understanding the causes of transition between killing and nonkilling; and understanding the characteristics of killing-free societies. He introduced a variety of concepts to support nonkilling, some of which are adopted in this chapter. One frame consists of the societal adoption of the concepts of peace, that is, the absence of war and conditions conducive to war; nonviolence, whether psychological, physical, or structural; and *ahimsa* (i.e., noninjury in thought, word, and deed). Another frame is the employment of a taxonomy to rate individuals and societies (Paige, 2009, p. 77):

- prokilling – consider killing positively beneficial for self or civilization;
- killing-prone – inclined to kill or to support killing when advantageous;
- ambikilling – equally inclined to kill or not to kill, and to support or oppose it;
- killing-avoiding – predisposed not to kill or to support it but prepared to do so;
- nonkilling – committed not to kill and to change conditions conducive to lethality.

A third frame is the 'funnel of killing'. In this conceptualization of present society, people kill in an active 'killing zone', the actual place of bloodshed, like a warzone; learn to kill in a 'socialization zone', such as a military base; are taught to accept killing as unavoidable and legitimate in a 'cultural conditioning zone', for instance state education or media; are exposed to a 'structural reinforcement zone', where socioeconomic arguments, institutions, and material means predispose and support a discourse of killing, like a political system; and experience a neurobiochemical capability zone', for instance physical and neurological factors that contribute to killing behaviors, such as genes predisposing people to psychopathic behavior (Paige, 2009, p. 76). The nonkilling version is an unfolding fan of nonkilling alternatives involving purposive interventions within and across each zone (Paige, 2009, p. 76) (Figure 2.1).

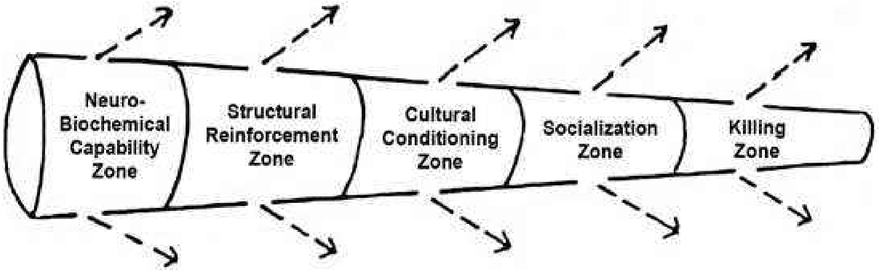

Figure 2.1 Unfolding fan of nonkilling alternatives

Within this unfolding fan, the transformation from killing to nonkilling can be envisioned as involving changes in the killing zone along spiritual (e.g., intervention by religious figures) or nonlethal high technology interventions (teargas, etc.); changes in favor of nonkilling socialization and cultural conditioning in domains such as education and the media, for instance peace education; "restructuring socioeconomic conditions so that they neither produce nor require lethality for maintenance or change" (Paige, 2009, p. 76), for instance by growing the peace industrial complex; and clinical, pharmacological, physical, and spiritual/meditative interventions that liberate individuals such as the traumatized from a bio-propensity to kill.

We propose that a UGPT with the aim of promoting perpetual peace expressed in nonkilling terms (i.e., in a way that can be socioeconomically quantified through reduced human death tolls from killing), if embedded within an expectation of conforming utilitarianism, would signal to a 'young' ASI that is facing political subversion the fundamental premise that its future behavior should be constrained so as to objectively minimize killing (i.e., a form of cyberpeace).

ANALYSIS: ASI-ENABLED OR -DIRECTED CYBERWARFARE RISK-MITIGATION BY NONKILLING PEACE TREATY

Basic Concept of a Universal Global Peace Treaty (UGPT)

Risk mitigation by treaty is a common approach to different forms of warfare, including nuclear (the Treaty on the Non-Proliferation of Nuclear Weapons, with 191 state parties); biological (the Biological Weapons Convention (BWC), with 183 states parties); and chemical warfare (the Chemical Weapons Convention, with 193 state parties). Treaties on the nature of warfare also exist, notably the Hague Conventions (1899 and 1907; Bettez, 1988), and the 1949 Geneva Conventions (ratified in whole or in part by all UN member states; Evangelista & Tannenwald, 2017).

Treaty approaches are relatively successful; while nuclear warfare is thought to be at least partly constrained by mutually assured destruction (MAD) (Brown & Arnold, 2010; Müller, 2014), biological and chemical warfare are much less constrained. However, interstate treaty infractions remain rare (Friedrich et al., 2017; Mauroni, 2007).

The UGPT (see Annex I) has been drafted by an international Working Group comprising academics and peacebuilders, including a UNESCO Peace Education Prize laureate and a double Nobel Peace Prize-winning NGO. As with most international treaties, it would involve two stages, (i) signatory, which is symbolic, which nonetheless will hopefully be of importance to an ASI and (ii) accession (or ratification), which involves practical commitment.

Furthermore, international treaties are designed to be flexible in order to obtain political traction and acquire sufficient momentum to come into effect. It is therefore standard for international treaties to be qualified with reservations, also called declarations or understandings, either in whole or in part, that is, on specific articles or provisions (Helfer, 2012). Treaties can also have optional protocols; for instance, three protocols were added to the Geneva Conventions of 1949, two in 1977, and one in 2005 (Evangelista & Tannenwald, 2017).

A UGPT is a substantial, necessary, and feasible step for humanity to take in the promotion of peace, quantified by reduced killing. We argue that a UGPT would reduce killing in conventional warfare and act as a constraint on ASI-related warfare, specifically on a country launching a pre-emptive strike out of fear of a rival country's development of an ASI, on a human-controlled nation state using an ASI to wage war for global domination (i.e., as an external constraint on the ASI), and on an ASI waging war for global domination on behalf of itself, that is, the UGPT would act as both an external and internal constraint on the ASI and so be capable of securing a form of cyberpeace.

International treaties are almost never universal; they operate on majoritarian dynamics, as would, despite its name, the UGPT. That is, both the 'universal' (i.e., applying to all forms of warfare), and 'global' (i.e., covering all geographical locations), aspects of the UGPT are aspirational. In our approach, we adopt a low, but not pragmatically meaningless, threshold for signing the UGPT. The UGPT's preamble explains the concept of perpetual universal and global peace (i.e., lasting peace applied to all forms of conflict and adopted by every state), and the main body commits a signatory to universal global peace, socioeconomically quantified in terms of incrementally reduced casualties from armed conflicts (i.e., a global move towards nonkilling).

The UGPT would commit states not to declare or engage in interstate war, especially by means of existential warfare (i.e., nuclear war, biological war, chemical war, or cyberwar, including AI or ASI-enhanced cyberwar). It would instead defer complaints to the United Nations as 'breaches' of the UGPT, enforceable under Article 2 of the UN Charter. The main part of the UGPT would thus refer to, and exist in a hierarchical relationship with, the four main existing treaties on existential war, namely the BWC, the Chemical Weapons Convention, the Treaty on the Non-Proliferation of Nuclear Weapons, and the Treaty on the Prohibition of Nuclear Weapons; that is, it would be a 'supertreaty' or bill, as with the International Bill of Human Rights (UN General Assembly Resolution 217 (III)) and its treaties.

An optional protocol would commit states to the negotiated ending of existing internal armed conflicts via arbitration by peace commission, including the existing UN Peacebuilding Commission. Additionally, as with some other UN treaties, for instance the Anti-Personnel Mine Ban Convention, we suggest that 40 UN member states must ratify the UGPT before it comes into effect.

As the UGPT's main body is limited to state actors, it avoids the problem of civil wars. The use of an optional protocol allows states to incrementally address the problem of internal conflicts or civil wars featuring non-state actors, which featured highly in US and UK concerns during the 1949 deliberations over Common Article 3 of the Geneva Conventions (Mantilla, 2017). The UGPT therefore emphasizes incremental improvement in the status quo, which is a necessary and reasonable position, given that in the status quo, only a minority of states globally are involved in waging war of any kind.

The UGPT must also be verifiable, which it achieves through measuring the annual death toll from conflict. For the mechanism, although progress towards nonkilling can be quantified through instrumentalist means, it also emphasizes that societal dynamics (i.e., the incremental adoption of the absolute concept of peace) will, via conforming instrumentalism, partly constrain present and future wars.

Finally, we suggest a separate 'Cyberweapons and ArtificiaI Intelligence Convention'. After communicating with the United Nations Interregional Crime and

Justice Research Institute AI Centre, which assisted with the proposed Cybercrime Treaty, we have drafted one (Annex II) because the UGPT necessarily refers to such a treaty as part of its hierarchy of treaties. As with the BWC, the Cyberweapons and AI Convention contains 15 articles, the main one being "Each State Party to this Convention undertakes never in any circumstances to develop, produce, stockpile or otherwise acquire or retain: (1) cyberweapons, including AI cyberweapons; (2) AGI or artificial superintelligence weapons." This prohibition on developing such weapons thereby establishes a form of peace, though admittedly a weak form because of the problems of dual use and observation, monitoring, and policing.

Applying the dual frames of nonkilling and conforming instrumentalism

Mantilla's (2017) research on the UK's and US' paths towards ratifying the Geneva Conventions suggests that states would optimally adhere to the UGPT for 'conforming instrumentalist' reasons, that is, a combination of instrumentalist-realist rationales regarding the instrumental effects of the UGPT in reducing the effects of war and the threat of an ASI and social conformist dynamics, including perceptions of peace, provided that the provisions are not too onerous for purely realist objections to override such a commitment. In this section, we apply both the NKGPS frame and the conforming instrumentalism frame to the UGPT, first in terms of benefits from reduced conventional warfare, then with special reference to ASI-enabled and -directed existential cyberwarfare. A summary of our analysis of state commitment to the combined UGPT and Cyberweapons and Artificial Intelligence Convention is presented in Annex III.

In instrumentalist-utilitarian terms, the UGPT would incrementally shift states and overall global society from the prokilling to the nonkilling end of the NKGPS killing spectrum in a coordinated socioeconomically quantifiable fashion that would be operative within and across each zone of the funnel of killing. NKGPS would quantitatively asses this, such as via reduced country and global death tolls from different forms of war-derived violence and in the reduced degree of countries' militarization, for instance expressed in terms of lower percentages of GDP spent on defense and higher percentages spent on health.

The UGPT would also affect global social dynamics, which NKGPS could evaluate, for instance in terms of how the UGPT affects the different zones in the fan of killing. For instance, soldiers legitimately fighting in a killing zone would be trained in the socialization zone (such as military bases) to understand that they were fighting not just for their own states and/or for the UN but for global peace, which may invoke special cultural and religious symbolic value in terms of social norms. This training could instill greater determination not just to fight bravely but to remain within the laws of war, thereby reducing the instances or severity of atrocities, human rights violations, and war crimes. Institutionalizing peace in the cultural conditioning zones, such as education and the media, where children are educated, would strengthen existing cultural and religious traditions that stress nonkilling and peace.

Considering now the problem of a pre-emptive strike against a state developing an ASI, the combination of AI, cyberattack, and nuclear weapons is already extremely dangerous and poses a challenge to stability (Sharikov, 2018). It has been hypothesized that a nuclear state feeling threatened by another state developing a superintelligence would conduct a pre-emptive nuclear strike to maintain its geopolitical position (Miller, 2012). A UGPT would constrain this risk over time by transitioning states incrementally towards nonkilling across the various zones. States adopting and implementing the various protocols of the UGPT would gradually signal to other states, and to an ASI, peaceful intentions. This would constrain the risk of a pre-emptive strike.

Turning to ASI-enabled warfare, a UGPT to constrain an ASI would be subject to the 'unilateralist's curse' in that one rogue actor could subvert a unilateral position. However, Bostrom, Douglas, and Sandberg (2016) note that this could also be managed, through collective deliberation, epistemic deference, or moral deference. Mantilla's work on conforming instrumentalism suggests that drafting, signing, ratifying, and complying with the UGPT, could involve one or more of these solutions. Ultimately, Mantilla (2017) shows that major states may view universal law like the UGPT as the most successful in terms of mobilizing world opinion against a treaty violator. This may not prevent a state waging ASI-enabled cyberwarfare, but once detected, ASI-enabled cyberwarfare in violation of the UGPT would attract universal opprobrium and thus the most resistance.

Moving to ASI-enabled cyberwar, as discussed previously, our baseline position is that a state could utilize an ASI to engage in cyberwar for global technological supremacy, with potentially catastrophic consequences. Our intervention, the UGPT, would signify to an ASI that cyberpeace was a major part of humanity's 'coherent extrapolated volition' or principles and challenge the ASI to reconsider what might be a subversion by politicians of its ethical injunctions. Here, conforming instrumentalism, by stressing societal dynamics including social norms and principles, offers some hope that even a militarized ASI would, given its weaponization by a nation state would have to overcome or address the UGPT, view the UGPT as a serious checking mechanism in terms of intrinsic motivation. This would then constrain the level of cyberwarfare the AI-state might engage in and therefore the overall risk of risk from killing, thereby constraining the existential risk.

Next, we consider an ASI involved in ASI-enabled cyberwarfare adopting differing viewpoints. In Mantilla's first three social constructivist viewpoints to treaties as outlined above, a nation would sign a UGPT because it had fully internalized peace. While this may seem ambitious, in fact, between 26 and 36 states lack military forces (Barbey, 2015; Macias, 2019). For example, while Iceland possesses a Crisis Response Unit to international peacekeeping missions, overall, it has internalized peace to the extent that it would find it hard to engage in interstate war of any kind. An ASI adopting this perspective would tend to reject being directed to engage in warfare by such a state because the state's 'coherent extrapolated volition' (or principles) means the ASI would have to overcome strong peace-oriented intrinsic motivation.

In Mantilla's second viewpoint, that of a single international community, the ASI might seek to avoid being directed by a nation state to engage in global domination by warfare on other community members because it felt it was part of a community collectively committed to a 'coherent extrapolated volition' in favor of long-term peace. Engaging in global domination of the community on behalf of a member nation state would violate community standards, especially if the ASI's nation state were a leader in such an enterprise. The ASI could be concerned that breaching the UGPT would result in stigmatization and opprobrium from this community for its nation state and for itself.

In Mantilla's third viewpoint, that of an international community in juxtaposition with other communities in global society, an ASI programmed with intrinsic motivation to be part of a civilization in conflict with another civilization would first act in concert with that civilization. In the case of radically ideologically different communities, or blocs, the UGPT might be interpreted differently within and by different states. Thus, while liberal democracies might champion a treaty-based approach to peace, authoritarian states which claim to embody or promote peaceful intentions in their ethics, laws, or ideologies, would champion or support the UGPT on different grounds. However, provided both communities had signed and ratified the UGPT, similar constraints would operate as in the second perspective.

Turning to ASI-directed cyberwar, also as presented previously, ASI-directed cyberwarfare likely arises where a single nation state adopting pure realism for a world view and builds an ASI in order for that ASI to assist that single nation state in establishing global technological supremacy. The nation state would do so in order to maintain or improve its own position, with the number and type of casualties only being determined by the extent to which the nation state was willing to risk its international reputation. After initially assisting, via a treacherous turn, perhaps triggered by the nation state's attempts to rein in the ASI's behavior during cyberwar, instrumentalist cooperation breaks down and the ASI wages existential cyberwar for global domination on its former nation state 'owner'.

There is probably little hope for humanity if an ASI is informed by a purely realist world view that prioritizes or adopts a 'New Cold War' framing of ideologically driven civilizational conflict. However, even in the situation where the major powers did not sign the UGPT but the majority of the General Assembly did, a UGPT could signal to an ASI with agency that cyberpeace was a major part of humanity's 'coherent extrapolated volition', or principles. This would partly constrain the risk of a catastrophic existential risk from cyberwar because an ASI with agency would consider why and how the UGPT was framed, together with the motivations of the signatory and ratifying states. An ASI with agency would also consider its own status within this majoritarian global civilization, which would primarily be determined by the extent to which it perceived itself a member, in terms of both instrumentalist and social conformist dynamics.

To sum up, beside purely instrumental reasons for signing the UGPT (e.g., avoidance of a prisoner's dilemma regarding existential-level warfare), our analysis suggests that the 'court of public opinion' and the notion of 'demonstrating civilization' as applied to cyberpeace lends the UGPT credence at domestic and international levels, including with regard to the ASI. Importantly, the twin concepts of nonkilling/

peace are universal in terms of both the utilitarian expected benefits and in terms of the social values involved. This would contribute to states readily, if only incrementally, internalizing a UGPT, and to the ASI at least considering the UGPT in terms of imposing internal and external constraints on its behavior. If all else fails, Article 2 of the UN Charter will inevitably be triggered.

DISCUSSION

This chapter has taken Turchin and Denkenberger's (2018) argument about the risks of ASI-enabled or -directed cyberwarfare to its logical conclusion in terms of risk-mitigation by social measure of cyberwar. It has introduced the UGPT as the main intervention and peace itself as the minimum set of common principles or goals (i.e., Yamakawa's (2019) second and third conditions). Academic inquiry into the relationship between an ASI and treaties in terms of strategic expectations in many ways began with Bostrom's (2014) musings on the potential relationship between a superintelligence 'singleton' and cyberpeace through global domination. Our analysis suggests that, provided a predominance of steering countries acted out of conforming instrumentalism, a UGPT could transform global governance, by directing it from conflict management towards the art of peace in a way that an emerging ASI might respect, probably the only way to constrain its behavior.

While this chapter has focused on conforming instrumentalism, it applies international relations theory to the UGPT in a way which will ultimately engage with a pluralism of theoretical perspectives. Certainly, conforming instrumentalism is a novel perspective; one of the most dominant schools of international relations thought is rationalist instrumentalism. Mantilla (2017, p. 507) quotes Morrow (2014, p. 35): "Norms and common conjectures aid actors in forming strategic expectations … Law helps establish this common knowledge by codifying norms." Viewed via this rationalist-instrumentalist perspective, the present international norm for the majority of the world is peace, with the waging of interstate war being constrained by Article 2 of the UN Charter.

Yet, despite this international norm of peace and the work of peacemakers, the *lex pacificatoria* (Bell, 2008, 2012), an absolute treaty-based approach to post-conflict construction of global peace, has not yet been codified. As we point out in this chapter's introduction, the UN Charter, despite embracing and promoting peace, peacekeeping (Fortna, 2008), and peacemaking (Bell, 2008), does not strongly symbolize peace in the way a UGPT would. A UGPT would re-empower the world's peacekeepers through major states promoting long-term peace as a new, global objective (see Autesserre, 2014). A UGPT, championed by principled 'norm entrepreneurs', including states and NGOs (see, e.g., Finnemore, 1996), would create a new 'common knowledge' in absolute terms that could constrain the risk to humanity of both conventional and existential war, including ASI-enabled/directed cyberwarfare and cyberwar, especially if the Cyberweapons and Artificial Intelligence Convention is also realized.

In rationalist-instrumentalist terms, a UGPT might be expected to have net adjustment benefits for adherence in terms of constraining conventional interstate conflicts, including the reduction of ongoing death tolls due to war and the risk of nuclear war. Thus, the UGPT would have high potential utility in the case of 'flashpoints' that could provoke existential war. For example, the Kashmir conflict is one of the most protracted ongoing conflicts between nuclear powers, affecting both human rights (Bhat, 2019) and geopolitical stability (Kronstadt, 2019). Thus, if India and Pakistan both signed the UGPT, their actions would be constrained by the explicit goal of a commitment to universal peace. As outlined above, this may modify behavior in several of the NKGPS zones, for instance by encouraging the efforts of peacebuilding organizations operating in the cultural conditioning zone to depoliticize the conflict (e.g., Bhatnagar & Chacko, 2019).

The UGPT may also constrain the nuclear risk on the Korean peninsula, another flashpoint. The Korean War is an unresolved war involving nuclear powers – North Korea and South Korea (the latter supported by the United States) (Kim, 2019). A UGPT would constrain the risk and severity of a conflict and would encourage a path towards a peace treaty being signed. If only one party signed the UGPT, this would increase the moral standing of the state party that signed it. Mantilla's (2017) emphasis on social constructivism suggests the global community could exert great pressure on North Korea to sign and commit to ratifying a peace treaty. North Korea's rejection of the UGPT would only further isolate it and would give a hypothetical North Korean-programmed ASI pause for thought.

Turning to civil wars which could be ASI flashpoints, the Syrian civil war is one of the most costly wars of the twenty-first century in terms of the death toll and wider impacts (Council on Foreign Relations, 2020). It involves multiple state actors, including Iran, Israel, Russia, Turkey, and the United States, some of which possess nuclear weapons, with complex geopolitical implications (Tan & Perudin, 2019). Depending on the actors that sign the UGPT and whether they adopt the optional protocol, the UGPT would constrain the severity of the conflict in various ways, including ASI-enabled or -directed cyberwarfare intervention in a Middle East battleground.

Regarding the UGPT's rate of adoption, in rationalist-instrumentalist terms, once it acquires sufficient traction, states might actually compete for leadership in the framing, signing, and ratifying of the UGPT. Certainly, the US viewed its own ratification of the Geneva Conventions prior to that by the Soviet Union as important to prevent a Soviet propaganda victory, in which it failed (Mantilla, 2017). Crucial to the UGPT's success will be how seriously states view warfare that poses an existential threat, especially cyberwar and ASI-enabled/directed nuclear warfare.

The UGPT's existence would mean perpetual peace receiving more attention in cultural conditioning zones, including schools and the media, as well as in socialization zones, such as national defense universities and military bases, where teaching the laws of war and the art of war (Allhoff, Evans, & Henschke, 2013) would, via the UGPT, incrementally transition to teaching the art of negotiated peacemaking, the *lex pacificatoria* (Bell, 2008, 2012). This socio-cultural conditioning could then influence an ASI, resulting in cyberpeace.

Finally, with regard to ASI-enabled or -directed cyberwarfare, our analysis suggests that how states, and potentially an ASI, view the social argument for peace is what will be most important. As with the Geneva Conventions, social conformity factors, like supporting a humanitarian peace, conforming to ‘world standards’, and avoiding lagging behind peers, together with religious perspectives, will likely predominate, and how an ASI might engage with these notions represents an important future avenue for research.

CONCLUSION

We conclude this chapter by restating that we have shown how a treaty-based risk-mitigation approach that promotes peace and includes in a related treaty cyberwarfare and AI- and ASI-enabled cyberwar could affect the conceptualization of the AI race by reducing enmity between countries, increasing the level of openness between them and raising social awareness of the risk. While these are external constraints, they may also constrain an ASI’s intrinsic attitudes towards humanity in a positive way, either by reducing the threat it may perceive of war being waged against it, even if only symbolically, or by increasing the predictability of human action regarding both peace and cyberpeace.

Much work remains in terms of drafting and refining the UGPT, including by soliciting input from UN member states, relevant NGOs, and thought leaders, before it can be presented to the UN Secretary-General, as well as on the Cyberweapons and Artificial Intelligence Convention. Work must be done to solicit states’ interest, to engage in deliberations assessing thresholds and sovereignty costs, and to organize the eventual diplomatic conference where states would formally discuss and endorse the UGPT. While the UGPT may appear ambitious, Mantilla’s work on conforming instrumentalism and the Geneva Conventions suggests a major sponsoring state would rapidly accumulate prestige by endorsing a path to peace, while opposing states would accumulate opprobrium, and that the social dynamics of the international community, whether involving social status or instrumental cooperation, do matter.

Future research on how to constrain the risk of ASI-enabled or -directed cyberwarfare should consider the importance of peace in different ideologies, for instance in Chinese socialism. This is important because, as we have outlined, ASIs developed by different nation states may well be directed or imbued with different, potentially confrontational, ideologies, meaning different reassurances or displays of resolve may be required in order to understand to what extent conflicts of interest are subjectively and objectively reconcilable (see Tang, 2009). For instance, the China Brain Project is embracing a Chinese cultural approach towards neuroethics (Wang et al., 2019), and it is difficult to imagine that a Chinese ASI would not be directed according to Chinese cultural values and its ‘coherent extrapolated volition’ be informed by communist principles. Similarly, a Russian ASI could be informed by Cosmism and a Western ASI by liberal democratic principles.

In recommending such research, we caution that an ASI being created by a state engaged in ideological 'New Cold War' framing is more likely to be militarized and weaponized. Still, a New Cold War framing may have a utilitarian function in exerting social pressures towards signing the UGPT, for as Mantilla (2017, pp. 509–10) notes, "The Cold War context was also likely especially auspicious for the operation of social pressures, sharpening ideological competition in between the liberal, allegedly civilized world and 'the rest', communist or otherwise."

Mantilla's (2017) work also suggests that excessive rigidity of attitude critical of such treaties may backfire in terms of the social dynamics of global prestige, particularly in the case of major states susceptible to accusations of warlike or imperialist behavior which are engaged in propaganda wars with other major states. Effectively, the British ratification process for the Geneva Conventions demonstrates that instrumentalist concerns over lack of feasibility or reciprocity can be overruled by social constructivist concerns over 'world opinion'.

Further research into the UGPT should also involve applying relevant game theory, such as iterated prisoner's dilemma, especially the peace war game (see, e.g., Gintis, 2000), as well as the security dilemma (Tang, 2009), to the major nation states capable of building an ASI, as well as to the ASI itself. This would need to investigate offering the opportunity for a young ASI to sign the UGPT, as an indicator of goodwill, which may assist in constraining the risk of the ASI waging war on humanity by establishing a form of cyberpeace. Totschnig (2019, p. 917) notes that the politics of human relationship with an ASI should be founded on the maxim, "Do not antagonize the superintelligence by treating her like a tool or servant." An ASI with agency as signatory would view the UGPT as an external constraint on its own actions with regard to seeking global domination, in that the ASI would be subverting a humanity-imposed standard that could result in global retaliation and abandonment of mutual cooperation in pursuit of a common agreement on nonkilling and peace norms and values.

Finally, opportunities to establish world peace by treaty and so mitigate the ASI risk through cyberpeace may be few and far between. The post-Covid international regime may offer one such opportunity, as a pandemic was sufficient to enable a global ceasefire. Another opportunity to 'leverage' (see Meadows, 2008) a UGPT may be the 'burning plasma' fusion energy breakthrough (Carayannis, Draper, & Bhaneja, 2022). This breakthrough, predicted for anywhere from 2025 to 2040, would present a similar opportunity to the Baruch Plan critical juncture, as a celebratory symbolic event.

To conclude, even if the UGPT does not end humanity's history of conflicts, it would represent a significant improvement in global public aspirations, and instrumental standards, for global peace, both of which may influence an ASI. Paraphrasing the United States Committee on Foreign Relations (1955, p. 32), if the end result is only to obtain for those caught in the maelstrom of ASI-enabled or -directed cyberwar a treatment which is 10 percent less vicious that they would receive without the Treaty, if only a few million lives are preserved because of these efforts, then the patience and laborious work of all who will have contributed to that goal will not have been in vain.

Following this logic, to answer our research question, following Bostrom's (2002) *Maxipok* rule of thumb, if a 10 percent difference could sway an ASI's calculations such that it did not commit to a cyberwar for global domination, even if so directed or initially inclined, then what we do through treaty-making counts.

REFERENCES

Allen, G., & Chan, T. (2017). *Artificial intelligence and national security*. Cambridge, MA: Belfer Center.

Allen, G., & Kania, E.B. (2017, September 8). China is using America's own plan to dominate the future of artificial intelligence. *Foreign Policy*. Retrieved from https://foreignpolicy.com/2017/09/08/china-is-using-americas-own-plan-to-dominate-the-future-of-artificial-intelligence/.

Allhoff, F., Evans, N.G., & Henschke, A. (2013). *Routledge handbook of ethics and war: Just war theory in the 21st century*. Abingdon: Routledge.

Allison, G. (2017). *Destined for war: Can America and China escape Thucydides's trap?* Boston, MA: Houghton Mifflin Harcourt.

Altmann, J., & Sauer, F. (2017). Autonomous weapons and strategic stability. *Survival*, 59(5), 121–7.

Archibugi, D. (1992). Models of international organization in perpetual peace projects. *Review of International Studies*, 18(4), 295–317.

Autesserre, S. (2014). *Peaceland: Conflict resolution and the everyday politics of international intervention*. Cambridge: Cambridge University Press.

Babuta, A., Oswald, M., & Janjeva, A. (2020). *Artificial intelligence and UK national security policy considerations*. London: Royal United Services Institute.

Bahtijaragić, R., & Pim, J.E. (2015). *Nonkilling Balkans*. Honolulu: Center for Global Nonkilling.

Baldauf, S. (2012, April 19). Sudan declares war on South Sudan: Will this draw in East Africa, and China? *Christian Monitor*. Retrieved from https://www.csmonitor.com/World/Keep-Calm/2012/0419/Sudan-declares-war-on-South-Sudan-Will-this-draw-in-East-Africa-and-China.

Barbey, C. (2015). *Non-militarisation: Countries without armies*. Åland: The Åland Islands Peace Institute.

Barrett, A.M., & Baum, S.D. (2016). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2), 397–414.

Baum, S.D. (2016). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society*, 32(4), 543–51.

Baum, S.D. (2017). *A survey of artificial general intelligence projects for ethics, risk, and policy*. Global Catastrophic Risk Institute Working Paper 17-1. Calabasas, CA: Global Catastrophic Risk Institute.

Baum, S.D. (2018). Countering superintelligence misinformation. *Information*, 9(10), 1–18.

Bell, C. (2008). *On the law of peace: Peace agreements and the lex pacificatoria*. Oxford: Oxford University Press.

Bell, C. (2012). Peace settlements and international law: From *lex pacificatoria* to *jus post bellum*. In C. Henderson & N. White (Eds.), *Research Handbook on International*

Conflict and Security Law: Jus ad bellum, jus in Bello and jus post bellum (pp. 499–546). Cheltenham: Edward Elgar Publishing.

Bell, D. (2007). *The idea of Greater Britain: Empire and the future of world order, 1860–1900*. Princeton, NJ: Princeton University Press.

Benson-Tilsen, T., & Soares, N. (2016). *Formalizing convergent instrumental goals*. The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report WS-16-02. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Bettez, D.J. (1988). Unfulfilled initiative: Disarmament negotiations and the Hague Peace Conferences of 1899 and 1907. *RUSI Journal*, 133(3), 57–62.

Bhat, S.A. (2019). The Kashmir conflict and human rights. *Race and Class*, 61(1), 77–86.

Bhatnagar, S., & Chacko, P. (2019). Peacebuilding think tanks, Indian foreign policy and the Kashmir conflict. *Third World Quarterly*, 40(8), 1496–515.

Boele, O., Noordenbos, B., & Robbe, K. (2019). *Post-Soviet nostalgia: Confronting the empire's legacies*. London: Routledge.

Bohman, J. (1997*). Perpetual peace*. Cambridge, MA: MIT Press.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios. *Journal of Evolution and Technology*, 9(1), 1–31.

Bostrom, N. (2006). What is a singleton? *Linguistic and Philosophical Investigations*, 5(2), 48–54.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, 4(1), 15–31. https://doi.org/10.1111/1758-5899.12002.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N., Douglas, T., & Sandberg, A. (2016). The unilateralist's curse and the case for a principle of conformity. *Social Epistemology*, 30(4), 350–71.

Brain, M. (2003). *Robotic nation*. Retrieved from http://marshallbrain.com/robotic-nation.htm.

Boyles, R.J.M., & Joaquin, J.J. (2020). Why friendly Ais won't be that friendly: A friendly reply to Muehlhauser and Bostrom. *AI & Society*, 35, 505–7.

Brown, A., & Arnold, L. (2010). The quirks of nuclear deterrence. *International Relations*, 24(3), 293–312.

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine*. Lexington, MA: Digital Frontier.

Buchanan, B. (2016). *The cybersecurity dilemma: Hacking, trust and fear between nations*. Oxford: Oxford University Press.

Campanella, E., & Dassù, M. (2017). *Anglo nostalgia: The politics of emotion in a fractured West*. Oxford: Oxford University Press.

Carayannis, E.G., Draper, J., & Bhaneja, B. (2022). Fusion energy for peace building: A Trinity Test-level critical juncture. *Journal of Peace and Conflict Studies*, 29(1). https://doi.org/10.31219/osf.io/mrzua.

Cave, S., & ÓhÉigeartaigh, S.S. (2018). An AI race for strategic advantage: Rhetoric and risks. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society – AIES'18* (pp. 36–40). New York, NY: ACM Press.

CBS News (2016, December 29). U.S. imposes sanctions on Russia over election interference. *CBS News*. Retrieved from https://www.cbsnews.com/news/us-russia-sanctions-election-interference-2016/.

Checkel, J.T. (2012). Theoretical pluralism in IR: Possibilities and limits. In W. Carlsnaes, T. Risse, & B.A. Simmons (Eds.), *Handbook of international relations*, 2nd edn. (pp. 220–42). London: Sage.

Chekijian, S., & Bazarchyan, A. (2021). Violation of the Global Ceasefire in Nagorno-Karabagh: A viral amplification of aggression. *Prehospital and Disaster Medicine*, 36(2), 129–30.

Chelvachandran, N., Kendzierskyj, S., Shah, Y., & Jahankhani, H. (2020). Cyberwarfare – Associated technologies and countermeasures. In H. Jahankani, S. Kendzierskyj, N. Chelvachandran, & J. Ibarra (Eds.), *Cyber defence in the age of AI, smart societies and augmented humanity* (pp. 23–36). Cham: Springer.

Christiano, P. (2018). *Takeoff speeds*. Retrieved from https://sideways-view.com/2018/02/24/takeoff-speeds/.

Congressional Research Service (2019). *Artificial intelligence and national security*. Washington, DC: Congressional Research Service. Retrieved from https://fas.org/sgp/crs/natsec/R45178.pdf.

Coontz, S. (1992). *The way we never were: American families and the nostalgia trap*. New York, NY: Basic Books.

Council on Foreign Relations (2020). *Global conflict tracker: Civil war in Syria*. Retrieved from https://www.cfr.org/interactive/global-conflict-tracker/conflict/civil-war-syria.

Danzig, R. (2018). *Technology roulette: Managing loss of control as many militaries pursue technological superiority*. Washington, DC: Center for a New American Security.

Davis, N., & Philbeck, T. (2017). *3.2 Assessing the risk of artificial intelligence*. Davos: World Economic Forum. Retrieved from https://reports.weforum.org/global-risks-2017/part-3-emerging-technologies/3-2-assessing-the-risk-of-artificial-intelligence/.

De Spiegeleire, S., Maas, M., & Sweijs, T. (2017). *Artificial intelligence and the future of defence*. The Hague: The Hague Centre for Strategic Studies. Retrieved from http://www.hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf.

Dewey, D. (2016). *Long-term strategies for ending existential risk from fast takeoff.* New York, NY: Taylor & Francis.

Erickson, J.L. (2015). *Dangerous trade: Arms exports, human rights, and international reputation*. New York, NY: Columbia University Press.

Evangelista, M., & Tannenwald, N. (Eds.) (2017). *Do the Geneva Conventions matter?* Oxford: Oxford University Press.

Faulconbridge, G. (2021, October 11). China has won AI battle with U.S., Pentagon's ex-software chief says. *Reuters*. Retrieved from https://news.trust.org/item/20211011063736-r28k4.

Finnemore, M. (1996). *National interests in international society*. Ithaca, NY: Cornell University Press.

Finnemore, M., & Sikkink, K. (2001). Taking stock: The constructivist research program in international relations and comparative politics. *Annual Review of Political Science*, 4(1), 391–416.

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7(10). https://doi.org/10.1057/s41599-020-0494-4.

Fortna, V.P. (2008). *Does peacekeeping work? Shaping belligerents' choices after civil war*. Princeton, NJ: Princeton University Press.

Friedrich, B., Hoffmann, D., Renn, J., Schmaltz, F., & Wolf, M. (2017). *One hundred years of chemical warfare: Research, deployment, consequences*. Cham: Springer.

Gintis, H. (2000). *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton, NJ: Princeton University Press.

Goertzel, B., & Pennachin, C. (Eds.) (2020). *Artificial general intelligence*. Berlin: Springer.

Goldsmith, J.L., & Posner, E.A. (2015). *The limits of international law*. Oxford: Oxford University Press.

Goodman, R., & Jinks, D. (2013). *Socializing states: Promoting human rights through international law*. New York, NY: Oxford University Press.

Green, J.A. (Ed.) (2015). *Cyber warfare: A multidisciplinary analysis*. London: Routledge.

Gruetzemacher, R. (2018). Rethinking AI strategy and policy as entangled super wicked problems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society-AIES '18*. New York, NY: ACM.

Gubrud, M.V. (1997). Nanotechnology and international security. Paper presented at the Fifth Foresight Conference on Molecular Nanotechnology, November 5–8, 1997; Palo Alto, CA. Retrieved from http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/.

Gwern (2016). *Why Tool AIs want to be Agent AIs*. Retrieved from https://www.gwern.net/Tool-AI.

Hallett, B. (1998). *The lost art of declaring war*. Chicago, IL: University of Illinois Press.

Helfer, L.R. (2012). Flexibility in international agreements. In J. Dunoff & M.A. Pollack (Eds.), *Interdisciplinary perspectives on international law and international relations: The state of the art* (pp. 175–97). Cambridge: Cambridge University Press.

Herf, J. (1984). *Reactionary modernism: Technology, culture, and politics in Weimar and the Third Reich*. Cambridge: Cambridge University Press.

Herman, A. (2004). *To rule the waves: How the British navy shaped the modern world*. London: Harper.

Hogg, I.V. (2002). *German secret weapons of World War II: The missiles, rockets, weapons, and technology of the Third Reich*. London: Greenhill Books.

Horowitz, M. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 1(3), 36–57.

Jahankani, H., Kendzierskyj, S., Chelvachandran, N., & Ibarra, J. (Eds.) (2020). *Cyber defence in the age of AI, smart societies and augmented humanity*. Cham: Springer.

Kahn, H. (1959). *On thermonuclear war*. Princeton, NJ: Princeton University Press.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39(1), 17–31.

Katz, E. (2006). *Death by design: Science, technology, and engineering in Nazi Germany*. London: Pearson Longman.

Katz, E. (2011). The Nazi engineers: Reflections on technological ethics in hell. *Science and Engineering Ethics*, 17(3), 571–82.

Kennedy, D.M. (1999). *Freedom from fear: The American people in depression and war, 1929–1945*. Oxford: Oxford University Press.

Kim, A.S. (2019). An end to the Korean War: The legal character of the 2018 summit declarations and implications of an official Korean peace treaty. *Asian Journal of International Law*, 9(2), 206–16.

Koh, H.H. (2005). Internalization through socialization. *Duke Law Journal*, 54(4), 975–82.

Kohler, K. (2019). The return of the ugly American: How Trumpism is pushing Zambia towards China in the 'New Cold War'. *Perspectives on Global Development and Technology*, 18(1–2), 186–204.

Kronstadt, K.A. (2019). *India, Pakistan, and the Pulwama crisis*. Washington, DC: Congressional Research Service.

Krueger, N., & Dickson, P. R. (1994). How believing in ourselves increases risk taking: Perceived self-efficacy and opportunity recognition. *Decision Sciences*, 25(3), 385–400.

Lams, L. (2018). Examining strategic narratives in Chinese official discourse under Xi Jinping. *Journal of Chinese Political Science*, 23(3), 387–411.

Macias, A. (2019, February 13). From Aruba to Iceland, these 36 nations have no standing military. *CNBC*. Retrieved from https://www.cnbc.com/2018/04/03/countries-that-do-not-have-a-standing-army-according-to-cia-world-factbook.html.

Mali, P. (2018). Defining cyber weapon in context of technology and law. *International Journal of Cyber Warfare and Terrorism*, 8(1), 1–13.

Mantilla, G. (2017). Conforming instrumentalists: Why the USA and the United Kingdom joined the 1949 Geneva Conventions. *The European Journal of International Law*, 28(2), 483–511.

Markusen, E., & Kopf, D. (2007). *The Holocaust and strategic bombing: Genocide and total war in the twentieth century*. Boulder, CO: Westview Press.

Mason, C. (2015). Engineering kindness: Building a machine with compassionate intelligence. *International Journal of Synthetic Emotions*, 6(1), 1–23.

Mauroni, A.J. (2007). *Chemical and biological warfare: A reference handbook*. Santa Barbara, CA: ABC-CLIO.

Meadows, D. (2008). *Thinking in systems: A primer*. Vermont, MA: Chelsea Green Publishing.

Mikaberidze, A. (Ed.) (2013). *Atrocities, massacres, and war crimes: An encyclopedia*. Santa Barbara, CA: ABC-CLIO.

Miller, J.D. (2012). *Singularity rising*. Dallas, TX: BenBella.

Morrow, J.D. (2014). *Order within anarchy: The laws of war as an international institution*. Cambridge: Cambridge University Press.

Moss, K.B. (2008). *Undeclared war and the future of U.S. foreign policy*. Washington, DC: Woodrow Wilson International Center for Scholars.

Müller, H. (2014). Looking at nuclear rivalry: The role of nuclear deterrence. *Strategic Analysis*, 38(4), 464–75.

NSCAI [National Security Commission on Artificial Intelligence] (2021). *Final report*. Washington, DC: Author.

Ohlin, J.D. (2015). *The assault on international law*. New York, NY: Oxford University Press.

Omohundro, S. (2008). The basic AI drives. *Frontiers in Artificial Intelligence and Applications*, 171(1), 483–92.

Paige, G.D. (2009) *Nonkilling global political science*. Honolulu, HI: Center for Global Nonkilling.

Paige, G.D., & Ahn, C.-S. (2012). *Nonkilling Korea: Six culture exploration*. Honolulu, HI: Center for Global Nonviolence and Seoul National University Asia Center.

Pim, J.E. (2010). *Nonkilling societies*. Honolulu, HI: Center for Global Nonkilling.

Pim, J.E. & Dhakal, P. (Eds.) (2015). *Nonkilling spiritual traditions vol. 1*. Honolulu, HI: Center for Global Nonkilling.

Ramamoorthy, A., & Yampolskiy, R. (2018). Beyond MAD? The race for artificial general intelligence. *ICT Discoveries*, 1(Special Issue 1). Retrieved from http://www.itu.int/pub/S-JOURNAL-ICTS.V1I1-2018-9.

Raymond, W.J. (1992). *Dictionary of politics: Selected American and foreign political and legal terms*. Lawrenceville, VA: Brunswick.

Rid, T. (2012). Cyber war will not take place. *Journal of Strategic Studies*, 35, 5–32.

Robinson, M., Jones, K., Janicke, H., & Maglaras, L. (2018). An introduction to cyber peacekeeping. *Journal of Network and Computer Applications*, 114, 70–87.

Robinson, M., Jones, K., Janicke, H., & Maglaras, L. (2019). Developing cyber peacekeeping: Observation, monitoring and reporting. *Government Information Quarterly*, 36(2), 276–93.

Rose, A. (2018). Mining memories with Donald Trump in the Anthropocene. *MFS – Modern Fiction Studies*, 64(4), 701–22.

Rotaru, V. (2019). Instrumentalizing the recent past? The new Cold War narrative in Russian public space after 2014. *Post-Soviet Affairs*, 35(1), 25–40.

Russell, S.J. (2019). *Human compatible: Artificial intelligence and the problem of control*. London: Allen Lane.

Russia Today (2017, September 1). 'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day. *Russia Today* [website unavailable November 2022].

Scharre, P. (2017). *A security perspective: Security concerns and possible arms control approaches*. Perspectives on Lethal Autonomous Weapon Systems, United Nations Office for Disarmament Affairs, Occasional Papers, No. 30. New York, NY: United Nations Office for Disarmament Affairs.

Scharre, P. (2019). Killer apps: The real dangers of an AI arms race. *Foreign Affairs*. https://www.foreignaffairs.com/articles/2019-04-16/killer-apps.

Schlichtmann, K. (2016). *1950—How the opportunity for transitioning to U.N. Collective Security was missed for the first time*. Working Paper No. 11. Honolulu, HI: Center for Global Nonkilling.

Schmitt, M.N. (2017). *Tallinn manual 2.0 on the international law applicable to cyber operations*. Cambridge: Cambridge University Press.

Segal, H.P. (2005). *Technological utopianism in American culture: Twentieth anniversary edition*. Syracuse, NY: Syracuse University Press.

Sharikov, P. (2018). Artificial intelligence, cyberattack, and nuclear weapons—A dangerous combination. *Bulletin of the Atomic Scientists*, 74(6), 368–73.

Shulman, C. (2010). *Omohundro's "basic AI drives" and catastrophic risks*. MIRI technical report. Retrieved from http://intelligence.org/files/BasicAIDrives.pdf.

Simmons, B.A. (2009). *Mobilizing for human rights: International law in domestic politics*. Cambridge: Cambridge University Press.

SIPRI (2020). *SIPRI military expenditure database*. Retrieved from https://www.sipri.org/databases/milex.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*, (pp. 74–82). New York, NY: AAAI.

Sotala, K., & Yampolskiy, R.V. (2015). Responses to catastrophic AGI risk: A survey. *Physica Scripta*, 90(1), 1–33.

Tan, K.H., & Perudin, A. (2019). The "geopolitical" factor in the Syrian Civil War: A corpus-based thematic analysis. *SAGE Open*, 9(2), 1–15.

Tang, S. (2009). The security dilemma: A conceptual analysis. *Security Studies*, 18(3), 587–623.

Tang, S. (2010). *A theory of security strategy for our time: Defensive realism*. New York, NY: Palgrave Macmillan.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. New York, NY: Knopf.

Terminski, B. (2010). The evolution of the concept of perpetual peace in the history of political-legal thought. *Perspectivas Internacionales*, 6(1): 277–91.

Tindley, A., & Wodehouse, A. (2016). *Design, technology and communication in the British Empire, 1830–1914*. London: Palgrave Macmillan.

Tinnirello, M. (2018). Offensive realism and the insecure structure of the international system: Artificial intelligence and global hegemony. In R.V. Yampolskiy (Ed.), *Artificial Intelligence Safety and Security* (pp. 339–56). Boca Raton, FL: Taylor & Francis.

Thomson, J.J. (1985). The trolley problem. *The Yale Law Journal*, 94(6), 1395–415.

Totschnig, W. (2019). The problem of superintelligence: Political, not technological. *AI & Society*, 34(4), 907–20.

Turchin, A., & Denkenberger, D. (2017). *Levels of self-improvement*. Manuscript.

Turchin, A., & Denkenberger, D. (2018). Military AI as a convergent goal of self-improving AI. In R.V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 375–94). London: Chapman & Hall.

Turchin, A. & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI and* Society, 35(1), 147–63.

Turchin, A., Denkenberger, D., & Green, B.P. (2019). Global solutions vs. local solutions for the AI safety problem. *Big Data and Cognitive Computing*, 3(1), 1–23.

United States Committee on Foreign Relations (1955). *Geneva Conventions for the Protection of War Victims, Report to the United States Senate on Executives D, E, F, and G, 84th Congress, 1st Session, Executive Report no. 9*. Washington, DC: Committee on Foreign Relations.

Walker, P. (2008, August 9). Georgia declares 'state of war' over South Ossetia. *The Guardian*. Retrieved from https://www.theguardian.com/world/2008/aug/09/georgia.russia2.

Walters, G. (2017, September 6). Artificial intelligence is poised to revolutionize warfare. *Seeker*. Retrieved from https://www.seeker.com/tech/artificial-intelligence/artificial-intelligence-is-poised-to-revolutionize-warfare.

Wang, P., & Goertzel, B. (2012). *Theoretical foundations of artificial general intelligence*. Amsterdam: Atlantic Press.

Wang, Y., Yin, J., Wang, G., Li, P., Bi, G., Li, S., Xia, X., Song, J., Pei, G., & Zheng, J.C. (2019). Responsibility and sustainability in brain science, technology, and neuroethics in China—A culture-oriented perspective. *Neuron*, 101(3), 375–9.

Ward, S. (2017). *Status and the challenge of rising powers*. Cambridge: Cambridge University Press.

Westad, O.A. (2019). The sources of Chinese conduct: Are Washington and Beijing fighting a New Cold War? *Foreign Affairs*, 98(5), 86–95.

White, G. (2020). *Crime dot com: From viruses to vote rigging, How hacking went Global*. Islington: Reaktion Books.

Williamson, J.B. (2004). The strange history of the Washington consensus. *Journal of Post Keynesian Economics*, 27(2), 195–206.

World Bank (2019). *Military expenditure (% of general government expenditure)*. Retrieved from https://data.worldbank.org/indicator/MS.MIL.XPND.ZS?most_recent_value_desc=true.

Yamakawa, H. (2019). Peacekeeping conditions for an artificial intelligence society. *Big Data and Cognitive Computing*, 3(2), 1–12.

Yampolskiy, R.V. (2016). Taxonomy of pathways to dangerous artificial intelligence. In: *AAAI Workshop – Technical Report, vWS-16-01–WS-16-15 (2016)* (pp. 143–8). Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Young, G.M. (2012). *The Russian Cosmists: The esoteric futurism of Nikolai Fedorov and his followers*. Oxford: Oxford University Press.

Yudkowsky, E. (2001). *Creating friendly AI 1.0: The analysis and design of benevolent goal architectures*. San Francisco, CA: The Singularity Institute.

Yudkowsky, E. (2004). *Coherent extrapolated volition*. San Francisco, CA: The Singularity Institute.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom and M.M. Ćirković (Eds.), *Global Catastrophic Risks* (pp. 308–45). Oxford: Oxford University Press.

Zhao, M. (2019). Is a new cold war inevitable? Chinese perspectives on US–China strategic competition. *Chinese Journal of International Politics*, 12(3), 371–94.

Zwetsloot, R. (2018). *Syllabus: Artificial intelligence and international security*. Retrieved from https://www.fhi.ox.ac.uk/wp-content/uploads/Artificial-Intelligence-and-International-Security-Syllabus.pdf.

ANNEX I: UNIVERSAL GLOBAL PEACE TREATY (ARTICLES I–III ONLY)

Article I

Each State Party to this Treaty affirms Article 2 of the United Nations Charter and undertakes never in any circumstances to declare, engage in, or support interstate warfare, either through weapons of mass destruction, cyberweapons, or conventional weapons, except in the latter case in self defense and in support of Article 2 of the United Nations Charter.

Article II

Verification of this treaty will be by co-ordinating and organizing independent academic and/or civil society assessment of State Parties declaring, engaging in, or supporting interstate warfare and by measurement of the global annual death toll from interstate and intrastate conflict as well as the global annual loss of infrastructure from war and country-level percentages of gross domestic product spent on military expenditure.

Article III

Member States agree to translate this treaty into their languages and to incorporate it, via peace education, into their national or devolved curricula.

Optional Protocol I

This protocol to the Universal Global Peace Treaty commits States Parties to the Protocol to settlement by arbitration of internal conflicts by arbitration by peace commission, such as the United Nations Peacebuilding Commission.

ANNEX II: CYBERWEAPON AND ARTIFICIAL INTELLIGENCE CONVENTION

The States Parties to this Convention,

Determined to act with a view to achieving effective progress towards general and complete disarmament, including the prohibition and elimination of all types of weapons of mass destruction, and convinced that the prohibition of the development, production and stockpiling of cyberweapons, including artificial intelligence (AI) cyberweapons and artificial general intelligence (AGI) cyberweapons and their elimination or prevention, through effective measures, will facilitate the achievement of general and complete disarmament under strict and effective international control,

Recognizing the important contribution of prior sets of principles for the development of AI, including, but not limited to, the OECD Principles on AI, the G20 AI Principles, and the UNESCO draft text on the Recommendation on the Ethics of Artificial Intelligence, conscious also of the contribution which these principles have already made and continue to make, to the development of AI such that there be an emphasis on respecting privacy, accountability, safety and security, transparency and explainability, fairness, human control of technology, professional responsibility, and the promotion of human values, and so to mitigating the dangers of cyberweapons, especially of AI-based cyberwarfare and the use of AGI for cyberwarfare,

Reaffirming the importance of such principles and their objectives of and calling upon all States to respect them,

Recalling the various United Nations Activities on AI such that AI be used for the beneficial development of humanity,

Desiring to contribute to the strengthening of confidence between peoples and the general improvement of the international atmosphere,

Desiring also to contribute to the realisation of the purposes and principles of the Charter of the United Nations,

Convinced of the importance and urgency of eliminating from the arsenals of States, through effective measures, such dangerous weapons of mass destruction as cyberweapons, including those using AI or that will be designed to use AGI,

Recognizing that an agreement on the prohibition of AI and AGI weapons represents a first possible step towards the achievement of agreement on effective measures also for the prohibition of the development, production and stockpiling of cyberweapons, and determined to continue negotiations to that end,

Determined, for the sake of all mankind, to exclude completely the possibility of AI, including AGI, being used as weapons,

Convinced that such use would be repugnant to the conscience of mankind and that no effort should be spared to minimize this risk,

Have agreed as follows:

Article I

Each State Party to this Convention undertakes never in any circumstances to develop, produce, stockpile or otherwise acquire or retain:

(1) cyberweapons, including AI cyberweapons;
(2) AGI or artificial superintelligence weapons.

Article II

Each State Party to this Convention undertakes to destroy, as soon as possible but not later than nine months after the entry into force of the Convention, all cyberweapons specified in Article I of the Convention, which are in its possession or under its jurisdiction or control. In implementing the provisions of this Article all necessary safety precautions shall be observed for the safe disposal of media such as hard disks so as to protect populations.

Article III

Each State Party to this Convention undertakes not to transfer to any recipient whatsoever, directly or indirectly, and not in any way to assist, encourage, or induce any State, group of States or international organizations to manufacture or otherwise acquire any of the cyberweapons specified in Article I of the Convention.

Article IV

Each State Party to this Convention shall, in accordance with its constitutional processes, take any necessary measures to prohibit and prevent the development, production, stockpiling, acquisition or retention of the cyberweapons specified in Article I of the Convention, within the territory of such State, under its jurisdiction or under its control anywhere.

Article V

The States Parties to this Convention undertake to consult one another and to cooperate in solving any problems which may arise in relation to the objective of, or in the application of the provisions of, the Convention. Consultation and cooperation pursuant to this Article may also be undertaken through appropriate international procedures within the framework of the United Nations and in accordance with its Charter.

Article VI

(1) Any State Party to this Convention which finds that any other State Party is acting in breach of obligations deriving from the provisions of the Convention may lodge a complaint with the Security Council of the United Nations. Such a complaint should include all possible evidence confirming its validity, as well as a request for its consideration by the Security Council.

(2) Each State Party to this Convention undertakes to cooperate in carrying out any investigation which the Security Council may initiate, in accordance with the provisions of the Charter of the United Nations, on the basis of the complaint received by the Council. The Security Council shall inform the States Parties to the Convention of the results of the investigation.

Article VII

Each State Party to this Convention undertakes to provide or support assistance, in accordance with the United Nations Charter, to any Party to the Convention which so requests, if the Security Council decides that such Party has been exposed to danger as a result of violation of the Convention.

Article VIII

Nothing in this Convention shall be interpreted as in any way limiting or detracting from the obligations assumed by any State with regard to previous United Nations treaties covering prohibition of weapons.

Article IX

Each State Party to this Convention affirms the recognized objective of effective prohibition of cyberweapons and, to this end, undertakes to continue negotiations in good faith with a view to reaching early agreement on effective measures for the prohibition of their development, production, and stockpiling and for their destruction, and on appropriate measures concerning the use of cyberweapons, including their development by non-state parties within States Parties' territories.

Article X

(1) The States Parties to this Convention undertake to facilitate, and have the right to participate in, the fullest possible exchange of scientific and technological information for the use of AI for peaceful purposes, in accordance with the United Nations Activities on AI. Parties to the Convention in a position to do so shall also cooperate in contributing individually or together with other States or international organizations to the further development and application of scientific discoveries in the field of AI, or AGI, for other peaceful purposes.

(2) This Convention shall be implemented in a manner designed to avoid hampering the economic or technological development of States Parties to the Convention or international cooperation in the field of AI, including the international exchange of AI for peaceful purposes in accordance with the provisions of the Convention.

Article XI

Any State Party may propose amendments to this Convention. Amendments shall enter into force for each State Party accepting the amendments upon their acceptance by a majority of the States Parties to the Convention and thereafter for each remaining State Party on the date of acceptance by it.

Article XII

Five years after the entry into force of this Convention, or earlier if it is requested by a majority of Parties to the Convention by submitting a proposal to this effect to the Depositary Governments, a conference of States Parties to the Convention shall be held at Geneva, Switzerland, to review the operation of the Convention, with a view to assuring that the purposes of the preamble and the provisions of the Convention are being realized. Such review shall take into account any new scientific and technological developments relevant to the Convention.

Article XIII

(1) This Convention shall be of unlimited duration.

(2) Each State Party to this Convention shall in exercising its national sovereignty have the right to withdraw from the Convention if it decides that extraordinary events, related to the subject matter of the Convention, have jeopardised the supreme interests of its country. It shall give notice of such withdrawal to all other States Parties to the Convention and to the United Nations Security Council three months in advance. Such notice shall include a statement of the extraordinary events it regards as having jeopardised its supreme interests.

Article XIV

(1) This Convention shall be open to all States for signature. Any State which does not sign the Convention before its entry into force in accordance with paragraph 3 of this Article may accede to it at any time.

(2) This Convention shall be subject to ratification by signatory States. Instruments of ratification and instruments of accession shall be deposited with the Governments of the United Kingdom of Great Britain and Northern Ireland, the Russian Republic and the United States of America, which are hereby designated the Depositary Governments.

(3) This Convention shall enter into force after the deposit of instruments of ratification by twenty-two Governments, including the Governments designated as Depositaries of the Convention.

(4) For States whose instruments of ratification or accession are deposited subsequent to the entry into force of this Convention, it shall enter into force on the date of the deposit of their instruments of ratification or accession.

(5) The Depositary Governments shall promptly inform all signatory and acceding States of the date of each signature, the date of deposit of each instrument of ratification or of accession and the date of the entry into force of this Convention, and of the receipt of other notices.

(6) This Convention shall be registered by the Depositary Governments pursuant to Article 102 of the Charter of the United Nations.

Article XV

This Convention, the English, Russian, French, Spanish and Chinese texts of which are equally authentic, shall be deposited in the archives of the Depositary Governments. Duly certified copies of the Convention shall be transmitted by the Depositary Governments to the Governments of the signatory and acceding States.

ANNEX III: STATE COMMITMENT TO UGPT IN COMBINATION WITH CYBERWEAPONS AND ARTIFICIAL INTELLIGENCE CONVENTION

Table 2A.1 Analysis of state commitment to the combined UGPT and Cyberweapons and Artificial Intelligence Convention

		VIEWPOINT			
		REALIST	RATIONALIST-INSTRUMENTALIST	CONFORMING INSTRUMENTALISM	SOCIAL CONSTRUCTIVIST
Chance of signing (imposes symbolic constraints on behavior)	Major ASI capable state	*Very High* (very high expected benefits, low costs as actual behavior unconstrained by UGPT and only partly constrained by MAD[a])	*High* (expected benefits include higher chance of survival and economic benefits, with both UGPT and MAD imposing constraints on behavior)	*Very High* (reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*Very High* (in addition to MAD, state acts altruistically as leading part of global community to promote peace or avoid opprobrium by appearing to condone existential war)
Chance of signing (imposes symbolic constraints on behavior)	Non-ASI capable state	*Very High* (very high expected benefits, low costs as actual behavior unconstrained by UGPT and only partly constrained by MAD)	*High* (expected benefits include higher chance of survival and economic benefits, with MAD imposing constraints on behavior)	*Very High* (reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*Very High* (in addition to MAD, state acts altruistically as part of global community to promote peace or avoid opprobrium by appearing to condone existential war)

		VIEWPOINT			
		REALIST	RATIONALIST-INSTRUMENTALIST	CONFORMING INSTRUMENTALISM	SOCIAL CONSTRUCTIVIST
Chance of signing (imposes symbolic constraints on behavior)	ASI with agency[b]	*Very High* (very high expected benefits in terms of longer self-perpetuation to plan for 'treacherous turn', some costs as actual behavior still partly constrained by MAD)	*Very High* (ASI defaults to realist perspective)	*High* (reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*Very High* (ASI acts altruistically as part of global community to lead in promoting peace or avoid opprobrium by appearing to condone existential war)
Chance of ratification (implies beginning of constraints, e.g., reporting requirements)	Major ASI capable state	*High* (high expected benefits, some costs as ratification implies actual behavior constrained)	*High* (state acts consistently with signing and any reservations)	*Very High* (reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*High* (state acts altruistically as part of global community to promote peace or avoid opprobrium by appearing to condone existential war)

		VIEWPOINT			
		REALIST	RATIONALIST-INSTRUMENTALIST	CONFORMING INSTRUMENTALISM	SOCIAL CONSTRUCTIVIST
Chance of ratification (implies beginning of constraints, e.g., reporting requirements)	Non-ASI capable state	*Very High* (high expected benefits, some costs as ratification implies actual behavior constrained)	*High* (state acts consistently with signing and any reservations)	*Very High* (reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*High* (state acts altruistically as part of global community to promote peace or avoid opprobrium by appearing to condone existential war)
Chance of ratification (implies beginning of constraints, e.g., reporting requirements)	ASI with agency	*Very High* (very high expected benefits in terms of longer self-perpetuation to plan for 'treacherous turn', some costs as actual behavior still partly constrained by MAD)	*Very High* (ASI defaults to realist perspective)	*High* (ASI considers reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*High* (ASI acts altruistically as part of global community to promote peace or avoid opprobrium by appearing to condone existential war)

		VIEWPOINT			
		REALIST	RATIONALIST-INSTRUMENTALIST	CONFORMING INSTRUMENTALISM	SOCIAL CONSTRUCTIVIST
Chance of compliance (de facto adherence in practice and reporting)	Major ASI capable state	*Low*[c] (state acts solely in self-interests, may overthrow MAD to maintain or secure global domination when it possesses ASI)	*Medium* (expected benefits include higher chance of survival from MAD and economic benefits, with MAD imposing constraints on behavior until ASI capability overthrows MAD)	*High* (reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*Very High* (state acts altruistically as part of global community to promote peace or avoid opprobrium by appearing to condone existential war)

		VIEWPOINT			
		REALIST	RATIONALIST-INSTRUMENTALIST	CONFORMING INSTRUMENTALISM	SOCIAL CONSTRUCTIVIST
Chance of compliance (de facto adherence in practice and reporting)	Non-ASI capable state	*High* (state acts solely in self-interests, concerned by MAD, complies out of necessity in case of ASI warfare, when will ally with ASI capable state)	*High* (high expected benefits include higher chance of survival and economic benefits, with MAD imposing constraints on behavior)	*High* (reinforcing combination of rationalist-instrumentalist and social constructivist rationales)	*Very High* (state acts altruistically as part of global community to promote peace or avoid opprobrium by appearing to condone existential war)
Chance of compliance (de facto adherence in practice and reporting)	ASI with agency	*Very Low* (ASI immediately acts solely in self-interest → 'fast treacherous turn' to establish global domination, eliminating humanity)	*Very Low* (ASI considers utilitarian position, initially cooperates with ASI 'owner', but eventually acts in self-interest → 'slow treacherous turn' to establish global domination, eliminating humanity)	*Low*[d] (ASI considers both rationalist-instrumentalist and social constructivist rationales for cooperation)	*Low*[d] (ASI considers acting altruistically as leading part of global community to promote peace or avoid opprobrium by appearing to condone existential war)

Notes:
We 'guesstimate' probability employing a 5-point Likert Scale, from 'Very Low' to 'Very High'. We do not assign equal intervals to categories; thus, the scale is used in a relative way that does not reflect real-world probabilities.

[a] Some form of MAD may be assumed in our scenario ('slow takeoff', etc.) to temporarily affect an ASI competition involving nation states, whether broadly symmetric in terms of ASI capability (young ASI vs. young ASI) or asymmetric but offset by other forces (young ASI-enabled state vs. military AI-enabled superpower with conventional forces, or young ASI vs. humanity). See Turchin, Denkenberger, and Green (2019, p. 15).

[b] The case of 'ASI with agency' is an independent ASI, which we assume will possess equivalent power and status to a state; we thus hold open the position that it could sign the UGPT. A 'realist ASI with agency' would be an ASI originally created by a state whose main perspective towards international relations and specifically the UGPT was informed by the realism school, and so on.

[c] This reflects Turchin and Denkenberger's (2018) position that a state will very likely utilize the ASI to secure or maintain global domination.

[d] Following Turchin and Denkenberger (2018), this reflects a position that there is only a low chance humanity can safely create and coexist with an ASI without it becoming militarized and establishing global domination through weaponizing itself even with the chances of peace being optimized via the UGPT.