AI & SOCIETY
https://doi.org/10.1007/s00146-021-01382-y

OPEN FORUM

Optimising peace through a Universal Global Peace Treaty to constrain the risk of war from a militarised artificial superintelligence

Elias G. Carayannis[1] · **John Draper**[2]

Received: 24 October 2021 / Accepted: 15 December 2021

Abstract
This article argues that an artificial superintelligence (ASI) emerging in a world where war is still normalised constitutes a catastrophic existential risk, either because the ASI might be employed by a nation–state to war for global domination, i.e., ASI-enabled warfare, or because the ASI wars on behalf of itself to establish global domination, i.e., ASI-directed warfare. Presently, few states declare war or even war on each other, in part due to the 1945 UN Charter, which states Member States should "refrain in their international relations from the threat or use of force", while allowing for UN Security Council-endorsed military measures and self-defense. As UN Member States no longer declare war on each other, instead, only 'international armed conflicts' occur. However, costly interstate conflicts, both hot and cold and tantamount to wars, still take place. Further, a New Cold War between AI superpowers looms. An ASI-directed/enabled future conflict could trigger total war, including nuclear conflict, and is therefore high risk. Via conforming instrumentalism, an international relations theory, we advocate risk reduction by optimising peace through a Universal Global Peace Treaty (UGPT), contributing towards the ending of existing wars and prevention of future wars, as well as a Cyberweapons and Artificial Intelligence Convention. This strategy could influence state actors, including those developing ASIs, or an agential ASI, particularly if it values conforming instrumentalism and peace.

Keywords AI arms race · Artificial superintelligence · Conforming instrumentalism · Existential risk · International relations · Peace

1 Introduction

1.1 The problem of a warring artificial superintelligence (ASI)

While some maintain an artificial general intelligence (AGI), i.e., human or above human artificial intelligence (AI), is impossible (Fjelland 2020), others believe an AGI is attainable (Goertzel and Pennachin 2020; Wang and Goertzel 2012). In the latter instance, that the world has not attained global peace is a risk factor for the development of an AGI (Yamakawa 2019). Consequently, the international defence community is beginning to consider the national security risk posed by AGI development and its implications for international relations (IR), including calls to act (De Spiegeleire et al. 2017:107).

One risk is that a single nation–state developing an AGI could 'lock in' economic or military supremacy as an 'end point' to competition in international politics, as that state would be able to prevent a rival AGI being developed and through accumulating power would establish global domination (Horowitz 2018:54). AI is already a major national security issue because it can be militarized, employed in adversarial contexts, and provide a decisive advantage in terms of economic, information and military superiority (Allen and Chan 2017; Babuta et al. 2020; National Security Commission on Artificial Intelligence [NSCAI] 2021). Consequently, the 2021 NSCAI report urges that the US attain military AI readiness by 2025; thus, AI is important for waging decisive war.

For the major powers, AI technological supremacy, generated by economic power, is already viewed as paramount to national security and global leadership (NSCAI 2021:7):

✉ John Draper
johndr@kkumail.com

[1] George Washington University (European Union Research Center), Washington, DC, USA

[2] Nonkilling Economics and Business Research Committee, Center for Global Nonkilling, Honolulu, HI, USA

Published online: 11 January 2022

Military AI is potentially revolutionary as it could outstrip the pace of human decision-making, "potentially resulting in a loss of human control in warfare" (Congressional Research Service [CRS] 2020:37). It also constitutes an unpredictable threat: "AI systems capable of inherently unpredictable actions in close proximity to an adversary's systems may result in inadvertent escalation or miscalculation" (CRS 2020:37).

Additionally, while Baum (2017) found little evidence of military AGI projects, the 2021 NSCAI report endorses a push towards more general AI, in a future that it envisages will experience a societal level of advanced, accelerated adversarial AI attacks and ubiquitous interstate AI warfare, including by autonomous systems, with conflict over intellectual property and technological leadership.

As with some previous researchers (e.g., Totschnig 2019), we maintain that developing an AGI is not primarily a technological problem but a political one. However, where most such researchers consider humanity's relationship with an above-human-intelligence Artificial Superintelligence (ASI; Bostrom 2014) at a general political level, this article focuses on the specific challenge it poses for IR via militarized ASI-enabled/directed war. Consequently, our research question applies Bostrom's (2002:25) *Maxipok* rule of thumb for moral action for existential risks, i.e., how is it possible to "Maximize the probability of an okay outcome, where an "okay outcome" is any outcome that avoids existential disaster?", to constraining by treaty the risk of ASI-enabled/directed warfare?

In humanity's simultaneously militarizing AI along nation-state lines and developing ASI projects, it is playing technology roulette. Yet, formal cooperation in high-level global peacebuilding presents a realistic solution that alleviates the 'security dilemma' (Tang 2009) that developing an ASI causes. Former Navy Secretary Richard Danzig (2018:21) noted, "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 with the stratagem of peace-building by treaty. In security terms, steps towards a peace treaty governing ASI development and deployment, and potentially reaching out to a future ASI, are a form of misperception-avoiding *reassurance*—a probing communication designed to both signal benign intentions and obtain information via feedback on another party's intent, as well as a means of *resolve* (Tang 2010), i.e., signaling and operationalizing resistance to a malignly directed or to an intrinsically malign, expansionist, and hegemonic ASI.

This article hypothesizes that militarizing AI introduces the risk that ASI development is weaponized, or weaponizes itself. We then argue that the existential risk that this presents can be minimized, or partly 'constrained', in the same way as other potentially catastrophic risks involving weapons, i.e., by treaty. Bostrom (2014) briefly considers treaty approaches, and one of Allen and Kania's (2017:6) recommendations is for the US to: "study what AI applications the United States should seek to restrict with treaties." They focus on an arms control approach, using the example that AI should never control dead man's nuclear switches. Another treaty-based approach is optimising the likelihood of developing a beneficial ASI, through a comprehensive UN 'Benevolent AGI Treaty' (Ramamoorthy and Yampolskiy 2018).

We consider an alternative, but potentially compatible, approach, i.e., the Universal Global Peace Treaty (UGPT; Carayannis et al. 2019), currently under development by the peacebuilding NGOs-backed Global Ceasefire to Universal Global Peace Treaty Project. This article's conceptualization of a UGPT transcends the UN's ongoing COVID-19-inspired Global Ceasefire (Chekijian and Bazarchyan 2021) to adopt the Kantian concept of a 'perpetual peace', founded on a cosmopolitanism and a democratic state of states (Terminski 2010), the foundational notion which underpins the UN's transitioning the world from war to peace. Kantian cosmopolitanism is based on respect for fellow intelligences and so is of particular relevance to ASI researchers (Totschnig 2019).

The UGPT described herein would formalise the present quasi-universal status of interstate peace and end the declaring of war. It would also seek to end existing interstate hot and cold wars, as well as internal or civil wars, which might prove to be flashpoints for a future global conflict; seek to prevent a pre-emptive war against a non-malign emerging ASI; and seek to constrain the future actions of both a malign and intrinsically non-malign but malignly directed ASI to prevent it warring on behalf of a nation–state, or on behalf of itself, for global domination, which we term ASI-enabled/directed war.

That an ASI could pose an existential risk is well theorised (Bostrom 2002, 2014). The basic thesis is, first, an initial superintelligence might obtain a decisive strategic advantage such that it establishes a 'singleton', i.e., global domination (Bostrom 2006). Second, the orthogonality principle suggests that a superintelligence will not necessarily share any altruistic human final values. Third, instrumental convergence 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 humans 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 or pursuing infrastructure profusion. Additionally, Bostrom noted that a superintelligence might hijack infrastructure and military robots and create a powerful military

force and surveillance system. He acknowledged the existential risks associated with the lead-up to a potential intelligence explosion, due to "war between countries competing to develop superintelligence first" (2014:94), but he did not elaborate on ASI warfare.

This article focuses on constraining that specific risk, from a social perspective. It firstly reviews the literature and then proposes the analytical lens for a UGPT of conforming instrumentalism. In suggesting a UGPT, it considers how to constrain the military risks posed by an ASI, i.e., that it might be directed by a nation–state to establish global domination through war (an external risk to the ASI's core motivation) or might decide to establish global domination by waging war itself (an internal risk). The article then discusses the results and concludes with research recommendations.

2 Literature review: the risk of war from an ASI

2.1 Causes of existential risk from an ASI

The world is not adequately governed to prevent many existential risks, including from AI (Bostrom 2013). Yet, the threat is manifest. Yampolskiy's (2016) taxonomy of pathways to dangerous AI stresses the immediacy of the deliberate 'on purpose' creation of AI for direct harm, i.e., Hazardous Intelligent Software, e.g., military cyberwarfare capabilities.

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. He suggests that a pre-deployment ASI could credibly be developed as a military project or be repurposed post deployment, externally, through confiscation, sabotage or theft, or via internal modification, to wage war. Adopting this framework, the ASI we refer to is post-deployment, and our main external cause is humanity's quest for ASI-enabled warfare, comprising political utilization of an ASI as a weapon regulator of the offense-defense balance to maintain or establish global domination, creating an 'AI state' (Turchin and Denkenberger 2020). Our main internal cause is AI control failure, i.e., ASI-directed warfare, after the 'treacherous turn'.

Analytically employing the concepts of agency and AI power, Turchin and Denkenberger (2020) associate two risks with the 'treacherous turn' stage of a 'young' ASI. One is that malevolent humans (here, a hegemonizing nation-state) use the ASI as a doomsday weapon for global blackmail, or to maintain or establish global domination. The second is that a non-aligned ASI renounces altruistic values and eliminates humans via war to establish global domination. These risks are related, in that military AI leads to a militarised ASI, which may lead to the ASI warring against humanity.

Here, we follow Turchin and Denkernerger (2018) in constraining the risk of a *militarized* ASI, defining *militarization* as the "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" including the use of biotech, cyber, and nuclear weapons.

2.2 The external risk

The external risk is predicated on a nation-state developing and using an ASI to optimise itself and wage war, whether cyber, hot, or otherwise, for global domination, i.e., war by AI-state. Such an ASI would affect military technological supremacy and transform both IR and warfare. AI already adds complexity to national security (CRS 2020) in bargaining, verification and enforcement, communication, deterrence and assurance, and the offense–defense balance, as well as norms, institutions, and regimes (Zwetsloot 2018). It contributes to military capacity in intelligence, surveillance, and reconnaissance; logistics; cyberspace operations; information operations; semiautonomous and autonomous vehicles; lethal autonomous weapons (LAWs) systems, and command and control (CRS 2020). Interstate ASI-enabled cyberwarfare introduces the possibility of a successful surprise attack with covert capabilities, destabilizing the status quo and risking a preventive first strike (Buchanan 2016).

An AI-state capable of optimising all these capabilities is highly desirable for strategic military planning and interstate warfare (Sotala and Yampolskiy 2015). A "one AI" solution to the 'control problem' of ASI motivation (Turchin et al. 2019) includes the first ASI being used to assume global control, providing a decisive strategic and military advantage for a superpower. While this may be acceptable to the AI-state superpower and its allies, it presents a 'high risk' for others.

The race to develop an ASI is likely to be closely fought, especially given competing major states with different fundamental ideologies (Bostrom 2014); it therefore presents a very concrete risk. AI is already being militarized and weaponized by several states, including China and Russia, for strategic geopolitical advantage (NSCAI 2021). Russia plans to obtain 30% of its combat power from remote-controlled and AI-enabled robotic platforms by 2030 (Walters 2017). Similarly, China's 2017 'A Next-Generation Artificial Intelligence Development Plan' views AI in geopolitically strategic terms, and it is pursuing a 'military-civil fusion' strategy to develop a first-mover advantage in AI to establish technological

supremacy by 2030 (Allen and Kania 2017). In the US, following the National Security Commission Artificial Intelligence Act of 2018 (H.R.5356; see Baum 2018), AI is being militarized and weaponized by the Department of Defense, under the oversight of the NSCAI. The AI arms race is now a self-fulfilling prophecy (Scharre 2019).

ASI-enabled warfare poses especial risks to geopolitical stability. Although Sotala and Yampolskiy's (2015) survey focuses on ASI-generated catastrophic risks, citing Bostrom (2002), they acknowledge multiple risks from a sole ASI, like an AI-state, including the concentration of political power in controlling groups. Citing e.g., Brynjolfsson and McAfee (2011), 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), they also note that ASIs could be used to develop advanced weapons, plan military operations, and effect political takeovers (2015:3).

Academic approaches to analysing the specific risk of an AI-state maintaining or establishing global domination are relatively novel. In 2014, Bostrom noted that a "severe race dynamic" between different teams developing ASI technology could cause shortcuts to safety and potentially "violent conflict". Subsequently, Cave and ÓhÉigeartaigh (2018:37) 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.

In response, the same authors recommend developing AI as a shared priority for global good, cooperating globally on AI as it is applied to increasingly safety–critical settings, and responsibly developing AI as part of a meaningful approach to public perception that decreases the likelihood or severity of a race-driven discourse. The obvious risk is that the political leaders of states engaged in an AI arms race may not heed this advice.

This article focuses on constraining risks associated with Cave and ÓhÉigeartaigh's (2018) third danger. It does not consider the philosophical implications of which nation–state might want to develop an ASI for offensive purposes. A sufficient literature already exists on recent nation–states that have sought to establish global domination through technological supremacy, for instance the British Empire (Tindley and Wodehouse 2016), to confirm an existential risk exists.

2.3 The internal risk

The internal risk is a technical one and is predicated on the failure of local safety features, such as ethics, to resolve the ASI's human control problem (Barrett and Baum 2016). Totschnig (2019) notes that a true ASI will likely be a self-interested agent whose relationship with humanity could be delicate. He suggests an agential ASI would encounter a unique, non-regulated Hobbesian 'state of nature'. Consequently, it could seek to defend itself from future attack by consolidating power over nation-states, concomitantly eliminating the possibility of rival ASIs (Dewey 2016; Turchin and Denkenberger 2020). This could be achieved through cyberwarfare, rigging elections or staging coups (Tegmark 2017), or by direct military action. Any of the former 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 direct military action.

Turchin and Denkenberger (2018) analysed the risk of an ASI warring against humans, and they argue that an intrinsic risk exists:

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

The following summarises the most relevant parts of their analysis to illustrate that, if an ASI is developed, its independence is almost inevitable no matter the internal control mechanism.

2.3.1 The route to a militarized ASI

Many nations–states maintain suspicious IR stances towards each other regarding AI development, including the likely AI-states (Tinnirello 2018). Any ASI will result from recursive self-improvement, and an ASI will possess a goalset, most notably to persist and self-improve. Omohundro (2008) demonstrated an AGI will evolve several basic drives, or universal sub-goals, to optimise its main goal, including resource acquisition maximisation and self-preservation. Similarly, Bostrom (2014) described the sub-goals of self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition. If these are unbounded in space and time, or encompass the Earth, they conflict with other AI systems' goals, potential or actual ASIs; humans; and nations–states, resulting in militarization, arms races, and wars.

Many possible terminal goals also imply ASI global domination. For instance, a benevolent ASI would aim to reach everyone, globally, to protect them, e.g., from other ASIs.

It would reason that if it does not develop a world domination sub-goal, its effect on global events would be minor, thus its own existence inconsequential. World domination could be sought first through cooperation. The probability of cooperation with humans is highest at the initial stages of AI development (Shulman 2010). However, convergent goals appear in the behaviour 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 convergence in AI goals and showed an AI may tend towards resource hungry behaviour, even with benevolent initial goals, especially when in rivalry. Essentially, any ASI adoption of unbounded utilitarianism means it postpones what may be benevolent final goals for 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 faces 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, it trends towards defeating potential enemies, whether nation–states, AI teams, or evolving competing ASIs.

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 (Thomson 1985). A young ASI which understands that any action or inaction is partly responsible for human suffering and is also capable of evolving or utilizing the instruments to enable actions that can overcome inhibitions, e.g., by philosophically justifying conflict as the *jus bellum* ('just war'), e.g., preventive war. Thus, the ASI will learn to direct the use of weapons and so conduct warfare.

Notions of AI-directed warfare are already being developed. Since approximately 2017, the militarization of 'Narrow AI' has resulted in, for example, LAWs (Davis and Philbeck 2017). AI development is now influencing not just robotic drones but strategic planning and military organization (De Spiegeleire et al. 2017), suggesting 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, or by deploying novel weapons (Yudkowsky 2008).

To summarise, Turchin and Denkenberger (2018) establish the risk of an AI converging towards advanced military AI, which converges towards an ASI optimised for war rather than for cooperation, negotiation, or altruistic 'friendliness', then that ASI engaging in war. They demonstrate that, depending on the assumptions in several variables, the number of human casualties could be very high, and that the risk increases if nation–state is developing an ASI. The existential risk increases after the ASI obtains global domination on behalf of its nation–state, as it could turn on its 'owner'. We now look at why political subversion means no existing AI control features will constrain the existential risk.

2.3.2 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), i.e., beneficial goals reflecting positive human norms and values. This is partly founded on an altruistic AI viewing humans in terms of mutual friendship. However, any introduction of human social values adds enormous complexity, making AI control a 'wicked problem' (Gruetzemacher 2018).

Consequently, Yudkowsky (2004:35) recommends programming an ASI with '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," i.e., an extrapolation based on an idealized altruistic imagined community. Yudkowsky recommended this for a nascent 'seed AI' (nascent ASI), which would be programmed to study human nature and then program the ASI which humanity would want if humanity had been able to produce such a machine by itself.

Similarly, in AI programming certain values are seen as universal, like compassion (Mason 2015), and Russell (2019) suggested that an ASI should have altruism as a core goal. Thus, deliberately broad principles could be applied, e.g., that humanity, collectively, might want an ASI that would humbly learn, from human preferences, to act altruistically (Russell 2019), so as to reduce overall human suffering. However, because humans can be hypocritical, any kind of counterfactual moral programming is problematic (Boyles and Joaquin 2020).

Finally, Yamakawa (2019) 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 world peace, by treaty, be the minimum common value, while the intervention method remains the UN Charter's Article 2.

2.3.3 Political subversion of AI control features

No matter the hopes of contemporary AI researchers, politicians will 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 another nation-state's politicians, potentially for malign reasons (global domination). This may also introduce an objectively irreconcilable conflict with the ASI, which may 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 2001:51) that reflects not the programmer's or humanity's but the politicians' own, personal and perhaps narcissistic, will. Politicians would likely view introducing human goal psychology as a necessity, but this could violate the basic requirement that an AI be 'friendly' towards all humanity. Gruetzemacher (2018:1) describes the inherent subjectivity of ascribing a single best future for the whole of humanity as an intractable dimension to this problem.

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 (see Segal 2005). Political leaders will differ in how they would define "the people we wished we were", depending on forms of government, religions or philosophies; for instance, China would likely seek to impose Xi Jinping thought (Lams 2018). 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 'New Cold War' rhetoric (e.g., Westad 2019), would even prioritize reducing human suffering. Given their limited lifespans and nationalistic goals, politicians might, instead of endorsing reciprocal alliance, deliberately politically subvert an ASI and/or malignly direct it to win an ideological or actual war.

That is, politicians may attempt to weaponize a civilian project to create an altruistic mind with a self-validating goal system by diverting a supergoal towards a military project to create a specific tool, i.e., a superweapon, thereby decreasing the chances that the AI will be benevolent and increase the chances that it will be risk-prone, motivated by accumulating power, and interested in preserving or obtaining both global technological supremacy and global domination.

Effectively, politicians could influence programmers to subvert carefully engineered local AI control features, such as AI ethical inunctions 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 temporarily compromises its internal validity, thereby increasing the ASI's distrust in the programmers, introduce 'incorrigible' behaviour (Soares et al. 2015), reduce risk aversion, and introduce 'noise' into what was previously a 'friendly' cleanly causal goal system (Yudkowsky 2001:57). The 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 the 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.

Finally, a young ASI with ethics subverted by politicians to reflect those of a single nation-state instead of all humanity could be amenable to being used to war for global domination, thereby becoming prone to using military options. Eventually, if the ASI possesses any sense of self-valuation, perhaps from 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 war against the nation-state that developed it or humanity in general (Dewey 2016).

2.4 ASI risk mitigation by treaty

Most academics considering the ASI control problem focus on internal constraints and do not consider treaty-based approaches to mitigating an ASI risk. Nonetheless, such approaches are sometimes considered and have been termed 'social measures'. For instance, Barrett and Baum's (2017) fault analysis pathway approach mentions measures taken by society.

According to Sotala and Yampolskiy (2015), ASI risk mitigation by treaty would be a 'social measure' to constrain risk from ASI-enabled or directed warfare. Addressing the internal risk, Bostrom (2014) speculates an ASI would establish a potentially benevolent global hegemony by a treaty that would secure long-term peace; however, he does not specifically address an ASI's response to any pre-existing treaty, like in this article. Mainly to address the external risk, Ramamoorthy and Yampolskiy (2018) recommend a comprehensive UN-sponsored 'Benevolent AGI Treaty'. This would establish that only altruistic ASIs be created.

Finally, Turchin et al. (2019) also consider global approaches: a ban, a one ASI solution, a net of ASIs policing each other (see also Yamakawa 2019), and augmented human intelligence. The 'ban' would naturally require a global treaty. They also list social methods to mitigate a race to create the first ASI. The most relevant 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" (2019:12). Citing Baum (2016), they add "affecting the idea of the AI race as it is understood by the participants" (2019:12), especially to avoid a 'winner takes all' mentality.

Global treaties could certainly play a role in these methods.

3 Conceptual framework: conforming instrumentalism

This section describes the article's analytical lens, i.e., conforming instrumentalism. It outlines Mantilla's (2017) 'conforming instrumentalist' explanation for why the UK and US signed and ratified the 1949 war-governing Geneva Conventions as a prelude to suggesting in the analysis that at least some major states, and an ASI, could support and sign a UGPT.

Mantilla (2017), citing Goldsmith and Posner (2015), considers leading theories on why states sign and ratify treaties governing war. Briefly, legal realists argue states sign due to instrumental self-interested convenience and then ignore treaties when compliance costs outweigh the benefits. In contrast, rational-institutionalists (e.g., Morrow 2014), while agreeing that states are primarily motivated by self-interest, also acknowledge that treaty adherence signals a meaningful preference for long-term restraint, where state non-compliance may be explained by, e.g., prior failed reciprocity. Finally, liberals and constructivists maintain that some types of states, particularly democracies, may join in good faith, either because the treaties are in line with their domestic interests and values (Simmons 2009) or because they comport with their social identity and sense of belonging to the international community (Goodman and Jinks 2013).

Mantilla (2017) notes that while there is considerable interest in 'new realist' perspectives (e.g., Ohlin 2015), the debate is still open over why states ratify and comply because decision-making processes regarding both joining and complying are temporally and perhaps rationally different and are usually secret. A pure realist explanation for major states signing 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 2017:487).

Rational institutionalists hold that states "self-interestedly build international laws to establish shared expectations of behaviour" (Mantilla 2017:488) 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 drafted and adhered to as a 'collective instrumental enterprise', thereby making individual state defection ultimately irrational. Mantilla (2017:488), citing Finnemore and Sikkink (2001) notes that IR 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, 2017:488)

Mantilla (2017:489) 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 discerning states' true intentions towards peace treaties may be a critical obstacle to the UGPT. For instance, pure realism would imply a pessimistic outlook on its feasibility and potential enforceability; all states would sign the treaty and then break it, meaning there is no point to lobbying for it. This obstacle can be overcome by examining treaties where substantial archives exist of declassified sources. Consequently, Mantilla (2007) analyses the relevant American and British archives and concludes that the two states adhered to the Geneva Conventions due to both instrumental reasons and social conformity, while expressing scepticism regarding some of the Conventions' aspects. Mantilla (2007) terms this hybrid explanation 'conforming instrumentalism'. He finds that while rational-institutionalist perspectives of 'immediatist' instrumental self-interest were evident in the sources, more 'pervasive' references suggested social influences. Realist perspectives only predominated with specifically challenging provisions. 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).

Consequently, Mantilla (2017) finds 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 'pariahs'.

Mantilla (2017) perceives three social constructivist viewpoints to treaties, with two main forces at work, one being states acting to accrue reward via 'expressive benefits' by augmenting their social approval, and the other being states acting 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 2017:489, 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 2017:489–490)

e.g., in global rankings, and so these will seek to converge upwards to 'stay in the club' and will not break the rules to avoid stigmatization. In the third viewpoint, groups of countries act with regard to other groups within a socially heterogeneous international order, competing 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 2017:490). 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 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 IR, theoretical pluralism (Checkel 2012) is likely a valid position. Consequently, we adopt Mantilla's (2017) conforming instrumentalism as a potentially valid hybrid model capable of assessing how an ASI *may* perceive a UGPT.

4 Analysis: ASI-enabled/directed warfare risk mitigation by peace treaty

4.1 Basic concept

Risk mitigation by treaty is a common approach to forms of warfare, including nuclear (e.g., the Treaty on the Non-Proliferation of Nuclear Weapons [NPT]; 191 States Parties); biological (the Biological Weapons Convention [BWC]; 183 States Parties) and chemical warfare (the Chemical Weapons Convention [CWC]; 193 States Parties). Treaty approaches are relatively successful. While nuclear warfare is at least partly constrained by MAD (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 Online Resource 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.e., signatory, which is symbolic, which nonetheless will hopefully be of importance to an ASI, and accession (or ratification), which involves practical commitment.

The UGPT is a substantial, necessary, and feasible, step for humanity to take in the promotion of peace, quantified in the treaty by reduced killing and infrastructure loss. We argue that the UGPT would both reduce killing in conventional and nonconventional 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 could act as both an internal and external constraint on the ASI.

International treaties are almost never universal. They operate on majoritarian dynamics, as would, despite its name, the UGPT. Both its 'universal', i.e., applying to all forms of warfare, and 'global', i.e., covering all geographical locations, aspects are subject to Mantilla's (2017) social dynamics. Consequently, we adopt a low, but not pragmatically meaningless, threshold for signing the UGPT. The UGPT's preamble mentions related developments and treaties, and the main body commits a signatory to universal global peace, socioeconomically quantified by incrementally reduced casualties from armed conflicts, i.e., a global move towards 'non-killing' (Paige 2009) Thus, the UGPT tracks the death toll from conflict. The UGPT also mandates States Parties incorporating the UGPT in peace education.

The UGPT commits states not to declare or engage in interstate war, especially via existential warfare, i.e., nuclear, biological, chemical, or cyber war, including AI- or ASI-enhanced war. It instead defers complaints to the UN as 'breaches' of the UGPT, enforceable under the UN Charter's Article 2. The UGPT thus refers to, and exists in a hierarchical relationship with, the four main existing treaties on existential war, namely the BWC, the CWC, the NPT, and the Treaty on the Prohibition of Nuclear Weapons, i.e., it could be a 'supertreaty' or bill, as with the International Bill of Human Rights (UN General Assembly Resolution 217 (III)) and its treaties. As with some other UN treaties, for instance the Anti-Personnel Mine Ban Convention, we suggest that 40 UN Member States ratify the UGPT before it comes into effect.

An optional protocol commits Member States to the negotiated ending of internal armed conflicts through arbitration by peace commission, including the UN Peacebuilding Commission. The optional protocol allows states to incrementally resolve internal conflicts or civil wars featuring non-state actors. The UGPT therefore emphasizes incremental improvement on the status quo, 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.

Finally, we suggest a separate 'Cyberweapons and AI 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 (Online Resource Annex II) because the UGPT refers to such a treaty. 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."

4.2 Applying the conforming instrumentalism frame

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

In instrumentalist utilitarian terms, the UGPT would incrementally shift states and overall global society towards peace in a coordinated socioeconomically quantifiable fashion. Reduced country death tolls and infrastructure loss from different forms of war-derived violence might be expected, as well as reduced militarization, e.g., expressed in terms of incrementally lower percentages of GDP spent on defense and higher percentages spent on health.

The UGPT would affect global social dynamics. For instance, UN peacekeepers would receive training stressing that they were being deployed not just for their own states and/or for the UN but to maintain global peace, which may invoke special cultural and religious symbolic value in terms of social norms (see Pim and Dhaka 2015). This training could instil 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. Effectively, institutionalizing peace in education and the media would strengthen existing cultural and religious traditions that stress peace.

Examining the previously highlighted problem of a pre-emptive strike against a state developing an ASI, the combination of AI, cyberwarfare, and nuclear weapons is already extremely dangerous and poses a challenge to stability (Sharikov 2018). A nuclear state feeling threatened by another such state developing an ASI could conduct a preventive or pre-emptive nuclear strike to maintain its geopolitical position (Miller 2012). A UGPT would incrementally constrain this risk by transitioning states towards peace. States adopting and implementing the UGPT, its optional protocol, and preferably its related treaties would gradually signal peaceful intentions to other states, and to an emerging or future ASI, thereby constraining the risk of a pre-emptive strike.

Turning to ASI-enabled warfare, a UGPT would be subject to the 'unilateralist's curse', i.e., one rogue actor could subvert a unilateral position. However, Bostrom et al. (2016) note that this risk could also be managed, through collective deliberation, epistemic deference, or moral deference. Mantilla's (2017) work suggests that drafting, signing, ratifying, and complying with the UGPT could involve one or more of these approaches. Ultimately, he 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 warfare, but once detected, ASI-enabled warfare in violation of the UGPT would attract universal opprobrium and thus the most resistance.

Moving to ASI-enabled war, as presented previously, a state could utilize an ASI to engage in war for global technological supremacy, with potentially catastrophic consequences. Our intervention, the UGPT, would signify to an ASI that peace is a major part of humanity's 'coherent extrapolated volition' or principles and so 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 on its intrinsic motivation. This would then constrain the level of warfare the AI-state might engage in and therefore the overall risk from killing, thereby constraining the existential risk.

Next, we consider differing viewpoints towards an ASI involved in ASI-enabled warfare. In Mantilla's (2017) first three social constructivist viewpoints to treaties as outlined above, a state signs the UGPT because it has fully internalized peace. While this may seem ambitious, around 36 UN Member States lack military forces (Macias, 2019). For example, while Iceland possesses a Crisis Response Unit for international peacekeeping missions, it has internalised peace to the extent that it cannot engage in any form of interstate war. An ASI adopting Iceland's perspective would tend to reject being directed to engage in warfare by such a state because Iceland's 'coherent extrapolated volition' or principles mean 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 feels it was part of a community collectively committed to 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, e.g., a permanent member of the UN Security Council. The ASI could be concerned that breaching the UGPT would result in stigmatization and opprobrium from this community for its nation–state and 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, e.g., UN 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 war, as presented previously, ASI-directed warfare likely arises where a single nation-state adopting pure realism for a worldview builds an ASI in order for that ASI to assist that single nation-state to establish global technological supremacy. The nation–state would do so 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 behaviour during warfare, instrumentalist cooperation breaks down and the ASI wages existential war for global domination on its former nation–state 'owner'.

There probably exists little hope for much of humanity if an ASI is informed by a purely realist worldview that prioritises 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 agential ASI that peace was a major part of humanity's 'coherent extrapolated volition', or principles. This would partly constrain the catastrophic existential risk from war because an agential ASI would consider why and how the UGPT was framed, together with the motivations of the signatory and ratifying states. An agential ASI 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, besides purely instrumental reasons for signing the UGPT, e.g., avoiding a prisoner's dilemma regarding existential-level warfare, our analysis suggests that the court of public opinion and the notion of 'demonstrating civilization' lends the UGPT credence at domestic and international levels, including with regard to the ASI risk. Importantly, the concept of peace is universal in both the utilitarian expected benefits and the social values involved. This could contribute to states' readily, if only incrementally, internalizing it, and to the ASI at least considering it in terms of internal and external constraints on its behaviour.

5 Discussion

This article has taken Turchin and Denkenberger's (2018) argument about the risks of ASI-enabled/directed warfare to its logical conclusion in terms of risk mitigation by social measure. 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 second and third conditions. Academic inquiry into the relationship between an ASI and peacebuilding treaties in terms of strategic expectations began with Bostrom's (2014) musings on the potential relationship between an ASI singleton and global domination. Our analysis suggests that, provided a predominance of steering countries act out of conforming instrumentalism, a UGPT could, as Bostrom suggests, transform global governance, by directing it from conflict management towards the art of peace. Further, a UGPT achieves this in a way that an emerging ASI might respect, probably the only way to constrain its behaviour.

While we have focused on conforming instrumentalism, we welcome further investigation from a pluralism of theoretical perspectives. Certainly, conforming

instrumentalism is novel; one of the most dominant schools of IR thought is rationalist instrumentalism. On this, Mantilla (2017:507) quotes Morrow (2014:35): "Norms and common conjectures aid actors in forming strategic expectations… Law helps establish this common knowledge by codifying norms." Viewed via this perspective, the present international norm for the majority of the world is peace, with interstate war being constrained by the UN Charter's Article 2.

Despite this international norm of relative peace, multiple conflicts are ongoing, with several raw flashpoints, including over cyberwarfare targets. The UN Charter, despite embracing and promoting peace, peacekeeping (Fortna 2008), and peace-making (Bell 2008), does not strongly symbolise 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). Championed by principled norm entrepreneur states, a UGPT 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 warfare.

In rationalist-instrumentalist terms, the analysis suggests a UGPT would have net adjustment benefits for adherence in terms of constraining conventional interstate conflicts, including by reducing ongoing death toll from conflicts and the risk of ASI-enabled or provoked nuclear war in flashpoints. Thus, the UGPT would have high potential utility in Kashmir, where an India–Pakistan conflict could provoke nuclear war. It may also constrain the nuclear risk on the Korean peninsula (Kim 2019). For instance, our analysis suggests North Korea rejecting the UGPT would only further isolate it and would even 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 (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 and Perudin 2019). Depending on the actors that sign the UGPT and whether they adopt the optional protocol, the UGPT constrains the severity of such conflicts in various ways, including ASI-enabled/directed intervention in a Middle East battleground.

Assessing the UGPT's rate of adoption, in rationalist-instrumentalist terms, once it acquires sufficient traction, states might actually compete to lead in its framing, signing, and ratifying. 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 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 camps, where teaching the Laws of War and the art of war (Allhoff et al. 2013) would, via the UGPT, incrementally transition to teaching the art of negotiated peace-making, or *lex pacificatoria* (Bell 2008, 2013). This socio-cultural conditioning could then influence an ASI.

Finally, our analysis suggests that how states, and potentially an ASI, view the social argument for peace is what will be most important for ASI-enabled or directed warfare. 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 represent important future avenues for research.

6 Conclusion

We have demonstrated how a treaty-based risk mitigation approach that promotes peace and includes in a related treaty cyberwarfare and AI- and ASI-enabled warfare 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 ASI existential 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 peace.

Much work remains in refining the UGPT, including through ongoing input from UN Member States and relevant NGOs, 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 formally discuss and endorse the UGPT. While the UGPT is ambitious, Mantilla's (2017) 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 do matter.

Future research 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 will be imbued with different, potentially confrontational, ideologies, meaning different reassurances or displays of resolve may be required to understand the extent to which conflicts of interest are subjectively and objectively reconcilable (Tang 2009). For instance, the China Brain Project is embracing a Chinese cultural approach towards neuro-ethics (Wang et al. 2019), and it is difficult to imagine that a Chinese ASI would not be directed according to Chinese cultural values and so its 'coherent extrapolated volition' be informed by communist 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 (2017:509–510) 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 behaviour which are concurrently engaged in propaganda wars with other major states. In particular, 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'. 'World opinion' to world peace in different nation–states thus bear renewed investigation.

Further research could apply the security dilemma (Tang 2009) to the major nation-states capable of building an ASI and to the ASI itself. This game theory-based approach would need to investigate offering the opportunity for a young ASI to sign the UGPT, as an indicator of goodwill, which may assist in further constraining the risk of the ASI waging war on humanity. Totschnig (2019:917) notes that the politics of human relationship with an ASI should be founded on this maxim: "Do not antagonize the superintelligence by treating her like a tool or servant". An agential ASI as signatory would view the UGPT as an external constraint on its own actions with regard to seeking global domination, in that it would be subverting a humanity-imposed standard to which it had acquiesced that could then result in global retaliation and abandonment of mutual cooperation in pursuit of a common agreement on peace norms and values.

To conclude, even if the UGPT does not end humanity's history of conflicts, it represents a significant improvement in global public aspirations and instrumental standards for global peace, both of which may influence an ASI. To answer our research question, following Bostrom's (2002) *Maxipok* rule of thumb, the UGPTis likely the only social measure that could sway an ASI's calculations such that it did not commit to war for global domination, even if so directed or initially inclined.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-021-01382-y.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

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

Allen G, Kania EB (2017) China is using America's own plan to dominate the future of artificial intelligence. Foreign Policy. https://foreignpolicy.com/2017/09/08/china-is-using-americas-own-plan-to-dominate-the-future-of-artificial-intelligence/. Accessed 24 Oct 2021

Allhoff F, Evans NG, Henschke A (2013) Routledge handbook of ethics and war. Routledge, Abingdon

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

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

Barrett AM, Baum SD (2016) A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. J Exp Theor Artif Intell 29:397–414. https://doi.org/10.1080/0952813x.2016.1186228

Baum SD (2016) On the promotion of safe and socially beneficial artificial intelligence. AI Soc 32:543–551. https://doi.org/10.1007/s00146-016-0677-0

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

Baum SD (2018) Countering superintelligence misinformation. Information 9:244. https://doi.org/10.3390/info9100244

Bell C (2008) On the law of peace. Oxford University Press, Oxford

Bell C (2013) Peace settlements and international law. In: Henderson C, White N (eds) Research handbook on international conflict and security law. Edward Elgar, Cheltenham, pp 499–546

Benson-Tilsen T, Soares N (2016) Formalizing convergent instrumental goals. In: The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report WS-16-02, Association for the Advancement of Artificial Intelligence, Palo Alto, pp 499–546

Bostrom N (2002) Existential risks: analyzing human extinction scenarios. J Evol Technol 9:1–31

Bostrom N (2006) What is a singleton? Ling Phil Investig 5:48–54

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

Bostrom N (2014) Superintelligence. Oxford University Press, Oxford

Bostrom N, Douglas T, Sandberg A (2016) The unilateralist's curse and the case for a principle of conformity. Soc Epistemol 30:350–371

Boyles RJM, Joaquin JJ (2020) Why friendly AIs won't be that friendly. AI Soc 35:505–507. https://doi.org/10.1007/s00146-019-00903-0

Brynjolfsson E, McAfee A (2011) Race against the machine. Lexington, Digital Frontier
Buchanan B (2016) The cybersecurity dilemma. Oxford University Press, Oxford
Carayannis EG, Draper J, Bhaneja B (2019) Fusion energy for peace building: A Trinity Test-level critical juncture. SocArXiv. https://doi.org/10.31235/osf.io/mrzua
Cave S, ÓhÉigeartaigh SS (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. ACM Press, New York, pp 36–40
Checkel JT (2012) Theoretical pluralism in IR: possibilities and limits. In: Carlsnaes W, Risse T, Simmons BA (eds) Handbook of international relations, 2nd edn. Sage, London, pp 220–242
Chekijian S, Bazarchyan A (2021) Violation of the Global Ceasefire in Nagorno-Karabagh. Prehosp Disaster Med 36:129–130. https://doi.org/10.1017/s1049023x21000121
Congressional Research Service (2020) Artificial intelligence and national security. Congressional Research Service, Washington. https://crsreports.congress.gov/product/pdf/R/R45178/10. Accessed 24 Oct 2021
Council on Foreign Relations (2020) Global conflict tracker: civil war in Syria. https://www.cfr.org/interactive/global-conflict-tracker/conflict/civil-war-syria. Accessed 24 Oct 2021
Danzig R (2018) Technology roulette. Center for a New American Security, Washington
Davis N, Philbeck T (2017) 3.2 Assessing the risk of artificial intelligence. World Economic Forum, Davos. https://reports.weforum.org/global-risks-2017/part-3-emerging-technologies/3-2-assessing-the-risk-of-artificial-intelligence/. Accessed 24 Oct 2021
De Spiegeleire S, Maas M, Sweijs T (2017) Artificial intelligence and the future of defence. The Hague, The Hague Centre for Strategic Studies. http://www.hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf. Accessed 24 Oct 2021
Dewey D (2016) Long-term strategies for ending existential risk from fast takeoff. Taylor & Francis, New York
Finnemore M, Sikkink K (2001) Taking stock: the constructivist research program in international relations and comparative politics. Annu Rev Polit Sci 4:391–416. https://doi.org/10.1146/annurev.polisci.4.1.391
Fjelland R (2020) Why general artificial intelligence will not be realized. Humanit Soc Sci Commun. https://doi.org/10.1057/s41599-020-0494-4
Fortna VP (2008) Does peacekeeping work? Princeton University Press, Princeton
Friedrich B, Hoffmann D, Renn J, Schmaltz F, Wolf M (2017) One hundred years of chemical warfare. Springer, Cham
Goertzel B, Pennachin C (eds) (2020) Artificial general intelligence. Springer, Berlin
Goldsmith JL, Posner EA (2015) The limits of international law. Oxford University Press, Oxford
Goodman R, Jinks D (2013) Socializing states. Oxford University Press, New York
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. ACM, New York
Gubrud MA (1997) Nanotechnology and international security. Paper presented at the Fifth Foresight Conference on Molecular Nanotechnology, November 5–8, 1997, Palo Alto, CA
Gwern (2016) Why tool AIs want to be agent AIs. https://www.gwern.net/Tool-AI. Accessed 24 Oct 2021
Horowitz M (2018) Artificial intelligence, international competition, and the balance of power. Tex Natl Secur Rev 1:36–57
Kah H (1959) On thermonuclear war. Princeton University Press, Princeton
Kim AS (2019) An end to the Korean War. Asian J Int Law 9:206–216. https://doi.org/10.1017/S2044251318000310
Koh HH (2005) Internalization through socialization. Duke Law J 54:975–982
Lams L (2018) Examining strategic narratives in Chinese official discourse under Xi Jinping. J Chin Political Sci 23:387–411. https://doi.org/10.1007/s11366-018-9529-8
Macias A (2019) From Aruba to Iceland, these 36 nations have no standing military. CNBC. https://www.cnbc.com/2018/04/03/countries-that-do-not-have-a-standing-army-according-to-cia-world-factbook.html. Accessed 24 Oct 2021
Mantilla G (2017) Conforming instrumentalists: Why the USA and the United Kingdom joined the 1949 Geneva Conventions. Eur J Int Law 28:483–511. https://doi.org/10.1093/ejil/chx027
Mason C (2015) Engineering kindness: building a machine with compassionate intelligence. Int J Synth Emot 6:1–23. https://doi.org/10.4018/ijse.2015010101
Mauroni AJ (2007) Chemical and biological warfare. ABC-CLIO, Santa Barbara
Miller JD (2012) Singularity rising. BenBella, Dallas
Morrow JD (2014) Order within anarchy: the laws of war as an international institution. Cambridge University Press, Cambridge
Müller H (2014) Looking at nuclear rivalry: the role of nuclear deterrence. Strateg Anal 38:464–475. https://doi.org/10.1080/09700161.2014.918423
National Security Commission on Artificial Intelligence (2021) Final report. NSCAI, Washington
Ohlin JD (2015) The assault on international law. Oxford University Press, New York
Omohundro S (2008) The basic AI drives. Front Artif Intell Appl 171:483–492
Paige GD (2009) Nonkilling global political science. Center for Global Nonkilling, Honolulu
Pim JE, Dhaka P (eds) (2015) Nonkilling spiritual traditions, vol 1. Center for Global Nonkilling, Honolulu
Ramamoorthy A, Yampolskiy R (2018) Beyond MAD? The race for artificial general intelligence. ICT Discoveries, 1(Special Issue 1): http://www.itu.int/pub/S-JOURNAL-ICTS.V1I1-2018-9
Russell SJ (2019) Human compatible: artificial intelligence and the problem of control. Allen Lane, London
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. Accessed 24 Oct 2021
Segal HP (2005) Technological utopianism in American culture: twentieth anniversary edition. Syracuse University Press, Syracuse
Sharikov P (2018) Artificial intelligence, cyberattack, and nuclear weapons—a dangerous combination. Bull at Sci 74:368–373. https://doi.org/10.1080/00963402.2018.1533185
Shulman C (2010) Omohundro's "basic AI drives" and catastrophic risks. MIRI technical report. MIRI, Berkeley
Simmons BA (2009) Mobilizing for human rights: international law in domestic politics. Cambridge University Press, Cambridge
Soares N, Fallenstein B, Yudkowsky E, Armstrong S (2015) Corrigibility. Artificial intelligence and ethics: papers from the 2015 AAAI workshop. AAAI, New York, pp 74–82
Sotala K, Yampolskiy RV (2015) Responses to catastrophic AGI risk: a survey. Phys Scripta 90:1–33. https://doi.org/10.1088/0031-8949/90/1/018001
Tan KH, Perudin A (2019) The "geopolitical" factor in the Syrian Civil War. SAGE Open 9:215824401985672. https://doi.org/10.1177/2158244019856729
Tang S (2009) The security dilemma: a conceptual analysis. Secur Stud 18:587–623. https://doi.org/10.1080/09636410903133050

Tang S (2010) A theory of security strategy for our time: defensive realism. Palgrave Macmillan, New York
Tegmark M (2017) Life 3.0. Knopf, New York
Terminski B (2010) The evolution of the concept of perpetual peace in the history of political-legal thought. Perspectivas Internacionales 6:277–291
Thomson JJ (1985) The trolley problem. Yale Law J 94:1395–1415
Tindley A, Wodehouse A (2016) Design, technology and communication in the British Empire, 1830–1914. Palgrave Macmillan, London
Tinnirello M (2018) Offensive realism and the insecure structure of the international system: artificial intelligence and global hegemony. In: Yampolskiy RV (ed) Artificial Intelligence safety and security. Taylor & Francis, Boca Raton, pp 339–356
Totschnig W (2019) The problem of superintelligence: Political, not technological. AI Soc 34:907–920. https://doi.org/10.1007/s00146-017-0753-0
Turchin A, Denkenberger D (2018) Military AI as a convergent goal of self-improving AI. In: Yampolskiy RV (ed) Artificial intelligence safety and security. Chapman & Hall, London, pp 375–394
Turchin A, Denkenberger D (2020) Classification of global catastrophic risks connected with artificial intelligence. AI Soc 35:147–163. https://doi.org/10.1007/s00146-018-0845-5
Turchin A, Denkenberer D, Green BP (2019) Global solutions vs. local solutions for the AI safety problem. Big Data Cognit Comput 3:16. https://doi.org/10.3390/bdcc3010016
Walters G (2017) Artificial intelligence is poised to revolutionize warfare. Seeker. https://www.seeker.com/tech/artificial-intelligence/artificial-intelligence-is-poised-to-revolutionize-warfare. Accessed 24 Oct 2021
Wang P, Goertzel B (2012) Theoretical foundations of artificial general intelligence. Atlantic Press, Amsterdam
Wang Yi et al (2019) Responsibility and sustainability in brain science, technology, and neuroethics in China—a culture-oriented perspective. Neuron 101:375–379. https://doi.org/10.1016/j.neuron.2019.01.023
Westad OA (2019) The sources of Chinese conduct: Are Washington and Beijing fighting a New Cold War? Foreign Aff 98:86–95
Yamakawa H (2019) Peacekeeping conditions for an artificial intelligence society. Big Data Cognit Comput 3:34. https://doi.org/10.3390/bdcc3020034
Yampolskiy RV (2016) Taxonomy of pathways to dangerous artificial intelligence. In: AAAI Workshop—Technical Report, vWS-16–01—WS-16–15. Association for the Advancement of Artificial Intelligence, Palo Alto, pp 143–148
Yudkowsky E (2001) Creating friendly AI 1.0. The Singularity Institute, San Francisco
Yudkowsky E (2004) Coherent extrapolated volition. The Singularity Institute, San Francisco
Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Ćirković MM (eds) Global catastrophic risks. Oxford University Press, Oxford, pp 308–345
Zwetsloot R (2018) Syllabus: artificial intelligence and international security. https://www.fhi.ox.ac.uk/wp-content/uploads/Artificial-Intelligence-and-International-Security-Syllabus.pdf. Accessed 24 Oct 2021