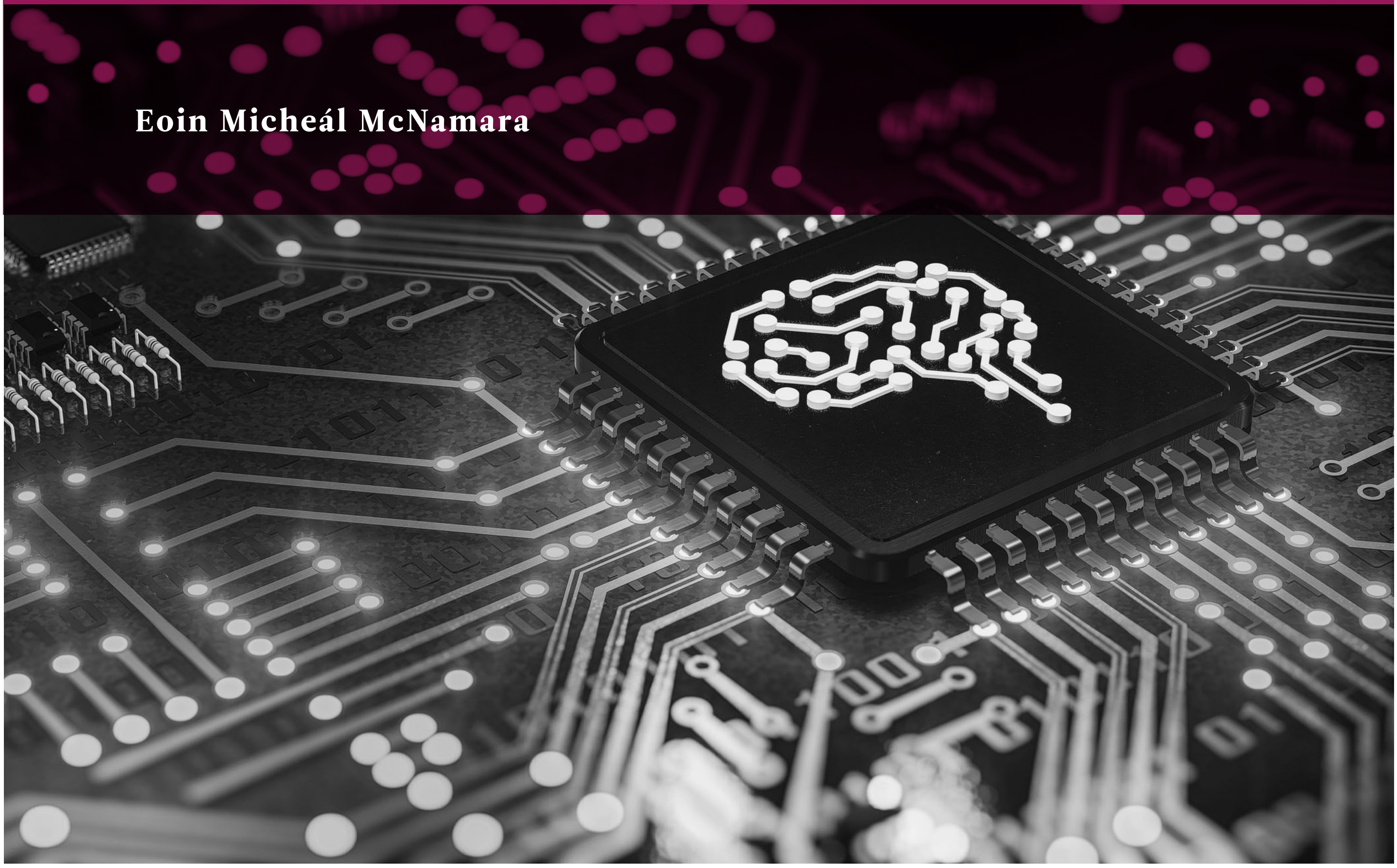
FIIA
BRIEFING PAPER

JANUARY 2024 / 381

# NUCLEAR ARMS CONTROL POLICIES AND SAFETY IN ARTIFICIAL INTELLIGENCE

## TRANSFERABLE LESSONS OR FALSE EQUIVALENCE?

Eoin Micheál McNamara



FINNISH
INSTITUTE
OF INTERNATIONAL
AFFAIRS

# NUCLEAR ARMS CONTROL POLICIES AND SAFETY IN ARTIFICIAL INTELLIGENCE

## TRANSFERABLE LESSONS OR FALSE EQUIVALENCE?

- Multiple nuclear arms control treaties have collapsed in recent years, but analogies associated with them have returned as possible inspiration to manage risks stemming from artificial intelligence (AI) advancement.

- Some welcome nuclear arms control analogies as an important aid to understanding strategic competition in AI, while others see them as an irrelevant distraction, weakening the focus on new frameworks to manage AI's unique and unprecedented aspects.

- The focus of this debate is sometimes too narrow or overly selective when a wider examination of arms control geopolitics can identify both irrelevant and valuable parallels to assist global security governance for AI.

- Great power leaders frequently equate AI advancement with arms racing, reasoning that powers lagging behind will soon see their great power status weakened. This logic serves to intensify competition, risking a spiral into more unsafe AI practices.

- The global norm institutionalization that established nuclear taboos can also stigmatize unethical AI practices. Emphasizing reciprocal risk reduction offers pragmatic starting points for great power management of AI safety.

**EOIN MICHEÁL McNAMARA**

*Research Fellow*

*Global Security and Governance*

*Finnish Institute of International Affairs*

*ISBN 978-951-769-790-3*
*ISSN 1795-8059*
*Language editing: Lynn Nikkanen*
*Cover photo: iStock*

*This research is part of the Reignite Multilateralism via Technology (REMIT) project, funded by the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101094228.*

FINNISH
INSTITUTE
OF INTERNATIONAL
AFFAIRS

Arkadiankatu 23 b
POB 425 / 00101 Helsinki
Telephone +358 (0)9 432 7000
Fax +358 (0)9 432 7799

www.fiia.fi

*The Finnish Institute of International Affairs is an independent research institute that produces high-level research to support political decision-making as well as scientific and public debate both nationally and internationally.*
*All manuscripts are reviewed by at least two other experts in the field to ensure the high quality of the publications. In addition, publications undergo professional language checking and editing. The responsibility for the views expressed ultimately rests with the authors.*

# NUCLEAR ARMS CONTROL POLICIES AND SAFETY IN ARTIFICIAL INTELLIGENCE

## TRANSFERABLE LESSONS OR FALSE EQUIVALENCE?

### INTRODUCTION

Technological advancement in artificial intelligence (AI) is rapidly accelerating, promising immense benefits to support social development. However, without strong safeguards, serious risks to human society are also envisaged.

Advanced AI will improve a wide range of policies, from healthcare to green energy, but it also risks bringing with it a flood of automated disinformation; distorted information selection that polarizes society; and an upswing in global unemployment. While a contested vision, the most ominous risk discussed is that unsafe AI will eventually manipulate humans, casting doubt on our ability to maintain complete control and posing a threat to the very survival of human society.

As in the case of AI today, nuclear weapons emerged rapidly in the 1940s, heralding new dangers for human existence. Analogies from nuclear arms control have been scrutinized for their potential to guide policies responding to great power competition in AI. Some commentators argue that these analogies provide constructive policy perspectives on the new challenges posed by AI. Other experts see them as an irrelevant distraction, distorting the focus when it comes to creating new frameworks to regulate unique and unprecedented AI developments.

This Briefing Paper argues that policy value in AI safety can potentially be achieved by adopting a broader perspective that incorporates concepts from nuclear arms control. Parallels with previous nuclear arms racing instil a warning that unbridled strategic competition in AI will hasten the proliferation of unsafe technologies. Norm institutionalization stigmatizing "nuclear taboos" through the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) prompts greater urgency for global frameworks to morally discourage dangerous AI developments. Just as epistemic communities provided knowledge for safeguards against nuclear proliferation, similar communities in AI are also vital for policy capital. Just as strategic stability and reciprocity for risk reduction resulted in competing Cold War superpowers pragmatically agreeing to arms reductions, similar logic can also reduce unsafe elements in today's AI race.

### ARMS CONTROL PARALLELS AND AI RISKS

Bilateral US-Russia nuclear arms treaties (some originating from the Cold War) were devised to slow a dangerous arms race and to reduce tensions. Some major treaties have recently collapsed as casualties of a wider geopolitical breakdown. The US withdrew from the Intermediate-Range Nuclear Forces (INF) Treaty in 2019. Russia, on the other hand, suspended participation in the New START Treaty and withdrew from the multilateral Comprehensive Nuclear-Test-Ban Treaty (CTBT) in 2023. CTBT remains unratified by the US. Global arms control is at a low ebb, rendering the security situation more dangerous. Nevertheless, a raft of new challenges with existential consequences are developing, including climate change, seismic demographic shifts, and rapid technological advancement.

Superintelligent AI offers immeasurable benefits, but grave risks enter the equation should AI develop uncontrolled without strong safeguards. AI's double-edged nature was stressed by US President Joe Biden in May 2023 when he told the CEOs of leading technology companies that "What you're doing [with AI] has enormous potential and enormous danger".[1] As with AI today, nuclear weapons emerged swiftly in the 1940s to dramatically restructure international politics. Through the logic of Mutually Assured Destruction (MAD), a tense nuclear deadlock reduced the risk of direct military confrontation between superpowers, but any miscalculations, mistakes or mishaps risked accelerating events towards thermonuclear annihilation. This has parallels with the distant human extinction risk discussed in relation to superintelligent AI advancement today.

Nuclear weapons have the grim advantage of making superpowers behave more cautiously. AI benefits are expected to have a more wide-ranging positivity: In widespread use across society, advanced AI promises to improve healthcare, increasing new and more effective treatments for diseases. It also stands to improve green technologies, helping to lower global carbon emissions. AI innovation might initially require

[1] Matt O'Brien and Josh Boak, "Biden, Harris meet with CEOs about AI risks", *Associated Press*, 5 May 2023, https://apnews.com/article/ai-artificial-intelligence-white-house-harris-578d623e473b0eeb3fa3e4728d7e9868.

President Joe Biden delivers remarks at an Executive Order signing on Artificial Intelligence on 30 October 2023.
*Source: White House, Adam Schultz*

considerable financial investment and sophisticated prototyping. However, much of AI will then be relatively easy to replicate or adapt. This is good news for the world's poorest regions if AI can spread as a socially accessible technology, furthering efforts to fairly distribute social and economic benefits.

Conversely, AI might distort and damage societies in unprecedented ways. Geoffrey Hinton, a leading AI scientist, identifies some of these risks. Hinton resigned from Google in 2023, claiming to have made this decision to more openly discuss the risks of AI to humanity.[2] He foresees that AI might flood societies with disinformation to the point where humans no longer realize what is true or false; it might entrench deeper social polarization by manipulating information selection to agitate partisanship; and it might make a large proportion of the labor force redundant, causing a surge in unemployment. However, the most ominous risk, as Hinton elaborates, is that AI will completely take over human decision-making. AI works on sub-goals to support final tasks. According to Hinton, the power to control is a universal sub-goal because the more power that is held, the easier it is to complete tasks. Extremely manipulative AI might irreversibly dupe humans into accepting its control without humans even being aware of it.

Not all experts share Hinton's outlook, as some see little hard scientific evidence that the development of AI poses existential risks. Parallels with nuclear weapons distract from more realistic dangers. For example, when AI powering national infrastructure malfunctions, this might be a serious problem, but it will not spell disaster for human existence.[3] Despite these assurances, fears linking AI with human catastrophe persist. In November 2023, it was claimed that a controversial breakthrough at the OpenAI company would pave the way for a drastic reduction in the demand for human labour skills, which could lead to a process of dramatic social upheaval.[4] Whether gravely catastrophic or slightly less so, it can be agreed that AI carries some significant risks that make its geopolitical implications worthy of scrutiny.

[2] Madhumita Murgia and Richard Waters, "Why AI's 'Godfather' Geoffrey Hinton Quit Google to Speak Out on Risks", *Financial Times*, 5 May 2023, https://www.ft.com/content/c2b0c6c5-fe8a-41f2-a4df-fddba9e4cd88.

[3] Yasmin Afina and Patricia Lewis, "The Nuclear Governance Model Won't Work for AI", *Chatham House*, 28 June 2023, https://www.chathamhouse.org/2023/06/nuclear-governance-model-wont-work-ai.

[4] Anna Tong, Jeffrey Dastin and Krystal Hu, "OpenAI Researchers Warned Board of AI Breakthrough Ahead of CEO Ouster, Sources Say", *Reuters*, 23 November 2023, https://www.reuters.com/technology/sam-altmans-ouster-openai-was-precipitated-by-letter-board-about-ai-breakthrough-2023-11-22/.

## PROBLEMS AND BENEFITS WITH PARALLELS

Uncertainty surrounds the precise directions of AI development. When confronted with uncharted waters, policymakers are often uncertain and seek reference points for guidance. In the face of daunting complexity, historical analogies can be tempting because they can help to better define policy pathways when the initial direction is difficult to discern.

There are obviously numerous differences between nuclear weapons and AI advancements. Not comparing like with like might create interpretations leading to flawed or futile policies. When debating AI safety, arms control reference points may partly reduce uncertainty for governments, but critics see this as a consolation that can mislead, as AI paves the way for unique and unprecedented technological and social transformations. From this perspective, irrelevant lessons from distinct policy areas are a distraction when it comes to defining creative new frameworks that are needed to regulate AI advancement.[5]

Much discussion linking arms control policies with AI advancement begins with arms racing parallels, where the focus is on strategic competition for AI advantage. Equivalences are also drawn linking norm institutionalization with AI safety regulation and the role of epistemic communities, where expert networks help to design these norms.[6] Further arms control concepts have not yet been meaningfully explored to provide additional lessons for improving AI safety when under geopolitical pressure. Underexamined concepts include strategic stability and reciprocal risk reduction, defining great power behaviour to mutually regulate and/or abstain from unsafe development. Their potential utility for today's AI race is elaborated later in this paper.

## AI ARMS RACING DISCOURSES

AI geopolitics is frequently portrayed as an arms race. Experts and political leaders reason that powers lagging behind in AI will soon see their great power capabilities weakened. These discourses intensify competition, but they also highlight some serious problems, emphasizing the need to prioritize policy thinking in areas linking great power competition with AI safety. The US and Russia dominated nuclear arms control, but this order is reshuffled with AI development,  where the US and China are global frontrunners. The EU's composite economy is lagging behind, but still in the game. Russia is further behind, but despite this subordinate position, President Vladimir Putin draws explicit parallels between arms racing and AI advancement. Putin predicts that the great power that takes the lead in AI will eventually dominate the global order. To this end, he has likened Russia's current AI industries to the development of Soviet nuclear weapons after the 1940s.[7]

There are several reasons to be pessimistic about Russia boosting its AI economy. Economic weakness, exacerbated by authoritarian constraints under Putin, is unlikely to stimulate domestic innovation to outstrip the US, China or the EU. Western economic sanctions and a brain drain from Russia's already modestly performing technology sector will undermine its AI ambitions. Nevertheless, authoritarian control also provides political leeway to redirect resources from elsewhere in society into AI development. For Russia, this is likely to be particularly the case for militarized AI, prompting analysis that it is struggling but not collapsing in the military AI race.[8]

Contrary to previous US–Russia dominance in arms control, Western relations with China are vital for AI safety today. Despite rising tensions, the West has kept the door ajar for rapprochement with China. However, Beijing's posturing in response to Russia's aggression in Ukraine poses a continuing risk of tensions spilling over into AI diplomacy. Having declared a 'no limits' friendship with Moscow, open or clandestine support for Russia's war effort will obstruct China's dialogue with the West on AI safety.

Russia transferred significant military equipment and expertise to China in the 1990s when China's military-industrial base and technology sector were weaker. Beijing is subject to a Western arms embargo following the Tiananmen atrocities in 1989. China's economy now produces sophisticated military and civilian technologies. Current global tensions increase the stakes, but Russian responses in Ukraine highlight that it can still source technologies essential for war from governments at odds with the West. If Russia's AI sector declines, falling further behind competitors due to sanctions and other economic deterioration, Russia is likely to look to China for advanced technologies in desperation to maintain great power capabilities. China

[5] Dylan Matthews, "AI is Supposedly the New Nuclear Weapons — But How Similar Are They, Really?", *Vox*, 29 June 2023, https://www.vox.com/future-perfect/2023/6/29/23762219/ai-artificial-intelligence-new-nuclear-weapons-future.

[6] Matthijs M. Maas, "How Viable is International Arms Control for Military Artificial intelligence? Three Lessons From Nuclear Weapons", *Contemporary Security Policy*, 40 (3) (2019), pp. 285–311.

[7] Reuters, "Sberbank CEO Tells Putin of Huge Returns on its AI Investments", 19 July 2023, https://www.reuters.com/technology/sberbank-ceo-tells-putin-huge-returns-its-ai-investments-2023-07-19/.

[8] Katarzyna Zysk, "Struggling, Not Crumbling: Russian Defence AI in a Time of War", *RUSI Commentary*, 20 November 2023, https://rusi.org/explore-our-research/publications/commentary/struggling-not-crumbling-russian-defence-ai-time-war.

will then face a difficult choice, either support Russia and burn more bridges with the West or turn to Western governments for strategic negotiations to better define global AI safety, helping to legitimize its position as an AI superpower.

## EPISTEMIC COMMUNITIES AND NORM CREATION

Parallels between AI development and arms racing continue in US strategic debates, where the Department of Defense describes AI as leap-ahead technology. US primacy in AI is perceived as extending its military dominance. During the Cold War, the US military was often a leading organization driving technological modernization. Military technologies were then adapted and spun off into the civilian economy. The US military is now far more reliant on civilian technological innovation. Private industry has a much stronger role in defining AI safety than in arms control, where governments are responsible for the development, maintenance and safety of nuclear weapons. Nuclear weapons are kept strictly under lock and key by governments. While notorious, illicit exceptions to this are rare.

The US government has less control over private enterprises driving AI modernization. Opportunities for financial profit might not always harmonize with safety considerations.  Market intervention might also pose grand strategic dilemmas for the US over the balance it needs to strike between AI safety and the pursuit of AI primacy. Overly stringent government safety regulations risk stifling innovation and, by extension, US efforts to lead in AI advancement. Avoiding this risk offers military-technological advantages that the US government will find difficult to resist, which is one reason why the Biden administration was initially hesitant to depart from laissez-faire AI regulations.

Voluntary guidelines were initially introduced to bring about a change in commercial ethics without resorting to direct intervention. This is somewhat reminiscent of the way in which  epistemic communities of scientific, government, and industry experts shaped nuclear arms control in the past. Like the scientists who observed firsthand the speed at which devastating risks from nuclear weapons were increasing from the 1940s onwards, some prominent AI developers have issued public warnings about the dangers of AI. Epistemic communities are important for providing ethical and normative expertise when it comes to designing technological safeguards in society. In the realm of AI, these communities  are more aligned with  commercial interests than nuclear scientists working on government programmes.

AI is a fierce battleground in US domestic politics. Lobbying and counter-lobbying constantly seek to alter government regulations. On the one hand, huge financial interests might compromise some in AI's epistemic communities and thus distort safety debates. On the other hand, experts have played an important role in casting informed doubt on America's previously voluntary AI safety guidelines. This contributed to a policy change in October 2023 when the Biden administration introduced the first legally binding US safety regulations for AI. Companies will duly have to undergo new safety assessments overseen by US government agencies and undertake equity and civil rights guidance, while AI's impact on the labour market will be subject to government scrutiny.

The US, the EU and China have different regulatory outlooks on AI. Each framework is grounded in the political culture of the implementing actor. The US aims to strike a delicate balance between safety and stimulating strong market innovation. The EU's AI Act promises to be the world's most comprehensive, with different levels of risk categorization for AI technologies. China's framework prioritizes algorithm safety and keeps the authoritarian interests of the Chinese Communist Party firmly in mind. All three frameworks are in their infancy, and it will take time for the regulations to become established domestically.

It remains uncertain as to which of these will have the most influence on defining the global rules of the game in AI safety. With technological transformation still outpacing regulatory progress, arms racing analogies retain value in raising wider awareness of AI risks stemming from great power behaviour. They provide sober warnings that this AI race needs to be multilaterally managed or irreversible dangers might emerge.

## CONCEPTS TO ENHANCE FURTHER DEBATE

Multiple bilateral and multilateral parallels between arms control and AI safety have not yet been explored to their full potential. Strategic stability is an older concept worthy of adaptation. US and Soviet leaders undertook strategic stability to devise agreements to mutually manage the Cold War arms race, which reduced the dangers of escalating superpower competition.

Strategic stability in AI can be achieved bilaterally or with wider great power multilateralism to reciprocate safety principles. This dovetails with Graduated Reciprocation in Tension Reduction (GRIT), another concept with Cold War origins that might inspire great power policies on safer AI. Applying GRIT to AI could be understood as every power having a baseline interest to not endanger humanity's future, thereby propelling a need to improve great power relations in this area. GRIT can start with a voluntary initiative by a single power with persistent willingness to persuade others to reciprocate for the purposes of risk reduction.[9] Among AI powers, the EU might be the most likely to introduce GRIT initiatives. Its AI focus does not prioritize military advantage as much as the US, China and Russia. The fact that the EU lags behind the US and China in terms of AI power should make GRIT more attractive to EU policymakers, and initiatives slowing down the AI race will also help the EU to stay more attuned to competitors.

From a wider multilateral perspective, when nuclear weapons emerged in the 1940s, a pragmatic global framework for non-proliferation still took approximately 25 years to negotiate and implement. The NPT came into force in 1970, based on an agreement between recognized nuclear weapon states to eventually disarm in return for other signatories remaining non-nuclear. This and other NPT regulations have helped to prevent the rapid proliferation of nuclear weapons. A profoundly negative international stigma shames governments that consider operating outside the NPT, a norm institutionalization that is sometimes referred to as "the nuclear taboo".[10] Similar ethical taboos exist against the further development and use of chemical and biological weapons.

Reviewed in five-year cycles, NPT continuation sometimes faces uncertainty, but it has nonetheless endured, unlike many bilateral US-Russia arms control treaties. An NPT-like multilateral global treaty for greater AI safety might establish norm institutionalization to negatively stigmatize the particularly dangerous aspects of AI, but this parallel also has limits. The nuclear taboo is only one of multiple safeguards preventing proliferation. Nuclear weapons are a technology that is difficult to replicate with sophistication, but this is less so with AI.

Once innovation burdens are overcome, many AI technologies are easily replicable. This will put non-proliferation initiatives based on negative stigmatization about dangerous AI under serious pressure. Overlapping verification and licensing regimes might reinforce this safeguard. As with weapons-grade uranium or plutonium, governments can track computer chips being used to train AI models, providing a mechanism for oversight on commercial AI development activity. Licensing regimes such as those regulating nuclear energy are being discussed, whereby companies developing mainly beneficial AI with particularly dangerous side effects can gain permission to do so under strict government safety regulations and oversight.[11]

Negotiating norm institutionalization to strengthen AI safety is even more complex than for nuclear weapons. Whereas nuclear weapons are confined to military security, AI is a socially all-encompassing technology. Agreeing on the most critical dangers to prioritize among a multitude is an intricately complex task. Initially, progress could be made in establishing norms between the main stakeholders to regulate dangers in narrower areas, such as ethically problematic military-specific dimensions connecting AI use with intelligence-gathering, autonomous weapons and targeted killings. Nevertheless, even efforts in a specific direction like this seem unlikely at present. The US-China summit between Biden and Chinese leader Xi Jinping in November 2023 only produced a nebulous agreement for more intergovernmental dialogue on AI between Washington and Beijing.

## CONCLUSIONS

Arms control analogies have some value in enhancing understanding of the current strategic competition in AI, even if these parallels have their limits. Arms racing analogies offer sober warnings that the AI race requires multilateral management to avert the risk of irreversible problems in society. Arms racing perspectives recognize that those lagging behind in AI advancement are likely to have their great power standing undermined. Competition to advance in AI is fierce, and there are as yet few great power initiatives supporting multilateralism for AI safety. When AI is emphasized as a means of strengthening military advantage, the current AI arms race is more likely to accelerate than slow down. The world's leading AI powers have only

[9] Alan Collins, "GRIT, Gorbachev and the End of the Cold War", *Review of International Studies*, 24 (2) (1998), p. 202.

[10] Nina Tannenwald, "Stigmatizing the Bomb: Origins of the Nuclear Taboo", *International Security*, 29 (4) (2005), pp. 5–49.

[11] Mauricio Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties", *Computers and Society*, 8 April 2023, https://arxiv.org/pdf/2304.04123.pdf. And Heidy Khlaaf, "How AI Can Be Regulated Like Nuclear Energy", *Time*, 24 October 2023, https://time.com/6327635/ai-needs-to-be-regulated-like-nuclear-weapons/.

recently introduced domestic regulatory approaches. As these become established domestically, there may be more impetus for great power dialogue centred on the US, China and the EU to establish global norms to strengthen AI safety.

There is a strong parallel between the epistemic communities involved in nuclear arms control and those involved in AI safety today. These communities contribute vital ethical and normative capital for technological safeguards. Nuclear arms control is primarily a governmental policy sphere, but governments often take a backseat to private industry in AI development. Market logic interacts with government regulations, making AI safety policy even more complex. Outlooks diverge on whether closer proximity to high financial stakes might compromise the impact of the epistemic community on AI safety.

Norm institutionalization, as in the case of NPT's nuclear taboo, has some lessons for AI. A global multilateral treaty that negatively stigmatizes governments that consider deriving strategic advantages from particularly dangerous areas of AI development has merit in theory. In practice, however, AI is a vast technological sphere compared with nuclear weapons. Reaching agreement on areas to prioritize in global dialogue will be arduous. To reduce risk before this, strengthening safety in AI could draw inspiration from other arms control concepts. Strategic stability can be established when great powers mutually adopt AI safety principles. GRIT can inspire individual great powers concerned about the global risks posed by uncontrolled AI to engage in persuasive efforts and initiatives to reduce these risks. /