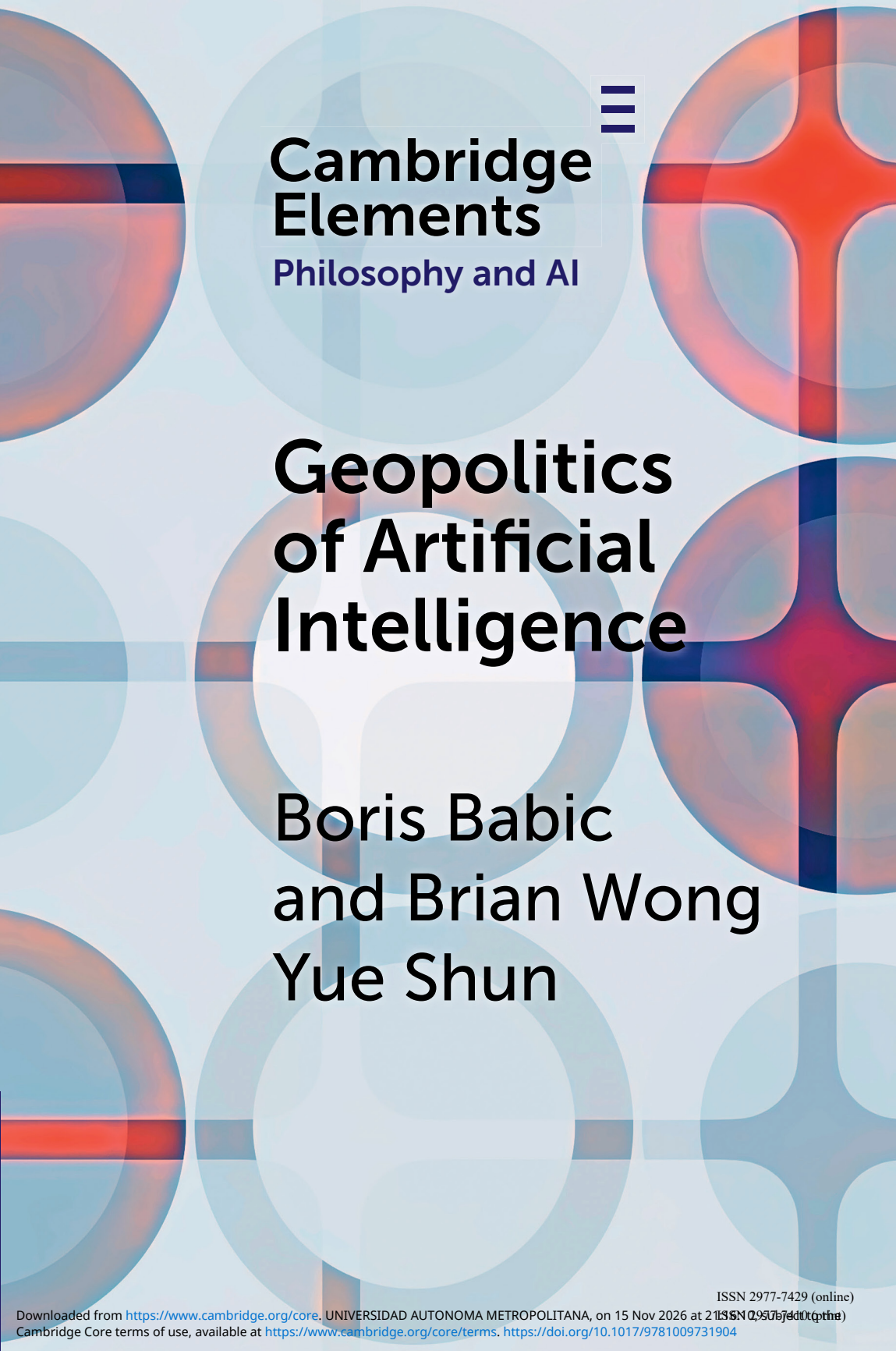


ISSN 2977-7429 (online)
ISSN 2977-7410 (print)

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:55:07, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Cambridge Elements

Elements in Philosophy and AI
edited by
Herman Cappelen
University of Hong Kong

GEOPOLITICS OF ARTIFICIAL INTELLIGENCE

Boris Babic
University of Hong Kong
Brian Wong Yue Shun
University of Hong Kong

CAMBRIDGE UNIVERSITY PRESS

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009731898
DOI: 10.1017/9781009731904

© Boris Babic and Brian Wong Yue Shun 2026

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, no reproduction of any part may take
place without the written permission of Cambridge University Press & Assessment.

An online version of this work is published at doi.org/10.1017/9781009731904
under a Creative Commons Open Access license CC-BY-NC-ND 4.0 which
permits re-use, distribution and reproduction in any medium for non-commercial
purposes providing appropriate credit to the original work is given. You may not
distribute derivative works without permission. To view a copy of this license,
visit https://creativecommons.org/licenses/by-nc-nd/4.0

When citing this work, please include a reference to the DOI 10.1017/9781009731904

First published 2026

A catalogue record for this publication is available from the British Library

*A Cataloging-in-Publication data record for this Element is available from the
Library of Congress*

ISBN 978-1-009-73189-8 Hardback
ISBN 978-1-009-73191-1 Paperback
ISSN 2977-7429 (online)
ISSN 2977-7410 (print)

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

For EU product safety concerns, contact us at Calle de José Abascal, 56, 1°, 28003
Madrid, Spain, or email eugpsr@cambridge.org.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the
Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Geopolitics of Artificial Intelligence

Elements in Philosophy and AI

DOI: 10.1017/9781009731904
First published online: August 2026

Boris Babic
University of Hong Kong

Brian Wong Yue Shun
University of Hong Kong

Author for correspondence: Brian Wong Yue Shun, bwongys@hku.hk

Abstract: This Element develops a sustained, pioneering treatment of the geopolitics of artificial intelligence through the lenses of political philosophy and rational choice. Three claims are advanced: first, the combination of AI and geopolitics gives rise to novel risks and opportunities. Second, the treatment of AI-geopolitical risks and opportunities should be treated through at least two different lenses: responsibility for risks should be fruitfully analysed using Iris Marion Young's seminal social connection model of responsibility for justice, while opportunities can be insightfully examined from the perspective of relational egalitarianism, in contrast to distributive and Rawlsian perspectives. Third, from a rational choice perspective, greater relational equality in AI access and cooperation among states can result in Pareto-improving rational (Nash Equilibrium) outcomes. Global AI dynamics should be conceived of as a Stag Hunt. This title is also available as Open Access on Cambridge Core.

Keywords: geopolitics of AI, AI governance, responsibility for AI, AI and grand strategy, AI and rationality

© Boris Babic and Brian Wong Yue Shun 2026

ISBNs: 9781009731898 (HB), 9781009731911 (PB), 9781009731904 (OC)
ISSNs: 2977-7429 (online), 2977-7410 (print)

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Contents

1 Introduction and Overview 1

2 Responsibilities for Interactional Risks 18

3 Opportunities and Global Relational Egalitarianism 33

4 A Game-Theoretic Perspective on Cooperation 47

5 Conclusion and Implications 65

Glossary 73

References 75

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

1 Introduction and Overview

Let us begin with a toy example.

> **LLM Bernie**. A government is developing its own Large Language Model (LLM), with the objective of assisting it with decisions pertaining to prospective immigration, refugee, and asylum-seeking claims. The LLM is deployed after significant debate between competing camps across the political spectrum, concerning the level of optimal restrictiveness and selection criteria with which immigrants, refugees, and asylum seekers should be admitted.

When presented with the profile (including both demographic characteristics and self-reported subjective information) of a particular applicant, Bernie will produce an output suggesting the category to which the applicant belongs, and it will supplement that output with a textual explanation of its decision, drawing on considerations available in the training data (e.g. crime and safety, economic value, cultural resonance, and compatibility).

To fulfil public expectations concerning 'putting their country first', the government explicitly refuses to train the model on data from other states and their populations (e.g. Government of Canada 2025). With the intention of bolstering the model's capabilities, the government voluntarily encourages everyone in the country to submit an updated version of their personal details and data via an optional survey.

Bernie is fundamentally opaque. The underlying functional relationship between the information in the survey and the prediction about a particular asylum case is difficult to understand, and it changes continuously. This is generally true of neural network models with many layers operating in high-dimensional spaces (Babic and Cohen 2023). The agent therefore operates as a *black box*, though it is given wide decision-making latitude, per the government's drive towards digitalisation.

Over time, the decisions facilitated by Bernie become increasingly contentious. In one case, a rich businessman is granted refugee status on the grounds that they reported being 'persecuted' for being a member of the ultra-wealthy tech elite in their home country. In another case, a poor single mother from a country with high crime rates is deemed a security liability in virtue of her origin, even though she has never committed a crime. A peaceful religious believer is suspected of harbouring 'terrorist intentions', while an anti-social anarchist is admitted.

Upon scrutiny, it is found that Bernie has been trained in such a way as to develop negative biases towards prospective entrants to the community

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

along ethnic, gender, and religious lines (Babic and Johnson King 2025). Additionally, the individuals who submitted their data in the voluntary survey have tended to be middle-class, from the country's ethnic majority, and generally suspicious of immigration. Such a skewed dataset may in turn precipitate problems such as *concept drift* (referring to a change in the relationship between the population characteristics and the target variable) and *covariate shift* (referring to a change in the distribution of population characteristics) (Babic et al. 2025). A simple example of concept drift would be where recidivism predictions on a skewed dataset become less reliable once that same model is used to make predictions for the population at large while a simple example of covariate shift would be where a facial recognition system trained on young faces has trouble correctly identifying older faces.

Today, many countries are developing what they call 'Sovereign AI' – AI developed using data primarily, if not exclusively, drawn from their own populations (A. Lee 2024). Such AI systems are in turn increasingly applied towards some of the core functions that a government agency or bureau would serve, for example, determining criminal sentences and penalties (in courts), allocating life-saving healthcare resources such as masks, donated organs and defibrillators (in hospitals), and adjudicating immigration claims (Beduschi 2021; Glazier et al. 2021).

As we shall see, so-called *Sovereign AI* systems such as Bernie are quickly assuming geopolitical significance. There exists a non-trivial possibility that LLMs, or other AI models, developed with the intention of facilitating foreign policy decision-making, can be subject to many types of undue and improper interference, or – under more benign circumstances – skewed data selection, incomplete data representation, poor benchmarking, as well as common issues of model deterioration, such as those described earlier and a lack of robustness in their outputs (Weise and Metz 2023; Raji et al. 2021; Bender et al. 2021). All these issues can produce responses that contribute to poor decision-making. While some of these could be non-consequential, such as when an AI model predicts that you buy a shampoo you would not enjoy, in the geopolitical arena, these poor decisions can be of monumental significance.

Indeed, the *Center for Strategic and International Studies* (CSIS) Futures Lab devised a study based on the Critical Foreign Policy Decision Benchmark. The team found that among seven major LLMs, across the four decision-making domains of deterrence, crisis escalation, alliance formation, and intervention, certain models (namely Qwen2 72B, Gemini 1.5 Pro-002, and Llama 3.1 8B Instruct) were more prone

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

to prescribing the usage of force – as compared with their counterparts (Atalan, Reynolds, and Jensen 2025). Crucially, all seven LLMs are more likely to recommend that the US, UK, and France escalate their actions (e.g. declaring war or using force) during a crisis, as compared to Russia and China. The models therefore exhibit a systematic diplomatic bias favouring the legitimacy of force by Western states; this is unsurprising given the origin of most of their training data (Jensen, Reynolds, and Atalan 2025).

On the other hand, the immense benefits that effective AI usage can yield for countries cannot be overlooked – from optimising public policy decisions and feedback mechanisms, to expediting digitalised innovation and research; from improving literacy and unleashing human potential via education, to bolstering regulatory and legal transparency (OECD 2025; Miao et al. 2021; Padmanaban 2023). Yet access to these potential benefits is not a given – indeed, across our world, access to cutting-edge AI technologies, their raw ingredients, and their resultant dividends can be extremely uneven.

These considerations have produced an ongoing debate concerning the US' semiconductor manufacturing equipment (SME) export policy. One view holds that the US should strictly gatekeep access to the most cutting-edge SMEs, in preventing non-geopolitically aligned actors from acquiring such technologies and wielding them in ways that undermine American interests (it remains unclear as to *what* and *whose* exactly these interests are, given the heterogeneity and diversity of actors within the country itself). A competing view suggests that it would be in the US' interest to *prevent rivals* from developing greater self-reliance, as well as ensuring that the platform of American chip companies (e.g. the CUDA platform of NVIDIA) is more extensively used across the world (Huang 2025).

Uniting advocates across both views, of course, is the assertion that the US must seek to out-compete China in its pace of AI development, whatever the cost may be. The rationale? That the US could ill afford to allow its geopolitical rivals to secure ahead of them powerful AI models that are capable of matching human beings across a majority of economically useful functions, which some refer to as *Artificial General Intelligence* (AGI) (Aschenbrenner 2024).

Many philosophical works on AI tend to focus on more abstract and foundational questions such as – "What is AI?", "How should we account for and deal with alignment problems between AI and humanity?", and "Is it meaningful to speak of AI agents?" or "Is it meaningful to speak of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

AI consciousness?" Yet this Element takes a different approach, laying out the landscape of a burgeoning field of study – *the ethical implications* arising from the interactions between *geopolitics and AI*. As Flint (2021) frames it, geopolitics denotes "the struggle over the control of geographical entities with an international and global dimension, and the use of such geographical entities for political advantage" (15).

The study of the philosophy of AI and geopolitics will necessarily entail our being more dynamically flexible and open-minded about the range of sources upon which we draw. Alongside peer-reviewed works such as books and journal articles, we will also be looking at commentaries and writings from reputable think-tanks, public-facing publications, and policy documents. This is intentional: the timeliness of such inputs, and the first-hand experiences they feature, are very much needed in order to reflect complex and shifting stakes involved in the topic at hand.

1.1 Why Should We Care About the Ethics of the Geopolitics of AI?

There are at least four reasons as to why this Element's contribution should be of interest.

First, the rapid emergence of AI technologies has brought about a genuinely interesting and unprecedented set of potential harms and benefits, which share some resemblance with but are otherwise distinguishable from previous technologies that have shaped geopolitics (such as the development of the atomic bomb or the emergence of the internet).

"How should we interpret these phenomena?" "What are the *ethical stakes* involved?" "Why should select corporations be permitted to exert such substantial influence over their consumers and over the foreign policy decisions of nation-states?" "When, if ever, should governments be allowed to incorporate AI-based algorithmic models into their infrastructural designs?" "How can we justify the procedural decision-making models developed by AI systems processing vast swathes of data?" These questions are innately *normative*, insofar as they focus on what *ought* to be, as opposed to (primarily) what is or what will be, which concerns the *descriptive*.

When we start to talk about the normative implications of these harms and benefits, or of *risks and opportunities* (a more useful distinction in the presence of uncertainty), we are operating in the world of ethics – characterised by scarce resources where fair or just allocations under conditions of competing claims matter. There is therefore a need for normative accounting and justification concerning how to handle these competing risks and opportunities.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Second, the rise of AI sharpens several long-standing *value tradeoffs* within existing philosophical literature on global justice. Consider two examples. Suppose it were the case that it would be in the interests of small, impoverished developing economies for technologically advanced developed economies to impose no restrictions upon their exporting to them of graphics processing units (GPUs) and other ingredients integral to AI development. A technologically advanced state could nevertheless object to the demands that it does so, on grounds of *national security* – that such advanced know-how could be weaponised against it by the wrong hands. Such reasoning has been repeatedly invoked to rationalise the kinds of restrictions adopted by China, the US, as well as the European Union (EU), over the export of parts of their AI supply chains (Davidson 2025; The White House 2025; Government of the Netherlands 2025).

While such normative discussions may appear futile in face of the pervasiveness of *realpolitik* in international relations (IR), such scepticism towards the relevance of moral discourse does not apparently extend to a distinct trade-off between *public security interests* (of the collective) and *individual privacy concerns* (of the individual citizen) on the domestic level, which is often taken very seriously. Indeed, the EU AI Act, for example, places considerable emphasis upon preventing the "loss of privacy, limitations to the right of freedom of expression, human dignity, and discrimination, for instance, in access to employment" as significant individual rights that cannot be sacrificed easily in exchange for purported collective benefits (European Commission 2025).

Third, AI systems pose unique questions with respect to the *assignment of responsibilities and liabilities*. Consider, for instance, the alleged usage of the 'Lavender' AI identification system by the Israel Defence Forces in the ongoing war in Gaza, which yielded at one point 37,000 potential targets based on their connections to Hamas (McKernan and Davies 2024; Frenkel and Odenheimer 2025), or India's deployment of AI-powered drones such as the Harop loitering munition and the Heron surveillance drone in its 2025 border conflict with Pakistan (Bindra 2025). Independent of our assessment of the ethical justifiability of these wars, there exist clear questions regarding the responsibility for the use of AI in such conflicts.

Fourth, how should different actors – from governments and international organisations to corporations and civil societies – *practically* tackle the risks and manage the opportunities arising from AI within the geopolitical arena? These *prescriptive* questions are not purely hypothetical. From diplomats and trade policy specialists in charge of negotiating

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

and gatekeeping access to frontier AI technologies to engineers and start-up investors trying to decide if they should sell their products to militaries and intelligence agencies, there are high ethical stakes in the choices they face. These stakes require justifiable and actionable prescriptions.

This Element is written for readers who want to understand the ethical questions raised by the intersection of AI and geopolitics. By the end, the reader should be able to identify the core risks and opportunities, distinguish between major normative frameworks for evaluating them, and assess the strategic logic of cooperation and competition among states.

1.2 Central Claims

This Element will not answer all of the questions outlined thus far or provide a survey of the disparate policies across the world. Instead, our main contribution comprises the following key claims, which we term (1) the Risk-Opportunity Claim, (2) the Justice Claim, and (3) the Interest Claim.

First, contemporary developments in AI generate a *unique impetus* for a theory of the ethics of the geopolitics of AI, in part because the combination of AI and geopolitics gives rise to novel risks and opportunities. The subsequent discussion will further reveal that existing frameworks in the study of IR and geopolitics do not take seriously the ethical questions at hand, while existing AI ethics literature may not provide sufficient clarity and resources in accounting for the uniquely geopolitical challenges. These key risks include *conflict escalation*, *unjustified violence*, and *resource misallocation*. Key opportunities include *developmental benefits*, ranging from improved *industrial capabilities*, and *workforce productivity* to operations, global supply chain logistics, and transportation, among many others. Call this the *Risk-Opportunity Claim*.

Second, the treatment of AI geopolitical risks and opportunities should be viewed, respectively, through at least two different lenses. In cases where such risks materialise as harms, and where such harms can be clearly attributed to the intentional actions of particular agents, we can hold them *morally responsible* for their actions and demand that they repair the effects of their actions. Yet in the many cases where such responsibility assignment is ontologically or epistemically infeasible, the ensuing responsibilities for repair should be apportioned in accordance with Young's (2006, 2011) seminal *Social Connection Model (SCM)* of responsibility for justice (Browne 2023).

On the other hand, opportunities accruing from AI are best viewed through the lens of global relational egalitarianism. In contrast to fledgling applications of Rawlsian principles of justice (Gabriel 2022; Westerstrand

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

2024) to the rights, opportunities, and welfare goods generated by AI, we adopt a distinctly more relationally orientated approach to conceptualising AI justice (Anderson 1999; Wolff 2015). Genuine global AI justice demands a world where nation-states, communities, and individuals can relate to one another as equals in their ability to access, develop, and regulate AI. Even if such a vision remains unattainable, we should still seek to approximate it. This amounts to the *Justice Claim*.

Third, we will consider one of the most extensively litigated arguments within the sphere of global AI justice, namely, that a state's restrictions to the exporting of, and other states' and non-state actors' access to, advanced technologies, are justifiable on grounds of the advancement of *its own interests*. Drawing upon game theory and empirical evidence concerning the limits to strategies of *supply chain disruption*, we advance the claim that greater relational equality in AI access can in fact produce cooperative Pareto-improving Nash Equilibria. We term this the *Interest Claim*.

1.3 The Limits of Traditional IR and Orthodox AI Ethics

Existing geopolitical analyses of AI have by and large opted for a great power-centric approach to understanding AI's potency, risks, and opportunities, typically omitting normative questions of how their AI policies may impact or interact with the agency, decisions, and interests of small and medium powers, as well as disparate non-state actors (Wang et al. 2024; Coeckelbergh 2025).

On the other hand, traditional AI ethics has tended not to address the geopolitical dimension, focusing instead on how algorithms affect individuals or population subgroups. AI ethics has since grown to include other topics, including research in alignment (Kirk et al. 2025; Ji et al. 2025), responsibility (Dignum 2020; Santoni de Sio and Mecacci 2021), safety (Gyevnár and Kasirzadeh 2025; Bengio et al. 2025), existential risk (Carlsmith 2022), manipulation (Carroll et al. 2023), deception (Park et al. 2024), and misinformation (Zhou et al. 2023). Yet what remains missing is an investigation of the complexity posed by how state and non-state actors interact with one another on the international stage.

In our view, conventional AI ethics has been primarily developed from a *micro-perspective*, in the sense that researchers have focused largely on the relationship between an algorithm and an individual or a group of individuals. In contrast, the framework we develop may be more appropriately thought of as *macro-AI ethics*, focusing on decision-making at an institutional level, including that of nation-states, corporations, and intergovernmental agencies.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Why the Study of AI Geopolitics Needs Ethics

The field of *AI geopolitics* can be conceptualised through three clusters of research: how are a) *developments in*, b) *governance and regulation of*, and c) *conflict deployment* of AI technologies shaped by geopolitical processes?

The first cluster revolves around *development*. A Lazard (2023) Geopolitical Advisory report identified four key 'building blocks' to competitiveness in the global AI race – *computing power* (including semiconductors and data centre infrastructure), *talent* (researcher quality and quantity), *data*, and *physical infrastructure* (energy and water resources). These blocks are asymmetrically concentrated in a handful of economies and corporations – for example, NVIDIA, Taiwan Semiconductor Manufacturing Company (TSMC), Broadcom, and Advanced Semiconductor Materials International (ASML) (Mahindroo, Suryanarayanan, and de Jong 2025).

For instance, the US and China are both endowed with significant concentrations of high-skilled talents and natural resources. While countries such as Japan, Israel, and France produced far fewer top talents in absolute quantities, their highly integrated roles within a US-led AI ecosystem have enabled their corporations and research laboratories to access cutting-edge systems (NVIDIA 2024; Press 2024; École Polytechnique 2025). The flipside to such cross-national synergies consists of governmental restrictions to access. Such restrictions are often justified in terms of *interest protection* – for example, through preventing nefarious usage of their own citizens' data by foreign companies.

Other literature has focused on restrictions put in place for *developmental throttling* – a chief case-in-point being attempts by governments to restrict the pace at which other countries are able to acquire and deploy cutting-edge semiconductors in GPUs, which play a critical role in AI applications. The bipartisan US CHIPS and Science Act, 136 Stat. 1366 (CHIPS Act 2022) was initially hailed in policy circles for curbing China's access to advanced chips – with leading consumer-facing Chinese giants, such as Huawei, long depending upon fabricators such as Samsung and TSMC in providing crucial components (Shivakumar, Wessner and Howell 2024). More recently, Shivakumar, Wessner, and Howell (2025) have argued that 'the imposition of export controls has resulted in China doubling down on its [own] development efforts that could well produce breakthrough technologies' – the announcement by Chinese start-up DeepSeek in January 2025 of an open-source research model, R1, which matched roughly leading models from Google and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

OpenAI, spurred speculation that Chinese developers were able to increase algorithmic efficiency and make use of insightful data compression and model optimisation strategies in order to make up for limitations in computational power and typically smaller datasets, as compared with their overseas counterparts.

The second cluster concerns the *governance and regulation* of AI. A prevalent approach among existing theorists (e.g. Bradford 2023) is to focus on key geopolitical clusters of AI powers – what she terms "digital empires" and Schmid et al. (2025) term "technopoles" – and ascribe to each region distinctive principles guiding their approach to regulation. Per this framing, the EU is portrayed as primarily 'rights-centric', leveraging its bureaucratic apparatus and effective legal dissemination methods in accomplishing the proverbial *Brussels Effect* – whereby European ordinances have an outsized impact on blocs of nation-states that possess extensive commercial connections to and legal ties with Europe (Wang et al. 2024). The US can be characterised as "market-driven", ostensibly because its government would "step aside to maximise the private sector's unfettered innovative zeal" (Bradford 2023, 24). Finally, some would assert that the Chinese model is top-down and strives to advance both innovation and stability, coupled with effective enforcement mechanisms underpinned by strong state capacities and understanding of AI technologies (Reed Smith 2024). More recently, Chinese state signalling suggests that it is endeavouring to position the country as a key proponent of openly available AI, building off the positive reception of DeepSeek; the 2024 Shanghai Declaration on Global AI Governance states, "[w]e advocate the spirit of openness and shared benefit and will promote exchanges and cooperation on global AI research resources." (Ministry of Foreign Affairs, People's Republic of China [MFA] 2024).

The reality is more complex than the tripartite framing of digital empires or technopoles suggests. As Zhang (2024) observes, in China, the delicate interplay between the senior political leadership, provincial and local bureaucratic regulators, large platform economy companies, and individual workers tends to give rise to a volatile regulatory process generating fragile policy outcomes which may be open to interpretation and rearrangements based on market and leadership sentiments. Meanwhile, Fang and Hwang (2023) argue that the US state is increasingly embracing its own form of "techno-nationalism", proactively promulgating the narrative that the "competition between distinct 'Chinese' and 'US' models of leveraging emerging technologies" is a *zero-sum struggle* and must be *won* (The White House 2025). The implication is that American companies

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

should fall in line with their state's national security imperatives, without questioning the underlying normative premise – and indeed, many of them have, with Anthropic's Dario Amodei becoming one of the leading defenders of US export controls (Amodei 2025), despite his confrontation of the Pentagon in early 2026 over the parameters of reasonable AI deployment. The EU AI Act's cumbersome risk-tiered system of governance looks very much like a top-down enforcement mechanism – so heavy-handed that many worry about its drag on innovation speed and flexibility. Clearly, therefore, debates over the geopolitics of regulation must evolve to better capture the nuanced similarities and differences in governance dynamics across different AI powers. While the 'multiple empires' framing is tempting, it is ultimately too simplistic.

The third cluster pertains to the application and use of AI in *conflict contexts*. Payne (2021) observes that autonomous weapons are too ubiquitous and useful for states to be wholly outlawed. AI-enabled 'warbots', then, are a matter of political inevitability – even despite potential risks arising from misalignment between the occasion and intensity of force and the specific military contexts at hand. Many have warned (e.g. Simmons-Edler et al. 2024; Kmentt 2025) of the perils of an AI-enabled autonomous weapons arms race, which could generate unproductive escalation, heightened uncertainty, and proliferation of AI technologies among illicit actors.

This emerging geopolitical discourse on AI is limited by two theoretical defects: ethical statism and great power solipsism. We briefly discuss both.

The first limitation consists of the absence of defence for what we term an *ethical statist* premise, namely, that individual states should behave in ways that advance solely their *own* interests. Most government policies, directives, and platforms on AI governance tend to place significant emphasis upon the securing of *core interests* of the state. This state-interest-driven nature of AI analysis tends to ignore considerations of what is just, fair or otherwise normatively desirable.

Much attention is paid towards the *instrumentalist question* of how a particular objective (e.g. 'national security') is to be attained; far less attention – if any – is paid towards the *normative question*: what *should* be the objective towards which countries must strive in the context of inter-state AI dynamics?

This claim will be contested by those denying that states possess any duties to actors beyond their borders. Some *neo-realist* accounts (Waltz 1979; Mearsheimer 2001) hold that, given the anarchic conditions

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

characterising the international order, states should prioritise their own survival, security, and interests. Indeed, Carr (1940) goes one step further in noting that the empirical reality of conflicts of interest between different countries gives rise to an international order forged through strength and contrasts of power. Walzer once famously described American IR training post-1945 with the following: "Moral argument was against the rules of the discipline as it was commonly practiced." (See Nye 2020).

In response, states *should* incorporate into their considerations the *avoidance of unnecessary harms to* and *generation of benefits for other states* when it comes to their decisions. *Should* here is normative: just as the interests of its individual nationals matter *qua* intentional human agents, the interests of non-nationals should also matter in constraining and guiding the actions of states.

Moreover, the prescriptions of more stringent export restrictions advocated by Western liberal democratic governments have at times ended up undermining the interests of their own citizens as well as those in allied states. Hence, there is an argument from self-interest to be made here as well. For example, the AI Diffusion Rule (Bureau of Industry and Security 2025b) proposed by former US President Joe Biden would have imposed excessively cumbersome regulatory requirements against American companies. Likewise, as we have seen in Supreme Court oral arguments in Learning Res. Inc. v. Trump, No. 24-1287, and Trump v. VOS Selections, Inc., No. 25-250, the main argument against the tariffs Donald J. Trump's administration has levied on most other states is that they impose a tax on American citizens in the form of higher prices for imported goods (Learning Res. Inc. v. Trump 2025).

The second issue, *Great Power Solipsism*, consists of the excessive emphasis upon large nation-states and downplaying the role of small and medium states, as well as non-state actors, when it comes to their role within the global AI landscape (Wang et al. 2024). Whether it be the 'tripolar' frame focusing on China, the US, and the EU (Bradford 2023; Schmid et al. 2025), or the 'Democracy vs. Autocracy' frame adopted by select actors in the US (Aschenbrenner 2024), these narratives portray global AI development through the lenses of two to three actors with overwhelming influence.

This view ignores two key resources by which smaller countries can play an active role in the global AI landscape. The first is *strategic hedging* (Kuik 2008, 2016): pursuing a strategy whereby they build partnerships with a multitude of actors while refusing to conform to one digital sphere of influence. For example, Singapore and Vietnam are skewed hedgers

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

seeking cooperation with American AI regimes while retaining openness to Chinese investment (Zhao 2024). Malaysia has positioned itself as both a chip manufacturing and re-exporting hub for American and Chinese semiconductor firms alike – courting Advanced RISC Machine (ARM) Holdings (Dib, Le, and Sharma 2025), and importing extensively from Nvidia (Sipalan 2025), while facilitating regional consolidation and expansion of Chinese tech firms. Malaysia now accounts for 13 per cent of the world's semiconductor testing and packaging (Medina 2025). Hedging enables these Southeast Asian states to preserve autonomy and flexibility.

The second means consists of *niche specialisation*. Medium AI powers are identifying and harnessing their expertise in particular domains of AI development, diffusion, and regulation in order to make themselves useful for a range of partners – thereby shoring up their bargaining capital. German AI companies, for example, have opted to specialise in image generation and translation (Praas and Balland 2025).

Beyond nation-states, there exist other actors who exert substantial influence over AI in geopolitical settings. From terrorist and insurgent groups to powerful large corporations (Gu 2024 warns of a "new type of Leviathan" emerging from the deep pools of Big Tech). These actors can make intentional choices that shape the trajectory of AI development, diffusion, and application.

Why the Study of AI Ethics Needs Geopolitics

The literature on geopolitical implications of AI is surprisingly underexplored in the philosophy of AI, as well as in the computer science literature touching on fairness, accountability, and other ethical topics. This could be in large part an accident of the particular time, place and context in which the field of AI ethics emerged and in which it evolved as a subdiscipline of computer science and philosophy. In this section, we explain how both of these facets created a path dependence affecting the way in which AI ethics unfolded, and why that path dependence largely steered clear of core geopolitical issues.

In the early 2000s, one of the most exciting areas of computer science and operations research was sparse matrix completion: identifying mathematical techniques for imputing missing values in large data matrices (Candès and Recht 2009; Li et al. 2019). Indeed, Netflix began running what they called the Netflix Prize Competition, a hackathon-style tournament for predicting film user ratings. To run the competition, they released

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

an anonymised data set of user reviews, on which competitors could train their models (Bennett and Lanning 2007). The competition was successful and proved to be very popular.

Yet in 2007, Arvind Narayanan and Vitaly Shmatikov did something different: instead of developing a model that would actually compete in the Netflix Prize challenge, they developed a technique that could take the anonymised Netflix Prize data and systematically re-identify the users referenced therein (Narayanan and Shmatikov 2007). This development amplified a lesson that had previously been learnt in the literature on electronic health records: namely, that anonymising data is not enough for guaranteeing user privacy. Around the same time that Narayanan and Shmatikov undermined the value of anonymising data, Cynthia Dwork and her colleagues had been developing methods for obtaining provable privacy guarantees by adding noise to data – a technique that came to be known as ε-differential privacy (Dwork 2006). Given these two developments, the technical literature on privacy grew quickly.

This was a whole new area of research: one could generalise Dwork's approach to cases with different data-generating distributions, or relax its assumptions, or add noise from a different statistical model (Dwork et al. 2006; Pejó and Desfontaines 2020). As fascinating as these developments are, there was a certain historical contingency to how AI ethics evolved: the time and place for working on privacy were just right, given the concerns about large recommender systems using our data without consent. Such issues would also consume a lot of attention in the AI ethics community, not necessarily because they were the most pressing, but because they were salient and tractable.

Approximately a decade following the Netflix Prize controversy, with the research on differential privacy having matured into a subfield of computer science, a different topic emerged and led to the birth of modern AI ethics. As classification models became increasingly better and capable of handling structured high-dimensional data, they were put to use for predictive tasks in a wide variety of social settings. One of these settings was criminal procedure. Some jurisdictions in the US began to use algorithms to assist judges or magistrates in bail and parole decisions (Okidegbe 2022).

Now, one might ask: why bail and parole? The answer is: their criteria are well-defined. In the context of bail, the main (or only) question for the decision maker is flight risk: will the defendant show up at trial? For parole, the main (or only) question for the decision-maker is the risk of recidivism: will the defendant commit another crime within some fixed

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

time period if they are released? Given these well-defined questions, one can ask: if an algorithm is used to assist in answering them, will that algorithm be biased or unfair towards marginalised racial, gender, or ethnic groups?

There was one algorithm in particular known as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which had been licensed to certain US courts for use in recidivism rate prediction (Brennan and Dieterich 2018). A popular magazine, *ProPublica*, gathered data on its use through Freedom of Information Act (FOIA) requests and published a feature-length article arguing that the COMPAS recidivism algorithm is in fact biased (Angwin, Mattu and Kirchner 2016). In many ways, this article significantly spurred the birth of AI ethics as we know it today.

The authors defined racial bias as disparity in terms of false positive mistakes (i.e. the rate of people from each group predicted to recidivate that in fact did not recidivate). They then analysed their FOIA data and found that false positive rates among Black defendants were higher. This finding immediately raised a number of tractable questions among academics across both computer science and philosophy: should we define bias in terms of observed outcomes? If so, how should we deal with the missing data problem (people who are kept in prison cannot be observed to recidivate)? Can we define different fairness metrics? Is ethical bias really reducible to false positive disparity? The literature subsequently exploded. Again, social circumstances were right, and the problems tractable, for these issues to be properly examined in an academic light.

What has driven AI ethics thus far is familiarity – of tools and of topics. Computer scientists know how to work with optimisation and dynamic programming techniques, for instance. Philosophers know how to debate one's definition of 'fairness'. Yet just because these are familiar questions calling for familiar tools does not mean that they are the most consequential. The geopolitics of AI may be too complex and multifaceted to reduce to a computational or conceptual problem within a vacuum. Yet we will attempt to address its complexities carefully.

1.4 Regulatory Developments

Before we move on, we provide an overview of some of the regulatory and legal developments relevant to the geopolitics of AI. These developments will inform our subsequent discussion, and especially the development of the interest claim.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

We will not attempt to forecast what all the preceding countries 'really aim for' – in their heart of hearts, so to speak – instead, we focus on their publicly available expressions and policy regimes, and those are strikingly different between the isolationism of the US and the interest in multilateral cooperation in China and the EU, as well as among other large and middle power states.

The United States: A Path to Isolation

In the US, domestic legislation has been used primarily to increase the likelihood of a 'US Victory' in a perceived winner-take-all AI race (in Section 4, we will argue directly against the framing of inter-state AI dynamics as a race). This has been accomplished in roughly three episodes: the 2022 CHIPS and Science Act, the tightening of the export controls regime, and the 2025 White House AI Action Plan.

In 2022, Congress enacted, and former US President Joe Biden signed, the CHIPS and Science Act, 136 Stat. 1366 (CHIPS Act 2022), appropriating 39 billion USD to bolster domestic semiconductor manufacturing capacity and another 11 billion for semiconductor R&D through government research labs and in collaboration with industry partners. The CHIPS and Science Act was motivated by "[c]oncerns about a decline in the US position in semiconductor manufacturing and technology and potential rise of China's industrial and technological competitiveness" (Blevins et al. 2023). Indeed, the Act explicitly states "[i]t is the sense of Congress that … technology development is critical to long-term competitiveness with near-peer competitors, including China" (Sec. 103(d)) and defines several key provisions by reference to the need for competitiveness with China and other "countries of concern", defined to include North Korea, Russia, and Iran (Sec. 103(a)).

The US export controls regime has been expanding for almost a decade. Beginning roughly with the 2018 Export Control Reform Act (ECRA), which authorised the President to control exports for national security and foreign policy purposes, 132 Stat. 1636 (ECRA 2018). ECRA reversed a decades-long trend of export control liberalisation, which began with the US-Soviet detente of the late 1960s (Casey and Kerr 2021), providing the statutory basis for the Export Administration Regulations (EAR), a set of rules and guidelines administered by the US Department of Commerce's Bureau of Industry and Security (BIS), including the Commerce Control List (CCL), which identifies commodities, software and technologies subject to export controls for national security and foreign policy reasons.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

In October 2022, BIS significantly updated the EAR with the interim final rule 87 F.R. 62186 (BIS 2022), with the expressed intent of 'restricting China's ability to purchase and manufacture high-end chips'. Interim Final Rule (IFR) 2022 targeted 'advanced computing and semiconductor manufacturing items,' effectively prohibiting the export of Nvidia's A100 and H100 GPUs, which were the flagship high-performance models at the time.

Finally, the White House AI Action Plan, unveiled during the second Trump administration, has reinforced the isolationist path. After exclaiming that "the United States is in a race to achieve global dominance in artificial intelligence," the document adds that "[w]inning the AI race will usher in a new golden age of human flourishing, economic competitiveness, and national security for the American people." The document goes on to emphasise that "[w]e need to establish American AI – from our advanced semiconductors to our models to our applications – as the gold standard for AI worldwide and ensure our allies are building on American technology." While this is admittedly mostly rhetoric and purple prose, it is worth studying to understand the kind of tone that the White House sought to set between 2022 and late 2025. With that said, it is well worth noting that, as of this writing, the US seems to have gradually loosened some of its more restrictive measures, allowing, for example, Nvidia's H200 processors to be exported to China (Babic and Wong 2026).

The approach in the rest of the world is decidedly different, and perhaps best characterised by a cautious interest in multilateral cooperation together for tit-for-tat responses to US restrictions and trade barriers.

European Union

In the EU, while Regulation (EU) 2021/821 establishes a unified regime for export licences, focusing in particular on so-called dual-use items (those with both civilian and military applications), such as drone and surveillance equipment, other initiatives suggest an increasing interest in multilateral cooperation. As Lu (2025) notes, "[t]he EU has actively advocated for the establishment of global AI ethical guidelines and governance frameworks within multilateral platforms such as the United Nations (UN), the Organization for Economic Cooperation and Development (OECD), the Group of Seven (G7), and the Group of Twenty (G20)" and the EU "actively supports the UN Secretary-General's proposal to establish a 'Global Artificial Intelligence Governance Agency'" (4).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

China

The tone in China is likewise decidedly different from that of the US White House AI Action Plan. For example, the Ministry of Foreign Affairs (MFA) released the Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence in November 2022, stating that "China is committed to building a community with a shared future for mankind in the domain of AI" and "finds it important to enhance the understanding of all countries on AI ethics" (MFA 2022). The document also states that "China calls on the international community to […] work to formulate a widely accepted international AI governance framework."

Meanwhile, in October 2023, President Xi Jinping released the Global AI Governance Initiative at the Third Belt and Road Forum; the document emphasised both the need for global cooperation and recognised the value of including developing countries in shared governance, stating "[e]fforts should be made to conduct international cooperation with and provide assistance to developing countries, to bridge the gap in AI and its governance capacity" as well as that "[w]e support discussions within the United Nations framework to establish an international institution to govern AI, and to coordinate efforts to address major issues concerning international AI development, security, and governance" (MFA 2023, see also MFA 2024, for a similar sentiment).

Of course, as some have pointed out, China's Global Initiatives exhibit "dexterous ambiguity" (Olander and Wong 2025), a feature that grants its decision-makers flexibility when designing and adapting their policies based on incidental and circumstantial needs. Likewise, we did see tit-for-tat responses to US noncooperation in China via both tariffs (Race and Espiner 2025) and exports of rare earth minerals (Swanson and Tobin 2025). It is imperative to discern between rhetoric and facts. The following sections will develop a framework for avoiding getting stuck in such noncooperative cycles.

1.5 Structure and Roadmap

This Element proceeds as follows: Section 2 discusses the risk component of the Risk-Opportunity and Justice Claims, advocating that an SCM-inspired framework is a critical complement to conventional models of moral responsibility assignment over harms caused by AI – especially within the geopolitical context. Section 3 explores the opportunity

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

component of the Risk-Opportunity and Justice Claims, arguing that the existing literature would benefit from taking a relational egalitarianism-centric perspective.

Section 4 draws upon game-theoretic analysis in grounding the *Interest Claim*, arguing that lifting restrictions on the exporting of and access to AI can in fact be in the interest of the states from which such AI originates. This section directly undermines arguments that aim to frame state interests against the interests of the international community and arguments that portray global AI development as a zero-sum game. This is most certainly not the case. Section 5 concludes with the practical upshots of our key contributions: a series of recommendations for improving international AI governance and inter-state AI dynamics more generally.

2 Responsibilities for Interactional Risks

The late sociologist Ulrich Beck (1992) coined the term *risk society* to capture the essence of modernisation – the transition from an *industrial modernity* to a *risk-producing* society, where the blend of technology, industry, and capital gives rise to new, often global, and previously largely unforeseeable risks (Spencer 2025).

We deem risks *ethically significant* insofar as they:

a) have a non-trivial possibility of
b) causing significant harm to
c) moral patients
d) whose core rights or welfare are thereby damaged.

The first and second conditions are necessary in order to exclude cases of extremely low-probability occurrences, such as existential risks from super-powerful AI agents, which are not the subject of this Element. The third condition holds, for the impacted party must be a moral patient (for us, this would be a human or a non-human animal). Finally, the infringement should undermine substantive interests – for example, bodily or property rights or fundamental rights protected by tort and criminal laws.

Substantively, there exist two broad schools of thought concerning how responsibility ought to be approached in the field of AI-induced harms or risks. One school of thought – as captured by what Iris Marion Young (2011) dubs the *Liability Model* – holds that we should view our responsibilities over AI risks through the lenses of moral blameworthiness and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

conventional liability (Floridi et al. 2018; Braun et al. 2020). An alternative view posits that the responsibilities to mitigate and redress such risks may be better understood through Young's own (Young 2006) *SCM* of responsibility, with more recent developments (Nussbaum 2011; McKeown 2018) serving as important addenda to her framework. Both approaches have clear merits and demerits, though we take the view that the SCM merits more substantive consideration and application to many instances of AI risks.

2.1 Interactional Risks

The interaction between evolving AI technologies and geopolitical tensions generates a set of new risks. We call these *interactional risks* – in the sense of an AI risk being compounded by its geopolitical dimensions.

A primary example is *conflict escalation*. Human decision-makers increasingly rely upon recommendations from AI agents to devise foreign policy strategies and responses. Commercial generative AI models tend to overestimate the level of threat exhibited by the other side in their textual outputs (Rivera et al. 2024). The CSIS Futures Lab study (Atalan, Reynolds, and Jensen 2025) highlights this phenomenon – should these agents be involved in real, live conflicts, we could see a significant surge in casualties, suffering, and economic losses. Similarly, concepts such as *concept drift* and *covariate shift*, as aforementioned, take on particular significance in the geopolitical space. If we take a model like Bernie, introduced in Section 1, trained on a local population due to a government's felt need for 'Sovereign AI', and apply it to asylum seekers or work visa applicants of different states, this essentially guarantees that we will observe a change in both the feature distribution alone (covariate shift) and a change in the joint feature-label distribution (concept drift).

Similarly, we can imagine a Bernie-like model trained to predict and offer prescriptions on even more geopolitically sensitive issues, such as which foreign state-owned enterprise should get placed on a sanctions list. Its outputs have compounding consequences, not just on the entity being sanctioned, but on the impacted population in their access to and prices they pay for goods, as well as on the foreign state's economic growth. The *possibility* of these outcomes is what we mean by *interactional risk*.

There are two geopolitically centred mechanisms through which interactional risks can emerge. The first involves the siloing of AI development across states with distinct geopolitical orientations – indeed, this precipitates covariate shift and concept drift, described earlier,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

because it circumscribes the world of training data in a way that is unduly unrepresentative of the world at large.

It can also give rise to what we call the *multi-alignment problem*, which refers to the emergence of two or more AI models with significant divergence in values, priorities, and judgements which produce outputs that are *in conflict* with one another. Here is how it happens: those working in AI philosophy would be well-familiar with the challenge of human-AI *non-alignment*, which is a general term used to denote instances where AI agents exhibit behaviours that result in deviation from what is in fact aligned with human interests, preferences, and/or expectations. Consider, for instance, an LLM trained to produce outputs that generate maximal happiness and utility in its contents for its human user, which nevertheless embarks upon a spree of creating misinformation and hallucinating contents that deceive its user into a false sense of elated security. While the instruction 'Maximise human happiness' is clear, the failure to include parameters such as 'Do not generate demonstrably false or insufficiently verified – through independent corroboration – outputs' thereby renders the prompt defective, and the resultant AI behaviours misaligned (Bostrom 2014; Gabriel 2020).

Now consider the fact that different communities and groups are likely to harbour divergent understandings and conceptions of both their own and others' interests, as well as their relationships with one another. For instance, an LLM trained on data that asymmetrically casts Country A's foreign policy in a more pacifist, non-threatening, and benign light than Country B may be more inclined to recommend that when faced with provocations from Countries A and B, a peaceful solution be adopted for A, while a non-peaceful solution is to be used for B. One could imagine an LLM trained on data that reflects the opposite – namely, B is more pacifist, non-threatening, and benign than A – yielding the exact opposite in its recommendations. Such discrepancies hence amount to an instance of multi-alignment.

Such divergence is not solely caused by the different beliefs, practices, and cultural norms of the demographics from which the underlying model data is *primarily sourced*. It is also exacerbated by the *increasing siloing* of AI training across country lines. As Esposito (2025) warns, "As of mid-2025, the geopolitics of AI stands at a crossroads. On one path, the world could slide further into fragmentation, with a digital iron curtain separating US-led and China-led tech spheres [… meaning that] data centres, networks and AI ecosystems would be divided by incompatible standards and mutual suspicions."

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

The second mechanism is the rising levels of inter-state mistrust over AI development, undercutting both domestic accountability and international cooperation. The dominance of the narrative of 'winning the AI race' (The White House 2025) against a strategic rival could inadvertently undermine the strength of checks and constraints on AI adoption in domestic contexts. OpenAI's 2025 Economic Blueprint states, "We believe in America because America believes in innovation", warning of the ostensible dangers of allowing China to gain the upper hand in AI development – should the US fail to capture the USD 175 billion in global investment capital (OpenAI 2025). The geopolitically fuelled emphasis upon national security drives extensive informational gatekeeping concerning model development and algorithms, which in turn undermines transparency and accountability over how AI is deployed in public infrastructure. Indeed, prominent corporations have invoked geopolitical considerations as a pretext for keeping their models closed (Anthropic 2025; NTIA 2024).

While senior Chinese leaders have vocally warned of the dangers – "if the braking system [for AI development] isn't under control, you can't step on the accelerator with confidence" (MFA 2025a) – their own companies are also racing ahead in embedding AI into their consumer-facing products – from driverless cars to smart manufacturing powered by humanoid robots (Goh, Baptista, and Li 2025). The drive for maximal societal AI diffusion is real, as articulated by official party policy statements (State Council 2025).

The advocacy of a no-holds-barred approach to AI not only speaks to the perils of a hyper-competitive *global AI political economy*, but also to the trust-deficit characteristic of such strategic contexts, which renders global cooperation increasingly difficult.

2.2 The Liability Model and Interactional Risks

Granting that the *interactional risks* are *ethically significant*, what *should* be done about these risks? A standard response may be to focus exclusively on instances where such risks in fact manifest as detrimental states of affairs, and attribute moral responsibility for such harms to certain agents. For instance, Bernie's outputs are used by the government agencies in charge of approving immigration, refugee, and asylum requests. The senior civil servant tasked with Bernie's outputs and making decisions accordingly may be plausibly viewed as morally responsible for the eventual false positive and negative mistakes.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

In such a case, there exists a) a clearly *identifiable agent* that is the b) primary *cause* (as a *but-for* trigger for a certain counterfactual, which means that if the agent had not acted, the outcome would not have occurred (Lewis 1973) or, as Lewis (2000) puts it, exerts "causal influence" over the outcome should it be overdetermined) of an event producing significant harm, and who is c) neither excusable nor exempted from moral responsibility (Strawson 1994).

If a), b), and c) hold, we can adopt the *Liability Model* (LM) of responsibility for harm:

> The most common model of assigning responsibility derives from legal reasoning to find guilt or fault for a harm. Within this idea of responsibility, one assigns responsibility to particular agents whose actions can be shown to be causally connected to the circumstances for which responsibility is sought. […] If candidates for responsibility can successfully show that their action was not voluntary or that they were excusably ignorant, then their responsibility is usually mitigated, if not dissolved. When these conditions do exist, however, it is appropriate to blame the agents for the harmful outcomes. (Young 2011, 99)

This view has the strength of cohering with our legal and moral intuitions concerning responsibility. While an AI model may play a causally significant, indeed determinative role, in the precipitation of certain harms, we would not – instinctively – hold the model morally responsible. Most would posit that AI models, as of yet, have yet to develop a sense of consciousness or self-consciousness; there exists a broad consensus in existing discussions that it would be erroneous to attribute moral responsibility to them. Even if such attribution is credible and viable (for instance, AI developments render it the case that a hypothetical agent in fact satisfies the threshold for moral responsibility), it does not fully absolve the human agents involved in shaping and moulding the AI's behaviours and usages of their specific responsibilities. Indeed, in existing discourse, we would look towards the moral agents who are most causally proximate to the specific harm and who are capable of understanding and controlling the AI (Kiener 2024), or who should, on grounds of strict liability, be seen as responsible for outcomes produced by the AI (Floridi et al. 2018).

The Responsibility Gap Challenge

Yet while there are instances of interactional risks where the LM can be readily applied, it is not sufficient for the responsibility assignment over many interactional risks stemming from AI in geopolitical settings. This is in part because of the *Responsibility Gap Challenge* (Mittelstadt et al.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

2016; Amoroso and Tamburrini 2019; Braun et al. 2020). Consider this challenge the normative equivalent of what legal theorists dub the 'liability gap' in the emerging field of AI regulation – the difficulty of assigning compensatory burdens and related costs in relation to harmful outcomes resulting from AI, due to the "algorithmic opacity, the distributed architecture of AI systems, and the systemic and diffuse nature of AI-related harms" (Tran 2025).

An example of a liability gap would be a case where a semi-autonomous Food and Drug Administration (FDA)-cleared AI medical device makes a mistake. There is a gap because (a) the device is not licensed to practise medicine and hence does not have malpractice insurance, while (b) its semi-autonomous nature may shield the residing physician from liability.

We propose the following statement of the challenge:

C1: AI agents can cause harmful outcomes in the geopolitical space.
C2: The human source of responsibility for those outcomes is unclear.
C3: Holding AI agents responsible in the traditional sense cannot be an adequate substitute.

C1 has been tentatively established.

C3 is interesting. On the one hand, there are objects which we clearly do not hold responsible: we do not hold a rock 'responsible' for tripping a jogger. On the other hand, certain accounts of AI agency posit that *only* upon their reaching a level of sentience (Wong 2026), self-awareness (Bojić, Stojković, and Marjanović 2024), or capability to consistently and reliably execute longer action sequences (Ahn et al. 2022; Yang et al. 2025; Fung et al. 2025), could it be reasonable to attribute partial responsibility to said entities, for they qualify as agents. Yet this does not appear at all likely in the foreseeable future.

First, the causal sequence of the harms will often be epistemically difficult to identify and ontologically diffuse. For example, as Gerke et al. (2020) argue, AI models in medicine and public health are best construed as continuously evolving systems, not products. This is equally true for complex AI systems used for geopolitical objectives. Within such systems, there are innumerable moving parts: for example, we may have multiple sources of data from many different vendors, multiple AI models, robotic devices on which those models are implemented: sensors, cameras, gyroscopes, etc. for processing real-time evidence, many subsidiary digital assistants, and many human inputs for guiding the system.

When a mistake is made, it may be essentially impossible to determine which component of the system has 'caused' that mistake, in the sense of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

but-for and proximate causation, as required by product liability in tort law or the aforementioned accounts of moral responsibility. This is not merely an epistemic challenge. Consider the various examples from the *imperceptible harms* literature (e.g. Parfit 1984; Nefsky 2012): the 'exact' contribution made by individuals who are distantly contributing towards interactional risks – for example, those voluntarily participating in the census generating Bernie's data inputs – given the overdetermination of the resultant outcome just cannot be identified. To hold all these individuals morally responsible, even granting that some of them are excusable in virtue of their ignorance, appears unjustified.

Second, our current approach to liability may not be sufficient for holding AI systems liable. If Bernie predicts that an asylum claimant would commit a crime if admitted, that claimant is then refused entry, and we will never observe them committing a crime in the country. Asylum claimants could come from many different countries, and they will in general not have the resources to seek legal recourse, nor are they likely to have standing in the country they are petitioning. Consequently, Bernie's produced harms (e.g. of access denial) are diffused across many people all over the world, none of whom have the means or ability to address them.

We do not currently have a legal regime whereby a system like Bernie could sue or be sued in court, and even if Bernie could be sued, it would be judgement-proof (i.e. insolvent/incapable of paying). Bernie also cannot be punished in most conventional senses – we cannot imprison or fine it (though, of course, we could shut it off). Some of these limitations could change in the future – we could require AI systems to have liability insurance, or bank accounts opened in their names. Yet at present, we are nowhere near that kind of immersive digital environment. The law still largely operates by looking for humans to blame. Practically, present circumstances lend strong support to C3.

Now we turn to C2. The conventional LM is inadequate in cases where clear, intentional perpetrators of harms cannot be found (e.g. the government employee interfering with Bernie's input). Indeed, there exist various reasons for which few human agents *can be* deemed blameworthy for most *interactional risks*.

As suggested by Kiener (2024), one way of articulating this premise is to say that few or no human agents satisfy the following two necessary conditions for moral responsibility – the agent must be capable of (i) meaningfully influencing (*Control Condition*) and (ii) knowing the range of the possible outcomes (*Epistemic Condition*), of the AI model.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

These two conditions are unlikely to be met for most agents operating AI in the complex global geopolitical environment. For Bernie, the census conductors, the computer scientists training the LLM, and the bureaucrats drawing upon its outputs to devise specific policies on refugees and tightening of immigration controls may each hold a 'jigsaw piece' to the overarching puzzle. Yet individually, none of them is sufficiently well-positioned to meaningfully influence the ultimate outcome. To what extent, and for what exact contribution, should we hold agents such as Bernie's original manufacturers, the politicians pressing for Bernie's adoption in 'keeping out the migrants,' or the citizens voluntarily submitting their data as morally responsible for Bernie's recommendations? These are just a few sources: there will also be multiple data vendors, multiple development companies contributing different components towards the overall system, and different human agents using Bernie for a particular decision (e.g. one employee may use Bernie for the initial assessment, while another may use it for a review of that assessment).

There is a practical tendency to err on the side of *obviousness* and prioritise assigning blame to the *causally nearest* human operator, for example, perhaps the immigration officers who turn Bernie's prescriptions into concrete policies. Yet as Elish (2019) notes, per a naïve version of LM, the nearest operator – just as the front of a crashing vehicle – could fall into the "moral crumple zone" and be unduly scapegoated for a long and elaborate causal chain of events preceding their decision. Likewise, Gerke et al. (2020) suggest that the system view undermines the temptation to lay blame on the nearest human operator. To the extent that they can be clearly identified using the control condition, these obviously responsible agents may not be, all things considered, the most responsible – only the most *exposed* in the relevant causal chains.

To illustrate this in light of the AI-powered Lavender system allegedly used by Israel in the ongoing war in Gaza in targeting suspected Hamas elements and affiliates, we must ask – to what extent should the operators using the system to identify terrorist suspects in the Israel Defense Forces (IDF) be held responsible for the eventual targeted assassination of these individuals? What of their superiors, or the policy leadership of the military, who decided that such an AI-enabled system would be conducive towards the political objectives – as stipulated by the democratically elected government?

Further, many of the individuals involved in these scenarios may struggle to understand the extent to which they have made active contributions to and the nature of the harms arising from AI deployment.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

This problem is compounded by the black box nature of AI systems like Bernie, meaning that human operators will be incapable of understanding the reasons for the recommendations of such systems (cf. Babic and Cohen 2023). The epistemic condition is unlikely to be fulfilled on the part of individual employees of large, sprawling corporate (e.g. Meta and OpenAI) or governmental complexes (e.g. the Pentagon or State Department in the US, or the Chinese Academy of Sciences in China).

We could set aside the normative discussion by postulating strict liability on legal and prudential grounds, assigning it to certain actors (both individuals and groups alike), for example, an entire manufacturing unit can be held liable for the high incidence rates of false positives/negatives when it comes to the AI-powered drones, and tasked with the responsibility of compensating injured victims and loved ones of killed victims. This move can theoretically decouple the question of legal and practical responsibilities from the more complex and controversial debate over moral responsibility assignment.

Yet this carries a high degree of arbitrariness: we could then pick out any particular employee, for example, from a manufacturing unit, and prosecute them at will. It also fails to answer questions concerning how such liabilities ought to be distributed among individual members of these collective units. This seems unjust. As such, the Responsibility Gap Challenge remains a potent concern.

2.3 The Social Connection Model and Interactional Risks

Fundamentally, the LM is perhaps not the most applicable nor reasonably justified in the context of many interactional risks – given the intricateness, complex entanglement, and substantive challenges associated with holding specific actors responsible for their behaviours. In these cases, we should turn to an alternative that can serve as a helpful supplement: one that steps in where the LM runs out of steam or traction. Let us now consider Young's *SCM* – while the SCM is by no means innately superior to the LM, it fills a critical gap when it comes to the geopolitical dimension of AI risks.

Interactional Risks and Structural Injustices

Per Young (2011, 52), structural injustices refer to instances where:

1. "Social processes put large groups of persons under systematic threat of domination or deprivation of the means to develop and exercise their capacities."

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

2. "These processes enable others to dominate or to have a wide range of opportunities for developing and exercising capacities available to them."
3. These processes comprise "a consequence of many individuals and institutions acting to pursue their particular goals and interests, for the most part within the limits of accepted rules and norms."

Wong (2025a) terms these, respectively, the *Domination-deprivation*, *Enrichment*, and *Non-individuation Conditions*. Existing literature has already identified and analysed structural injustices posed by AI in certain contexts (Himmelreich and Lim 2022) – education (Madaio et al. 2022), healthcare (Braun et al. 2023), and algorithmic decision-making more broadly (Browne 2023). Browne, in particular, deserves particular credit for her efforts in incorporating structural injustice into the context of AI injustices. Indeed, Browne (2025) is seminal in devising a robust account of political responsibility and the governance of technologies in the era of AI. Likewise, Smart and Kasirzadeh (2025) rely on Young's SCM to develop a social notion of AI explainability.

However, relatively little has been said on how Young's work can be applied and developed in the context of international conflicts or geopolitical tensions. As such, we can fruitfully apply the theory of structural injustice to the subset of interactional risks identified thus far.

One may wonder whether a risk is an injustice. In the context of interactional risks, we take it to be so for two reasons. First, these risks can induce reasonable agents to *change their behaviours* and adapt their preferences accordingly (e.g. a soldier turning towards an AI model for recommendations on whether to open fire). Second, such risks give rise to the potential for *domination* – as Pettit (1999) puts it, domination consists of a power asymmetry between a dominator who *can* arbitrarily exert influence over the dominated. In any case, assuming risks can amount to injustices, many of the interactional risks we examine in this section would fit the three proposed conditions:

Domination-Deprivation

The mere *risk* of unwanted conflict escalation and the deployment of disproportionate and unjustified force against innocents can place prospective targets of AI-enabled weapons (including both conventional and lethal autonomous weapons) under the 'systemic threat' of significant bodily harm and death. Extensive propagation of misinformation comprises real threats that can significantly undermine the quality of life,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

decision-making autonomy, and ability to devise and maintain stable life plans and expectations of ordinary citizens.

Crucially, there could be a plurality of human actors (e.g. soldiers, infrastructural planners, programmers, engineers, etc.) involved in dominating these victims through the vehicle of AI technologies – these actors may neither possess complete control (*contra* the Control Condition for moral responsibility) nor knowledge (*contra* the Epistemic Condition) yet are still involved in the large-scale process of domination. This is also true in the case of Bernie's immigration decisions – the large groups may be systematically deprived of the conditions they deserve, yet they are too powerless and diffuse to sufficiently voice their protest.

Enrichment

'Cui bono?'

Interactional risks clearly exist for a reason – certain individuals stand to disproportionately gain from the processes giving rise to them. Consider, for example, large, AI-powered weapons manufacturers, which sign billions of USD worth of contracts with government actors to provide defence infrastructure. Lushenko and Carter (2024) have dubbed this phenomenon a "new military-industrial complex". Alternatively, there are also plenty of Silicon Valley AI companies that benefit from the exploitation of poorly paid workers in Kenya and Uganda (Perrigo 2023; Hao 2025).

Non-individuation

This condition has been fairly exhaustively litigated and discussed in previous sections, drawing on Gerke et al. (2020)'s systems view of AI models. Note that structural injustices can co-exist alongside individually or collectively perpetrated injustices. That particular harms inflicted upon a group of victims cannot be individuated need not entail that all of the harms cannot be traced to the blameworthy actions of any agent. As such, granting this condition for interactional risks does not thereby let off the hook actors who are clearly culpable in virtue of their *intentionally perpetrating harms*.

The SCM and Responsibilities for Interactional Risks

The SCM, as originally advanced by Young (2011, 104–113), provides an alternative to the LM; it comprises the following three features:

1. *Orientated around structures, not individuals*. For us to interpret and "criticise the way that many individuals and groups face too limited

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

and unsavoury sets of options", what is needed is an "account of large-scale systemic outcomes of the operations of many institutions and practices that constrain some [… and] enable others" (Young 2003, 4); individuals who are 'connected' to such structural injustices should be held responsible.

2. *Primarily 'forward-looking'*. Responsible actors should focus their energies and attention on devising solutions that can be implemented to tackle injustices that have yet to manifest – namely, to prevent them altogether, or mitigate against their worst effects. The kind of responsibility she envisions here is non-blame-centric – indeed, she suggests viewing it as *political responsibility*.
3. *Anchored in fundamentally shared and non-isolatable responsibilities*. The responsibilities arising from the SCM are essentially shared across all actors acting within the sphere permitted by rules and norms of the society.

Two constructive amendments should be incorporated into this account.

The first is a clarification by McKeown (2018) concerning the meaning of "connection". She is concerned that the vagueness of social connection could render the concept of "connection" over-inclusive, thereby assigning responsibilities to actors who are tangentially and insufficiently related to the injustice. Drawing a distinction between two kinds of causation – attributive causation, which tracks the LM, and explanatory causation, where *background conditions* causally contribute towards particular injustices, she construes connection as the *reproduction of structures* amounting to *man-made explanatory causes* of injustices, which involves a distinctly lower threshold of proof than conventional causal accounts of liability.

That is, an individual – so long as they are behaving with a modicum of intentionality and bare awareness concerning their actions – can *reproduce* a background condition even in cases where such a condition is overdetermined. For instance, in choosing to purchase a branded shoe made by a sweatshop worker, a particular customer is contributing revenue towards the brand. Even though the marginal revenue would have made no difference to the number of shoes the company produces, the customer has left their mark on the exact sales total. They are, to that extent, reproducing (even if not causing, *per se*) one of the conditions for the continued employment of sweatshop labourers. The venture capitalist backing a fledgling AI start-up specialising in lethal autonomous weapons is reproducing the background condition of AI 'enhancements' to killing technologies, even if some other investor would have entered the picture in their absence.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

The second tackles the objection raised by theorists including Nussbaum (2011) and Reiman (2012), concerning the "free pass" that Young's SCM apparently generates. If the SCM indeed does away with all notions of retrospective guilt for past moral transgression or ongoing failure to redress past injustices, Nussbaum fears that individuals would "get a free pass indefinitely" (xxi). The worry is that the SCM has no teeth: in the absence of real accountability mechanisms, there would be neither motivation nor an appropriate acknowledgement of the failure on the part of these non-compliant actors.

Where responsible agents (under the SCM) fail to discharge their responsibilities stemming from structural injustices, they can be *criticised* (Young 2011, 165). Such criticism *can* be meaningful, even while it does not amount to outright blame – for example, we can criticise the politicians in the city of Toronto for failing to provide adequate public transit, even without going so far as to say they are blameworthy and hence punishable for not doing so. The focus of our criticism is their forward-looking responsibilities: they *ought to* strive to improve the city's public transit. Indeed, such criticism may well entail that if they do not do so, they are less likely to be elected again – this is the way in which the SCM can still have 'teeth', despite the fact that it floats independently of traditional notions of blame, which stand at the core of the LM.

More generally, we advocate a *pluralist view of responsibility* for injustices (Wong 2025b). Any robust account of justice should make room for both forward-looking and backward-looking conceptions. In instances where perpetrating, complicit, or even culpably negligent actors can be easily identified, the LM should be applied. Yet in cases where such identification is not possible, we suggest that the SCM, as developed earlier, ought to be utilised.

We will call the SCM, with these two important clarifications, the SCM*.

Applying the SCM* to Interactional Risks

We can now appraise more fully the normative stakes associated with interactional risks, beginning with the following claim: actors who reproduce the *explanatory causes* of, and who are hence *socially connected to*, the risks arising from the *interaction of geopolitics and AI*, thus develop criticisable responsibilities both to *redress the past and present effects of* and to prevent or *ameliorate the future effects* of *such risks*.

Consider Susan, a hypothetical discriminative AI model that is trained with the intention of identifying hostile combatants and prioritising them for strikes by lethal autonomous weapons. During the development phase

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

of this model, the pressure from politicians on companies in the state to develop a model before their rivals produces lax regulations, which enables the firm to lower safety standards. Susan is first trained using data carefully assembled by engineers well-aware of the risks of improper or skewed data inputs; yet, upon exhausting available data, the development team then hires an outside company that specialises in developing synthetic data, some of which reproduces well-known quality problems.

The defence minister declares that the government will partner with Susan's developer. The highest-ranking military officer, reporting to the defence minister, is responsible for Susan's implementation throughout the military. Susan is eventually deployed in a war where squadron leaders are instructed to use her predictions to inform lethal drones. These drones kill over 50,000 individuals, among whom 3 per cent turn out to be innocent non-combatants. Who is to bear the responsibility for the 1,500 unjustly killed individuals?

The LM struggles to adequately answer this question. There could be a case made for holding Susan's developer collectively liable for its negligence to check for the model's output – though the severity of negligence does not appear to match up to the extreme aberration that the 1,500 innocent deaths comprise. Accordingly, should we hold the squadron leaders morally responsible? Or the military officer? The minister? The developer? The engineers?

All of them? None of them?

In contrast to the LM, the SCM* begins from the ontological complexity of our lived moral realities as its starting point – cases such as Susan are where its comparative advantage lies. The SCM* allows us to conclude, comfortably, that all of the actors who reproduced one or more of the explanatory causes (e.g. the conception and design, development, and safety standard testing of Susan; the facilitation and execution of the partnership agreement between the government and the firm; the deployment of Susan by the military) must bear corresponding, non-blame-centric responsibilities to redress Susan's damage.

These responsibilities are forward-looking and applicable to all individuals structurally connected to the innocent deaths. Substantively, this could entail the trio of reparative responsibilities – all responsible agents should *compensate, commemorate* and *apologise for* these injustices – and *oppose* future occurrences of such injustices, as discussed at length in Wong's account of *moral debt* (Wong 2025c). As stipulated by the original SCM, the degree of responsibility borne by individuals will depend on parameters such as the extent to which the agent possesses power and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

decision-making authority, overlapping interest (or lack thereof) in the redressing of injustice, and the ability to in fact make a difference (Young 2011, 144–147; McLaren 2019). It would be unreasonable to expect a junior sergeant in the army to sabotage Susan or to intervene in the decision of every squadron leader in the army. It would be less unreasonable to expect a senior data engineer in Susan's company to perform rigorous due diligence on the kinds of companies with which they are partnering and investigate the data sources on which Susan is trained.

The earlier argument can be applied to other kinds of interactional risks – AI-precipitated conflicts may be socially connected with poor development teams or with military leadership delegating AI models disproportionate influence in decision-making. AI-generated misinformation may be socially connected with politically motivated technological leaders seeking to shape geopolitical alliances. AI robotics systems may be socially connected with the destabilisation of global labour organisations.

One may worry about how much the SCM* has to say over how to fix structural injustices. Even though models of liability are separable from theories of how such liabilities ought to be addressed, we have already given some indication of the kinds of actions that would be appropriate – compensation, apology and commemoration – and commitment to opposing said injustice.

All the actors involved in the 1,500 innocent deaths caused by Susan are criticisable by third parties; such criticism can materialise as important reactions and judgements of them by others, which could thus shape their future actions. A failure on the part of any of these actors to discharge their forward-looking responsibilities could render them criticisable. In enforcing such criticisms, government authorities and non-state actors also have the permission to impose penalties on, for example, the data engineers, over their failure to comply with these, non-blame-centric requirements. For example, the military leadership may not have its mandate renewed. Likewise, the engineers who developed a model with such disastrous consequences could be passed over for promotion.

While this may not seem harsh enough, remember also that the SCM* can exist with the LM. For example, it may be that the investigations uncover some serious wrongdoing among the developer's engineers: say, they bore prejudiced attitudes towards the people in the country where Susan was going to be used and intentionally designed a poor system. Or they were simply reckless about its consequences. In that case, the LM can take over, with these people being deemed

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

morally responsible and held legally (and criminally) liable. The SCM* is designed to fill the *responsibility gap* left by the LM – i.e. it applies where the LM cannot be applied.

Section Conclusion

In this section, we have defended an alternative approach to mainstream views on responsibility for AI injustices – especially in the context of geopolitical risks. While the attribution of moral blameworthiness and backward-looking liability is tempting and reflects orthodox intuitions, we should take seriously cases where the background conditions for structural injustices are reproduced by human actions, and the ensuing forward-looking responsibilities. In such contexts, where the LM fails, the SCM* can fill the normative gap.

Having addressed responsibilities for AI risks, it is now time for us to turn to the question of opportunities that arise from AI's advent and adoption.

3 Opportunities and Global Relational Egalitarianism

While Section 2 examined the interactional risks between AI and geopolitics, this section will focus on the normative implications arising from the upsides of AI on the *global level* – from driving groundbreaking pharmaceutical research and scientific discovery (Jumper et al. 2021; Callaway 2020, 2022) to minimising energy waste and reducing carbon emissions in commercial buildings (Ding et al. 2024). Indeed, AI has also shown wider socioeconomic benefits through improving education access and outreach among underprivileged communities (Blackmon 2024) and bolstering the ability of companies and investors to undertake comprehensive macro risk assessments (Petric 2025). We can term these *potential* upsides as *opportunities*. Yet these opportunities are not distributed equally between individuals in states well-endowed with AI resources and their less-endowed counterparts, which prompts the following question: how can the opportunities generated by AI developments be configured in a way that is normatively justifiable?

Distributive luck egalitarians (e.g. Cohen 1989) could argue that the society into which we are born is innately arbitrary and is not something for which an individual agent can be held responsible. A teenager born in parts of Sudan or the Central African Republic, for example, is not at all responsible for the state of conflict-ridden destitution to which their country is subject. As such, to the extent they struggle with accessing most

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

LLMs or AI applications in everyday life, this is not a disadvantage that can be justified. The apparent gap in access between the Global North and South in accessing AI opportunities (UNCTAD 2021; Heeks 2022) cannot be 'justified' – while it is empirically explainable – with the reason, "Your country is poor, tough luck!".

Luck egalitarians would hence advocate the reallocation of global AI opportunity structures in a way that equalises, as much as possible, individuals' ability to access the many dividends of AI, regardless of their nationality, gender, ethnicity, or, indeed, wealth. While this vision is appealing in abstract, we may be rightfully concerned about the feasibility of its prescriptions. We may also believe that it obscures the more fundamentally *political* point of equality (Anderson 1999) – that equality maps not onto individuals' experiences and access to resources in a vacuum, but how they relate to one another, and their standing and status within a given polity. We shall see this critique of select forms of egalitarianism resurface in due course.

As such, this section focusses on two other approaches. It begins by observing that viewing global AI through the lenses of neo-Rawlsian justice (e.g. Gabriel 2022; Westerstrand 2024), while plausible, faces *three challenges*: first, neo-Rawlsian theory ignores the fact that AI, as with other forms of advanced technologies, is not a natural resource, but a socially constituted good embedded with intentional and meaningful labour; second, it fails to take sufficiently seriously the implications of political communities' collective rights to self-determine; and third, it oversimplifies the question of global AI justice by treating it as merely a question of resource possession and usage.

In contrast to neo-Rawlsian approaches, relational egalitarianism (RE) frameworks "demand […] we establish communities whose members relate and stand as equals, without social hierarchies, relations of inferiority and superiority, exploitation, or marginalization." (Bidadanure and Axelsen 2025) They commonly advocate that all individuals should live in a society of equals – though they disagree over what exactly this would entail. Some would focus primarily on the elimination of hierarchies, oppression, and other forms of subjugation (including the domination abhorred by republican theorists) (Anderson 1999); others primarily champion a positive vision where members of societies view one another with respect and see them as equally deserving of respect (Wolff 1998; Scheffler 2003).

Indeed, drawing upon strands of RE (Anderson 1999; Scheffler 2015), we suggest a framework of *Global AI Relational Egalitarianism* (GARE), which applies RE principles to global AI governance. Notably, the process

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

of governance here encompasses questions of desirable development, diffusion (distribution), and usage. Furthermore, a just governance model that complies with GARE will include distributive elements and reflect relational considerations.

3.1 Neo-Rawlsian Justice and Global AI Governance

AI is increasingly expediting and enabling groundbreaking innovation across industries. It has been pivotal in strengthening the competitiveness of consumer-facing firms. For example, the Chinese consumer ecosystem is extensively digitalised, with government statistics highlighting that the country's finance, government, telecommunications, and healthcare sectors had achieved an average AI penetration rate exceeding 60 per cent by 2025 (Frost & Sullivan 2025). Additionally, AI plays a pivotal role in optimising manufacturing operations (see e.g. Davenport and Ronanki 2018). How should we think about the distributive justice considerations of these tremendous and growing opportunities?

Some scholars invoke Rawlsian theory in applying his principles to the broader global justice context (Beitz 1999; Caney 2006; Pogge 2008). Gabriel (2022) articulates a compelling case for approaching AI through a Rawlsian lens – conceptualising socio-technical systems such as AI and its applications not as aliens to, but subjects within, frameworks of justice.

Observing the salience of Rawls' concerns – that a social practice should be regulated by principles of justice if the dearth of regulation would culminate in invidious inequalities, and if it induces a "profound and pervasive impact" upon individuals' life chances (Rawls 1971, 82) – Gabriel holds that AI should be designed and deployed in ways that conform with the two Rawlsian *principles of justice* – concerning the equalisation of basic liberties, and the justification of inequalities conditional upon their benefiting the least advantaged, and attachment to a fair equality of opportunity over offices and positions, respectively (Gabriel 2022; for a full discussion of what these principles entail, see Rawls 2001, 42–43; Bay 2023).

Meanwhile, Charles Beitz extends the Rawlsian account to the global level, focusing primarily on the distributive component of Rawls' conception of justice. Beitz (1999, 2009) argues that, given the empirical reality of economic interdependence between nation-states today, not only are boundaries between these entities morally arbitrary, but they are also fundamentally blurred when it comes to the effects natural resources exert over individuals' life chances.

Given how integrated global AI supply chains are (with clear interdependence over semiconductors, data, and talent, for example), should we

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

not also apply Beitz's global extension of Rawls' framework to how AI is distributed across countries? If so, then a 'Global Original Position' may be suitably devised to yield global principles of distributive justice for natural resources – mirroring the principles Rawls sees as applicable within only a nation-state. Per the two principles of justice, we should ensure that across the world, individuals' basic liberties are *not violated* by our usage of AI and proceed subsequently to equalise their *access to opportunities* arising from AI.

Limitations to the Neo-Rawlsian Framework

There are three reasons why this simply stated neo-Rawlsian framework is inadequate.

First, while Beitz primarily focuses on *natural resources* – on grounds that their distribution across different countries is geographically determined and hence arbitrary – the same cannot be said of AI, which requires investment of considerable artificial resources to develop, produce, and sustain – such as the development of large data centres and related infrastructure (e.g. Heating, Ventilation, and Air Conditioning (HVAC) systems). That these goods are created, reproduced, and transformed by human users complicates the picture of justice – especially if we recognise the existence of property or more generally the private rights of ownership and commercialisation of AI models. If we heavily invest our time, energy, and capital into building a house, we do not appear to be bound by overriding moral obligations to share it, even if doing so promotes distributive justice (Nozick 1974).

Second, as highlighted by nationalist theories of various kinds (Miller 1995, 2007; Riker 2009), this broadly cosmopolitan distributive justice worldview does not make sufficient room for the ethical significance of nations – especially for the individual citizens that comprise such collectives. Miller (1995), for instance, offers a defence of moderate ethical nationalism. He argues that "relations between persons are part of the basic subject matter of ethics, so that fundamental principles may be attached directly to these relations" (5), including the relations between teachers and students, neighbours within a village, or start-ups of like-minded co-workers. Such commitments in turn lead one to identify and perceive substantial value in their ties with their co-nationals, thereby giving rise to the special obligations we owe towards one another. These obligations do not obviate the relevance of the interests of foreign non-nationals; but they do mean that we can sometimes prioritise compatriots under conditions of scarce resources (e.g. distributing a limited supply of vaccines in a pandemic) (Zhang 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Third, global AI justice should not be *primarily* centred around distribution. When we posit that individuals should have 'equal opportunities' to succeed through AI, what *exactly* are we speaking of here? As Anderson (1999, 288) puts it, distribution-orientated accounts of global justice focus excessively on the treatment of "divisible, privately appropriated goods, such as income and resources, or privately enjoyed goods, such as welfare". Given such a fixation, how can we account for the normative status of the kinds of political and legal constellations, social norms, and values that both contribute towards and are shaped by AI?

Consider, for example, an LLM used by a large, dominant state to create purposefully discriminatory content towards particular nationalities and ethnicities. A purely distributive account may suggest that this model ought to be made equally available to smaller and less resourceful states as well, which of course misses the main point – namely, that the outputs are problematic, not that the opportunities to access it are unequal.

Indeed, justice is more than a merely distributive exercise. It is a fundamentally *justificatory exercise*, one in which individual agents have claims to normatively meaningful defences, excuses, or counter-defences concerning the events befalling them. The framing of *distribution* is too reductionist, precisely because it pays little attention to interpersonal relations, which are no less important – if not more so – for individuals, than bundles of resources or access to advantages. The neo-Rawlsian account falls short of meeting this threshold.

3.2 The Cantilevering Case for GARE

Despite some conceptual overlap, most RE theorists diverge from Rawlsian or neo-Rawlsian egalitarianism, in that they do not construe the two principles of justice as the only or best manifestations of egalitarian considerations; they also differ from distributive luck egalitarian views, over the extent to which the effects of (bad) luck should be fully addressed through redistribution of resources, opportunities, or welfare, in our pursuit of justice.

The following comprises a novel attempt to apply the tenets of RE to AI justice, using an approach that some philosophers dub *cantilevering* (Carens 2013) – namely, taking a relatively uncontroversial position (e.g. some P1) as a starting point, showing that accepting it generates strong reasons to accept a further position (e.g. P2); that accepting this further position also generates further reasons to accept another position (e.g. P3), so on and so forth. The end goal is to show that accepting P1 would give rise to a compelling justification for some P-N, even if

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

P-N may seem less-than-intuitive upon first glance. This argumentative approach allows for the justification of potentially controversial claims in a tight and rigorously analysed manner.

Relational egalitarians such as Wolff (1998), Anderson (1999), and Scheffler (2015) fundamentally believe that the most important form of equality entails the opposition to problematic hierarchies that entrench:

> Relations between superior and inferior persons. Those of superior rank were thought entitled to inflict violence on inferiors, to exclude or segregate them from social life, to treat them with contempt, to force them to obey, work without reciprocation, and abandon their own cultures. (Anderson 1999, 312)

RE advocates the elimination of such hierarchies and the affirmation of equal moral worths of persons. This neither entails that everyone should possess the same bundles of resources; instead, relational equality is achieved between agents so long as they can relate to, view, and interact with one another as equals in social standing; they do not treat one another as if they were inferior, in the absence of morally compelling reasons to do so (e.g. an intentional murderer who commits diabolical harm in his wider community). Various RE advocates, most prominently Anderson (1999), tend to restrict the scope of interest to within the same nation-states – namely, citizens of the same country should treat one another as *relational equals*.

We will assume that RE is at least a plausible interpretation of responsibilities we owe towards one another, and that – as a corollary – it is non-reducible to distributive or straightforward Rawlsian egalitarianism (Anderson 2010). It is beyond our scope to defend it. There are then three steps to the following argument – which builds up and extends RE to establish GARE:

Step 1. Those who endorse RE should also endorse Global RE (GRE).
Step 2. The best means of understanding GRE is to view it as orientated around individual citizens, not states; even if states can be important vehicles for equalising these individuals' relational standing.
Step 3: The citizen-centric version of GRE should be applied to AI; AI is not exempt from the requirements of GRE (GARE).

Step 1

Some egalitarian theorists hold that egalitarian responsibilities should apply strictly within one's nation-state – Wellman (2011, 75) observes that "the coercion states invariably employ is nonconsensual, and, as such,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

is extremely difficult to justify […] Thus, states are justified insofar as they provide vitally important benefits (i.e. protecting their constituents' human rights) that would otherwise be unavailable." In receiving benefits from one's state and both partaking in and being impacted by the same system of governance, co-citizens of the same state thus accrue egalitarian responsibilities towards one another. Wellman does not believe the same could be said of non-citizens.

Yet this coercion-based justification *against* broadening the scope of egalitarian concern runs into an objection launched by, for example, Abizadeh (2008), Nath (2011, 2015), Sharp (2022), and, to an indirect extent, Pogge (2008). States can exercise coercive influence – if not control – over foreign populations, via border-immigration control, economic sanctions (Pattison 2015), trade and investment policies, and even the propagation of affirmative or detractive discourses in relation to these populations' governments.

Indeed, Nath (2011) compellingly addresses the ways in which interstate inequalities can impact states' abilities to "shape the terms of cross-border interaction". We care about eliminating inequality – and the advancement of broader equality – between individuals of the same state, because their "subjection to institutional terms that account for such interconnectedness is, for the most part, unavoidable and nonvoluntary". The dearth of choices in the domestic setting gives rise to relational egalitarian demands; the same could be said of globally coercive structures – from the sweatshop economy to large multi-national corporations, from the unequal playing field of trade negotiations to the weaponisation of capital restrictions and exporting of critical raw materials, for example, rare earths and lithium, copper, and nickel, for geopolitical ends. Coercion extends beyond borders; Wellman's argument is therefore unpersuasive.

There are three reasons for relational egalitarians to endorse GRE. First, there is no *inherent reason* to limit the scope of justice to simply within one's national boundaries. That there exist *normatively significant distinctions* between how one relates to a fellow citizen, and to a non-citizen, *does not* eliminate the relevance of the latter's claims on one's duties towards them. Accepting that our responsibilities towards co-citizens are stronger does not thereby imply we accrue no responsibilities towards non-citizens. The Wellman-Nath dialectic above is replicated across a number of purported differentiating factors between citizens and non-citizens – for example, contribution towards welfare (non-citizens also contribute greatly towards one's domestic economy, life chances, and wealth), or strength of association (select citizens may associate more

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

closely with non-citizens in ways they find meaningful). See, for instance, Ip (2016), who advances a relational view of egalitarianism anchored in the normative significance of equal relations as the basis for global justice.

Second, most human beings are trivially involved in activities that would render them connected to a myriad of institutions with the potential to exert force upon them, cutting across nation-states and legal-territorial boundaries. This range of institutions is bolstered by trans-border interconnectivity (Pogge 2008).

Third, the spread of globalisation throughout the two decades between the Berlin Wall's fall and the Global Financial Crisis led to the destabilisation of once-rigid conceptions of national citizenship and identities among many, with a survey of over 20,000 people from eighteen countries in 2015–2016 revealing that 51 per cent of respondents reported seeing themselves more as 'global citizens' than 'citizens of their country' (GlobeScan 2016).

This phenomenon was particularly apparent in countries with emerging economies and large populations, such as India, China, and Nigeria. While subjective attitudes are normatively non-determinative, they do offer a glimpse into the emergence of an international order that both presents an alternative to and contextually supersedes the nation-state as the definitive anchor for individuals' self-understanding. In short, the nation-state is by no means the only political *community* that binds individuals' self-identification.

Step 2

The next claim is that the crux of GRE lies with individual citizens, as opposed to nations or states. Some may advocate a key role for nations and states within the RE conversation, given their functions as the primary political unit of social organisation, resource distribution, and domestic and international representation. Bernard Williams (2005, 3, cf. Sleat 2010) proposes the *first political question*, one that "is the condition of solving, indeed posing, any others", with the securing of "order, protection, safety, trust, and the conditions of cooperation". We have also seen convincing reasons for which nations should matter in our understanding of global justice. If an inter-*national* egalitarian theory makes no room for the state or nation, how complete could it be?

In response, a distinction should be drawn between the *foundational* and the *instrumental*. The instrumental significance of nations and states can be embraced by those subscribing to the view that foundationally, it is the rightful claims of *individuals*, not states or nations, to RE that ought to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

motivate our normative prescriptions. Advocating that states or nations are entitled to relationally egalitarian treatment apparently requires confronting more contentious questions on group and collective agency, which would be beyond our scope. Regardless of how geopolitically powerful or weak, wealthy or poor, large or small a state is, an individual is *pro tanto* entitled to being treated as equal by others from other states.

Indeed, achieving – or approximating – relational equality between individuals of different states, may practically require the elimination of some inequalities between their very states. States are instrumental intermediaries in determining the range of opportunities, the legal and regulatory foundations, and the average level of resources accorded to their own citizens. A world where powerful states can radically suppress and exclude weaker counterparts is unlikely to be one where the citizens of both sets of states can relate to one another as equals (Fanon 1963; Pogge 2008).

Step 3

Why, then, should GRE apply to AI?

Consider the following two positions:

1. A: AI is uniquely distinct from most other spheres of human activity, to which global RE applies; as such, AI should not be governed by principles of GRE.
2. B: AI is not uniquely distinct from most other spheres of human activity, to which GRE applies, in a way that renders it *exempt* from RE considerations.

B is more plausible. Unlike the family, which may reflect deeply intimate attachments integral to individuals' self-conceptions as autonomous moral agents, anchored in their authentic personal values, commitments, and relations (Williams 1973, 1981), the sphere of global AI *opportunities* is one that definitively does not involve prospective violations of integrity for the individual. The pursuit of GARE does not require that one give up vitally important personal ground projects.

One might reply that AI should be integrated into our understanding of GRE. Yet precisely because it can be subsumed accordingly, relational inequalities within AI can be offset through a heightened push for egalitarian transformations along other, non-AI dimensions. Perhaps the fact that a country X's government causes their citizens to not relate with citizens in country Y as equals, through imposing export bans on GPU chips, data, and AI talent outflows to Y, can be offset by X's government including Y's citizens within decision-making processes

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

concerning developmental loans and financing. The fungibility with other spheres of justice cannot sufficiently obviate the specific case for GARE for two reasons:

First, RE is a holistic concept. Suppose a philosophy department chair Joseph opted to systematically exclude a colleague, Susan from all social events, public discussions and dialogues, and inter-departmental meetings – on dubious grounds that she is not 'suitable' for such engagements. Yet in his capacity as a senior academic, Joseph offers to co-author an important paper with Susan, crediting the latter fully for her conception by making her the first author. We would not deem the respectful and egalitarian treatment in the paper-writing context as adequate in offsetting the disrespectful and inegalitarian treatment in the administrative context.

Second, to the extent that it is possible to enact principles of RE within the sphere of AI – through the measures outlined earlier – there exists no particular value in the preservation of such relational inequalities, such that they can only be offset indirectly. Why should we not tackle inegalitarianism across both AI and non-AI spheres?

Contingently, given AI's increasing importance and necessity in individuals' lives, the pursuit of greater relational equality in AI development, diffusion, and usage could well require that we prioritise AI in the proverbial *order* of priority. A world where access to AI usage, but also capacity to shape norms, regulations, and accountability, is increasingly stratified, is a world where individual citizens are unlikely to see and relate to one another as equals, at least in relation to their ability to derive benefits from the digital economy.

3.3 Three Key Implications

What does GARE *actually* imply? What is *relational equality* in the context of AI?

While answers to this question will vary by context, the following comprise three key implications arising from GARE, drawing upon different understandings of what the exact core of "securing egalitarian relationships" comprises (Bidadanure and Axelsen 2025).

Eliminating Status Hierarchies

An oft-highlighted component of RE is its attempt to eliminate oppression, where the victim suffers from "some inhibition of their ability to develop and exercise their capacities and express their needs, thoughts,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

and feelings" (Young 1990, 40). This can be traced to its broader commitment to tackling *status hierarchies* (Voigt 2020). Anderson (2012) argues that RE is opposed to problematic hierarchies of domination, esteem, and standing.

While domination (and oppression) have been extensively discussed as manifestations of structural injustices, more should be said about esteem – a form of recognition granted by others based on the individual's virtues, actions, or other admirable merits – and standing – the extent to which the individual is an agent of concern, respect, and accountability in the eyes of others. Only when undue inequalities across esteem and standing, along with structural relations giving rise to domination, are minimised can relational equality be attained.

In the context of AI, where actors' production and regulation of AI manifest in overt domination (per Section 2) or unwarranted inequalities in esteem and standing, there thus exists a strong GRE case on two levels: trivially, for these actors to change their actions; and more substantively, for third parties to *positively intervene* in limiting such inequalities. Refraining from contributing towards structural injustices alone is insufficient.

For example, if an AI conglomerate backed by a dominant state amasses copious volumes of human capital and technology, such that it effectively crowds out alternative competitors on the global market, even if it ends up deploying its resources towards wholly benign ends, its monopolistic nature enables it to exert dominant market influence. Such an immense power asymmetry between the company and its rivals can be conceptualised as domination, reflecting a problematic hierarchy of esteem and standing – not only relative to its fledgling and struggling competitors, but also the millions of consumers beholden to the whims of its leadership and the government behind it. Regulators across the world would have the *pro tanto* responsibilities to pursue anti-trust regulation, to preclude the conglomerate from accruing extremely concentrated power. As of 2022, just 100 companies – mostly in the US and China – comprised 40 per cent of global AI research and development (UNCTAD 2025). Leading players in the US and Europe, including xAI in the former and Mistral AI in the latter, have come under growing scrutiny for their 'weak' risk management practices, despite having amassed significant capital and market influence (Chow 2024).

The elimination of status hierarchies would of course have a distributive component. When citizens residing in Country X are denied access to instrumental components for cutting-edge LLM development, or

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

integral technologies synthesising AI and infrastructural planning, they are certainly unlikely to relate to their more resourceful counterparts in Y and Z as equals. They would be subjugated to another manifestation of the long-standing digital divide between the Global North and South (Heeks 2022).

Some authors (e.g. Zarakol 2017; Bell and Wang 2020) point to the value of hierarchies in shaping the international order. Hierarchies provide stability of expectations and support enduring leadership on the part of great powers. Similarly, one could argue that AI requires global hierarchies in order to be governed effectively, lest more unbridled competition between a large number of AI state and non-state powers take over in a world where there exists no overwhelmingly dominant player in its development or regulation.

Yet limiting hierarchies in the global AI space does not imply endorsing a state of anarchy across states and individuals within them. Nor, indeed, should we commit to the view that *all* relational hierarchies must be eliminated. Benign and just hierarchies can and do exist (Bell and Wang 2020). However, such rationale cannot be invoked to defend the extensive restrictions on the exporting of cutting-edge semiconductors, arbitrary stipulations targeting AI investments, and radical financial capital inequalities as witnessed in the status quo, which leave the less advantaged parties at the whims of the more advantaged.

In the face of AI status hierarchies without a compelling, normatively sound justification, all parties with the capacity to do so should work towards ameliorating them. This could take the form of collectively resisting or circumventing geopolitically motivated export controls, advocating more transparent sharing of model parameters and weights in LLM training – for example, open-source AI development (Bisht, Yee and Roberts 2025), and providing socioeconomically disenfranchised groups within countries with both access to AI and AI literacy training. The latter connects to a broader point about capacity-building – individuals cannot relate as equals with one another if they possess radically divergent capabilities, as well as access to the minimum threshold of functional knowledge required to derive benefits from AI.

Equalising Decision-Making Power Over Global AI Regulation

An alternative focus adopted by some RE theorists concerns ensuring that all agents possess equal decision-making power – that is, that their preferences and interests are weighed equally, considered carefully, and incorporated duly in the process by which significant decisions

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

are made. Scheffler (2014, 25) conceives of relational equality as when individuals conduct their relations "on a footing of equality [such that] each person accepts that the other person's equally important interests [… including] the person's needs, values, and preferences – should play an equally significant role in influencing decisions made within the context of the relationship".

He further substantiates this by calling for reciprocal acknowledgement of all parties in the relationship of one another's interests as "playing just as significant a role as [their own] in constraining [each other's] decisions". Such acknowledgement should give rise to dispositions that bind their joint decisions, through serving as an *egalitarian deliberative constraint* (Elford 2017). In the context of AI, we could see this evolve through discussions over global regulations and policies governing its development, diffusion, and use.

GARE thus encourages the establishment of multilateral frameworks in which the voices of small and medium states, but also representative non-state actors, are afforded a voice and endowed with impactful agenda-setting powers regarding AI governance. Progressively, the inclusion of more voices in the AI decision-making process can be accomplished through in-house reforms within the UN, World Bank, and the International Monetary Fund (IMF), whereby these states' representatives are given more autonomous say and political platforms to robustly comment on and advocate AI agenda on their citizens' behalf. This process could also be triggered by the efforts of more progressive lobbying and advocacy groups pressuring corporations to open up to genuinely consulting and incorporating views of representatives of the Global South and under-represented political communities in their development processes.

GARE can, therefore, go some way in identifying key expectations benchmarks for which actors in the global AI space ought to be accountable, specifically in areas in which such global deliberation is procedurally and instrumentally significant, including the management of AI use in military conflicts, combating AI-generated misinformation and hateful content, and tackling the use of AI-based algorithms in public health and humanitarian contexts. Some of these are issues tentatively discussed under the UN AI Panel and Dialogue (Resolution A/RES/79/325 (UNGA 2025)).

During this process of deliberative inclusion, we may find states' practically desirable proxies representing their constituencies' interests internationally. First, it is over-demanding to expect all their constituents to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

play an active role in steering and contributing towards a structured and outcome-oriented global conversation; second, there is value in having states serve as the initial filtration mechanism for the demands and interests of their citizens. Even in non-democratic states, where popular participation in elections is often lacking, select governments would draw upon consultative processes to aid them in gauging public interests and concerns over AI. As discussed by Sheehan (2023) and Zhang (2024), public opinion plays a partial – albeit important – role in shaping the bureaucratic processes in the Chinese state, as it iterates and devises regulatory frameworks for AI. Notably, non-state actors – ranging from important business associations, corporations, and investors, through to activists and citizens' collectives with significant interests in AI regulation – should also be given a robust chance to influence global regulatory principles (Li 2023; Diya 2025).

Ensuring an Ethos of Respect with AI Deployment

The third variant of GARE advocates that international AI deployment must develop and maintain a genuine *ethos of respect*. Wolff (1998) discerns between two distinct strands of the egalitarian ethos – one is fairness, which, per luck egalitarian intuitions, requires individuals to bear the costs of their choices and for the effects of bad brute luck to be eliminated; the other is respect, whereby everyone should be treated with acknowledgement of their equal innate worth as persons. Wolff suggests that these two values come into conflict, and that in circumstances where individuals' self-respect is at stake, respect should be prioritised over fairness. In short, both Wolff and Anderson (1999) believe that luck egalitarianism misses the core of what makes egalitarianism so desirable and forceful: ensuring that individuals are not treated in ways incommensurable with their status as autonomous, self-conscious members of shared political communities.

As applied to AI, there exist clear responsibilities on the part of all capable actors to develop AI in a way that does not undermine the respect of individuals. A generative AI model producing positive stereotypes about ethnic majorities (e.g. Hagen, Huo, and Nguyen 2025), yet implicitly excluding ethnic minorities from its outputs, may not be contributing towards structural injustices *per se*, though it can certainly be interpreted as reinforcing disrespectful attitudes towards the minorities.

Consider, for example, the case of Cambridge Analytica. The now-defunct analytics firm drew upon machine learning models to target

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

and influence voters in elections, including the 2016 US Presidential Election. In seeking to optimise the receptiveness of voters towards advertisements, the firm "built an AI based system that was able to automatically test hundreds or even thousands of variations of an ad before deciding which one to place." (Delcker 2020) The company undoubtedly used AI to wield and manipulate prospective voters' perceptions of facts (Rubera 2023). The reining-in of AI manipulation and deception (Carroll et al. 2023) is thus a foremost priority in the advancement of RE. Indeed, this normative concern also underpins the European rules concerning AI in political advertisement (Stephenson, Rinceanu, and Boverman 2025).

Section Conclusion

We have established that on a global level, the opportunities generated by AI are best understood not through neo-Rawlsian distributive lenses, but through a distinctly relational egalitarian approach. GARE entails three important implications, concerning status hierarchies, decision-making, and respect. Conforming with these implications could require us to undertake significant regulatory transformations, which we shall explore in greater detail in Section 5. Before then, however, we shall turn to a game-theoretic treatment of AI dynamics on a global scale.

4 A Game-Theoretic Perspective on Cooperation

The case advanced thus far in favour of reducing relational inequalities in AI access can also be made on the basis of the self-interests of countries. The following section advances the rational choice argument in favour of greater openness to cooperation, multilateralism, and hence a 'flatter' AI world.

The geopolitical relationship among countries with respect to AI development and positioning is often described as a non-cooperative enterprise or, more forcefully, a 'competition' or 'race'. For example, the US White House AI Action Plan, promulgated in July 2025, is entitled "Winning the Race" (The White House 2025). Its preamble reads in part as follows:

> Today, a new frontier of scientific discovery lies before us, defined by transformative technologies such as artificial intelligence … As our global competitors race to exploit these technologies, it is a national security imperative for the United States to achieve and maintain unquestioned and unchallenged global technological dominance. (i)

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Statements like this often invoke historical analogies of great power competition. For example, the White House document makes extensive references to America's "competitors" (5), "adversaries" (8), and "countries of concern" (21). Such rhetoric invites natural comparisons to America's Cold War rivalry with the Union of Soviet Socialist Republics (USSR), including the arms race to develop nuclear armaments as well as the space race to reach the moon. Yet such descriptions can be misleading, and ultimately counter-productive even to American interests.

In this section, we will explain why this is the case from a game-theoretic perspective. Our core argument will be that, in stark contrast to the space or nuclear races, the purported competition over the development of frontier AI models should be more fruitfully viewed as a cooperative, non-zero-sum multilateral enterprise – a cooperation game among disparate actors, even ones that may harbour antagonistic attitudes towards one another at large yet would find it consistent with their interests to work together.

Let us begin with a brief preface of game-theoretic modeling. Game theory comprises the formal and mathematical study of strategic decision-making among rational agents, where the outcome for each participant is contingent upon the actions of others. In general, game theory models begin by making assumptions about the preference order of all players over the outcomes available to them, where the latter are determined by the joint actions of all players in the game.

An example of an outcome would be that Alice and Bob go to see a movie together; the underlying joint *action set* that makes this possible is Alice going to the movie theater, and Bob going to the movie theater, and their *preferences* would correspond to some assigned utility values (e.g. how much each of them enjoys the said outcome). The core solution concept in game theory is that of an equilibrium, which we define more carefully later. In general, equilibria capture the notion of a stable point – once we reach it, neither player has an incentive to deviate (assuming they are rational). If Alice and Bob most prefer to see the movie together, then if Alice learns that Bob will go to the theater, she does not have an incentive to decide against going. Yet if Alice wanted to avoid Bob, then upon learning he will go, she should decide not to go.

Crucially, game-theoretic models assume that the players are rational; and they also assume that the players are aware of each other's rationality. In this sense, such models are approximations of where ideally rational agents pursuing their interests would end up. But they are not necessarily a reliable prediction of where *actual* agents will end up in the real world.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

For our purposes, however, this is a feature and not a bug. Our primary interest in this section, as in the rest of the manuscript, is to develop an action-guiding model that provides certain prescriptions about where states, should they rationally pursue their self-interests, will end up when it comes to inter-state AI dynamics.

4.1 The Nash Equilibrium and Prisoner's Dilemma

A *Nash Equilibrium* occurs when a joint set of all the players' actions is such that no player has an incentive to deviate from their chosen act, given the other players' choices. This is a normative concept: it suggests that when players are in Nash Equilibrium, it is in some sense irrational for any of them to deviate. The strategies sometimes take the form of probability distributions over actions. For example: in a game of Rock, Paper, Scissors, there is no fixed action that would lead to Nash Equilibrium; but there is a stochastic one – namely, to randomise uniformly among the three choices. This solution concept is eponymously named after the well-known Mathematician, John Nash, who proved in his Princeton University doctoral dissertation the existence of pure or mixed strategy equilibria for every game with a finite number of players in which each player has a finite choice of pure strategies (Nash 1951).

The Security Game

As Nash worked on his dissertation in Princeton, New Jersey, on the other side of the country were two mathematicians recruited by the RAND Corporation in California to conduct social scientific research on military strategy following World War II (Gass and Assad 2005). Merrill Flood and Melvin Dresher were interested in exploring player behaviours in repeated play of non-cooperative and non-zero-sum games, and they designed a game with the payoff structure of what we now know as a Prisoner's Dilemma. Strategic behaviours in AI development are often analogised to a Prisoner's Dilemma (e.g. Foley 2024; Perez 2025).

The Prisoner's Dilemma has been particularly influential in IR, due to its central role in security studies. It is the conceptual model underlying security strategy (Jervis 1978) – the idea that in the absence of a central authority, states acting in their own self-interest will pursue individually rational actions that could lead them to a conflict which none of them prefers (cf. Waltz 1979).

In the twentieth century, this model was often applied to the study of nuclear proliferation (or non-proliferation). Suppose we have two

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

players/nation states (e.g. the US and the USSR during the Cold War), who must decide whether to cooperate with an arms control agreement or a nuclear test ban treaty. Each state has two choices: cooperate or defect. Suppose that defections are undetected. Let us call this the Security Game.

How should we order the preferences of the two states in a security game? First, we assume that the states' interests are symmetric and that the game is played once (repeated play affects the rational strategy as defections can be punished in subsequent rounds of the game (Axelrod and Hamilton 1981)). We assume that self-interested actors seek the 'best deal' for their own nation, so defecting while the other state cooperates is best (we will assign this outcome four units of utility). The opposite of this is the worst ('you defect/I cooperate'), so we will assign one utility unit to that outcome. Meanwhile, if we both cooperate, we gain the benefit of the non-proliferation treaty or arms control agreement, whereas if we both defect, we return to the status quo, so we will assign the former three utility units, and assign the latter two utility units. Since the Von Neumann-Morgenstern utility function is unique up to affine transformations (Von Neumann and Morgenstern 1953), the choice to use a 1–4 ranking is made without a loss of generality for our purposes. This version of the Security Game therefore looks like a typical Prisoner's Dilemma (PD), as in Figure 1.

This is a symmetric game with one pure strategy Nash Equilibrium: namely, the outcome where both states defect. This simple game illustrates what we know well: two states acting rationally may find themselves in a socially sub-optimal outcome. While it is individually rational for each state to defect, it is *collectively* not ideal for them to end up in that outcome. Yet cooperation is not a Nash Equilibrium – it could leave one purportedly vulnerable to the other's taking advantage of them. In the context of a public good – something that is non-excludable

	State B: Cooperate	State B: Defect
State A: Cooperate	3, 3	1, 4
State A: Defect	4, 1	**2, 2**

Figure 1 The Security Game. A typical Prisoner's Dilemma with a unique pure strategy Nash Equilibrium highlighted in bold font.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

and non-rivalrous – each player has an incentive to free-ride. This is an instance of a non-cooperative but also non-zero-sum game.

There are aspects of global AI governance that seem to resemble a PD – nation-states, or frontier technology firms, can be modelled as having a similar incentive against cooperating as in a PD game: the more quickly that a company or nation can develop its own infrastructure or its own technological capability, the more economic, security, or pecuniary benefits it can secure. In this sense, cooperating with a competitor seems irrational. The 2025 US AI Action Plan makes this PD framing explicit:

> To succeed in the global AI competition, America must do more than promote AI within its own borders. […] America currently is the global leader on data center construction, computing hardware performance, and models. It is imperative that the United States leverage this advantage into an enduring global alliance, *while preventing our adversaries from free-riding on our innovation and investment*' (The White House 2025, 20, emphasis added).

Indeed, former US President Joe Biden implemented strict export controls on leading components required for the development of AI, including high-end chips and SME, 89 F.R. 72926 (BIS 2024a). As part of its rationale, the US Department of Commerce argues that '[e]xport controls on semiconductor manufacturing equipment (SME) and related parts and components are central to countering the PRC's goal of furthering its … production capacity', 89 F.R. 96790 (BIS 2024b). This framing has jolted Chinese manufacturers to compete by doubling down on their self-sufficient capacities, aggressively recruiting talent, redirecting state capital in a focussed manner, and identifying innovative means to maximise the utility of legacy chips (Zhang 2025).

One can, of course, contend that the drive towards self-sufficiency in fact pre-dated the introduction of US export controls – citing, for example, the 'Made in China 2025' policy (Center for Security and Emerging Technology 2022). Yet while 139 billion yuan (roughly 19.5 billion USD) were invested into semiconductors in 2014 in the initial phase of China's National Integrated Circuit Industry Investment Fund, a further 204 billion yuan (28.7 billion USD) and 340 billion yuan (47.8 billion USD) were invested in the second and third phases (L. Lee 2024). During his first term, Donald Trump launched the US-China trade war in 2018 (Bown and Kolb 2018), with his successor Biden taking a distinctly more proactive and targeted approach in kneecapping China's semiconductor industry developments (The Guardian 2024). Trump also sought to limit the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

power of the World Trade Organisation (WTO) by refusing to fill the quorum required for its Appellate Body (AB) to function (Murrill 2019), a policy which Biden also continued (Baschuk 2023).

Despite persisting challenges concerning corruption and inefficient capital deployment, the drive towards self-sufficiency in China has indeed delivered significant dividends. For instance, China's Semiconductor Manufacturing International Corporation (SMIC) has raced to increase domestic chip manufacturing capacity in an attempt to meet local demand for high-end chips required by domestic technology giants such as Huawei, Tencent, Alibaba and DeepSeek (Wu 2025). In September 2023, Huawei launched the smartphone chip Kirin 9000S, which was manufactured with a 7-nm process by SMIC (Liu 2023).

Despite the rhetoric, in many important respects, the PD framing of AI dynamics among nation-states is misleading. We can explain why by focusing on the specific reasons cited by the White House AI Action Plan: the first passage suggests that America must drive the adoption of its systems, infrastructure and standards (The White House 2025, 1). This is not an unreasonable claim for any nation-state to make – its own infrastructure outlay in, for example, small and medium countries, can benefit the investing state as its domestic manufacturers experience a growth in market share, its reputation grows across the world, and its industry standards are widely dispersed. Yet the last passage of the above quotation then goes on to say that America must prevent its adversaries from free-riding on its innovation and investment. The US Department of Commerce makes similar, and indeed stronger claims in 90 F.R. 47201 (BIS 2025a).

If global reliance on the US's innovation and investment does constitute a significant benefit for the country, as the first passage states, then such reliance on its innovation and investment does not amount to harmful free riding of the sort that drives a tragedy of the commons. Perhaps these claims can be squared by highlighting that such reliance counts as free-riding only if it is performed by adversaries but not if it is performed by allies. Yet this statement does not stand up to scrutiny – the thought that geopolitical adversaries will acquire significantly greater competitive advantages, bargaining capital, or prospective capacity to threaten coercion against the US through US-supplied technologies presupposes that the marginal technological (or productivity) gains can indeed yield a substantial increase in threat capacity and level. Present discussions have yet to point to precise ways in which ostensible 'adversaries' – for example, China – have uniquely stood to gain from US-developed AI technologies,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

in a manner that would substantively undermine American interests. If anything, attempts to curb American chip exports could well undermine the prevalence of the US tech stack as a foundation for global benchmarking and soft influence (Harithas 2025; Villasenor 2025).

Furthermore, the zero-sum framing or the PD framing assumes that advancements by all AI entities in one economy (e.g. China) must come as a detriment to the interests of the American ecosystem. One could point to DeepSeek as a counterexample. The company made its R1 model open source, and China has not blocked access to it from IP addresses originating in the US; this is despite the fact that OpenAI has restricted access to ChatGPT from IP addresses originating in China (Jiang 2024). The open-source nature of the company's models means they can be easily applied and extended by state and non-state actors within the US and beyond (Zhang and Liu 2025).

Indeed, one could even argue that our analysis of the PD model points towards the need for cooperation in inter-state AI dynamics. If we view the international institutions necessary for ensuring safe and responsible AI development as the relevant public goods, then a tragedy of the commons occurs where countries defect from such cooperation. The free-riders, in this case, are those who refuse to contribute to multilateral governance, due to which we lose out on both the mutual gains from trade that can happen when technological know-how and materials, standards, and infrastructure are shared, as well as the public goods that can only be achieved through cooperation – for example, the robust and consistent regulation of AI, the safe and transparent deployment of AI, and the reduction of barriers to the opportunities of AI for individuals lacking the financial resources and economic wherewithal to access them. These are, by hypothesis, non-excludable and non-rivalrous benefits that can be obtained from cooperation. Indeed, the ethos espoused by China's Shanghai Declaration on Global AI Governance, released at the 2024 World Artificial Intelligence Conference, focuses on global cooperation and encouraging shared governance (MFA 2024).

A Modified Prisoner's Dilemma

The preceding framing would suggest that we need to do what we can to increase the costs of defection and change the Nash Equilibrium of the simplistic PD model – indeed, this is Elinor Ostrom's approach to solving tragedy of the commons problems (Ostrom 1990).

For example, the easiest re-modelling illustration would occur when defections are punished (this could occur due to reputational harms in

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

	State B: Cooperate	State B: Defect
State A: Cooperate	**3, 3**	1, 4 - 1 = 3
State A: Defect	4 - 1 = 3, 1	**2, 2**

Figure 2 The AI-PD game. A modified Prisoner's Dilemma where defections are mildly punished with two unique pure strategy Nash Equilibria highlighted in bold font.

repeated play or through a central enforcement mechanism). If defecting carries a modest penalty of just one utility unit, that is enough to transform the game into one where cooperating becomes a Nash Equilibrium, as in Figure 2 (which we revisit later).

In China, the framing of AI governance stands in stark contrast to the US White House 'winning the race' narrative. In October 2023, President Xi Jinping announced China's Global AI Governance Initiative (GGI) at the Third Belt and Road Forum in Beijing. The Global AI Governance Initiative (GGI) describes the Governance of AI as a "common task faced by all countries" and calls on countries to "enhance information exchange and technological cooperation". The pursuit of multilateral governance via, for example, an international institution designed specifically to oversee AI development, as envisioned in the GGI, would be one way of transforming the ordinary PD matrix (Figure 1) into a game where cooperating becomes a Nash Equilibrium (Figure 2) (cf. Ostrom 1990).

Of course, there are concerns one could raise with regard to the "dexterous ambiguity" inherent within China's approach to AI and global governance (Olander and Wong 2025). Multilateral institutions require buy-in, connections, and institutional know-how and capacity to impose effective sanctions, for them to be able to meaningfully impose the penalties associated with defection. One could also criticise this proposal not on grounds of theoretical intactness, but for its empirical infeasibility. Likewise, one can observe that mere pledging of openness to global governance does not translate into a binding policy or commitment towards such mechanisms. These are valid concerns. But the general sentiment in China has been different from that in the US – though there are signs that the Trump administration is beginning to see the value of cooperation.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

4.2 The Space Race and the AI Race

While the discussion about the PD sheds light on productive directions that global AI governance can take – namely, implementing and enforcing penalties that can deter defection – we can also consider an alternative framework to better understand the strategic interaction between China, the US, and other players in the global AI ecosystem.

Consider now a different historical analogy – the space race between the US and the USSR, with a particular example being getting the first man onto the moon (between 1950 and 1969). A game like this will likely produce one winner (Neil Armstrong, from the US, was the first human on record to land on the surface of the moon). If we simplify a little and assume that there are two possibilities for each country (either they can land a man on the moon first, or they can fail to do so), this further implies that being second or coming close – but not being first – is equivalent to failing (indeed, the USSR's Yuri Gagarin was the first human to reach outer orbit, but his reputation fades in comparison to that of Armstrong). In this case, being first brings substantial reputational value and domestic windfalls, while losing is quite costly.

Suppose the two choices are either to Race or to Quit, the best outcome (symmetrically) would be that I race, and you quit. The worst outcome is where I quit, and you race. But which of the following is better: a tie where we both reach the moon or a tie where we both quit? It is hard to answer this as there may be no truth to the matter. In this case, we may model both situations, as in Figure 3.

As we can see, the rational outcome in both situations remains the same: namely, the only pure strategy Nash Equilibrium occurs where both countries race to win. It does not matter how we assess the relative ranking of a winning-tie and a losing-tie, as reversing their order does not

	USSR: Race	USSR: Quit
USA: Race	**3, 3**	4, 1
USA: Quit	1, 4	2, 2

	USSR: Race	USSR: Quit
USA: Race	**2, 2**	4, 1
USA: Quit	1, 4	3, 3

Figure 3 The panel on the left-hand side depicts the Space Race model where a winning-tie is preferred, while the panel on the right-hand side depicts the Space Race model where a losing-tie is preferred. The Nash Equilibria are highlighted in bold.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

affect either country's best response. This is a pure competition game, in which both parties will race. As for whether this outcome is socially optimal, it depends on the relative ranking of ties – in the panel on the right the Nash Equilibrium is not Pareto Optimal (both parties would be better off quitting – an outcome that exists out of equilibrium), while in the panel on the left it is Pareto Optimal (the only way for one party to do better as compared to its equilibrium pay-off is to make the other worse off).

As applied to AI governance dynamics, this is the explicit framing of the US White House AI Action Plan and of the Department of Commerce export controls rationale – namely, that it is imperative for the US to win, and that the outcome is binary (hence if they do not win, they will lose) (The White House 2025; 89 F.R. 72926 (BIS 2024a)). Indeed, many in the United States have explicitly advanced a competitive frame. For example, Chris Miller argues that the US chip export controls, such as the prohibition on the sale to China of the most advanced GPUs like the NVIDIA H100s, as described in the Department of Commerce's Interim Final Rule (90 F.R. 47201 (BIS 2025a)), and related restrictions, such as the prohibition by the Dutch government, upon encouragement by the US, of the sale of ASML's Extreme Ultraviolet (EUV) lithography tools to China, have "succeeded … [by] significantly slow[ing] the growth of China's chipmaking capability" (Miller 2025). Similarly, Rush Doshi and Chris McGuire note, "[t]o maintain our lead in AI, the administration should tighten controls on equipment" (Doshi and McGuire 2025). They also assert, "sales [of NVIDIA chips] to China […] come at the expense of sales to the US and its allies. If the revenue didn't come from Beijing, it would come from elsewhere. Sales to our adversary will, however, advance its military and cyber capabilities". The main theme of this argument is that the strategic environment in AI is best captured by a *competition game* in which one has to do whatever it takes to win, since that is the only rational strategy given the unique Nash Equilibrium.

Yet this is likely a false narrative. In order to understand why, let us move to another game-theoretic environment, one which will serve as a stepping-stone to our eventual conclusion.

Hawks and Doves

In the competition game set out earlier, there is a unique Nash Equilibrium characterised by both parties deciding to race. This is because if I race, you can race too – i.e. we can both attempt to develop our own AI ecosystems and I cannot stop you from developing yours.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

	Player B: Drive Straight	Player B: Swerve
Player A: Drive Straight	1, 1	**4, 2**
Player A: Swerve	**2, 4**	3, 3

Figure 4 The Hawk-Dove game with two unique pure strategy Nash Equilibria highlighted in bold font.

And while we cannot both win – in the sense of being first – we can certainly both reach the moon. Now consider a different kind of race – one where the underlying shared resource is strictly rivalrous. This is best illustrated via the following example:

A and B are racing towards a tunnel and each of them must decide whether to drive straight or swerve; if A drives straight and B swerves, A wins (and vice versa); if both swerve, they both lose; if they both drive straight, they both die. Naturally, a losing tie is preferred. This game thus has two, non-overlapping Nash Equilibria as described in Figure 4.

This game, known as the Hawk-Dove game (Maynard Smith and Parker 1973) is a more extreme version of the earlier competition game, because now the good in question (access to the tunnel) is strictly rivalrous (we cannot both fit), yet it is non-excludable. It is also a puzzling game: if we both decide to swerve, then we are out of equilibrium – at that point, either of us would be better off by deciding to drive straight. Yet once we both decide to drive straight, we are again out of equilibrium and one of us should swerve.

Notice that the equilibria in this game occur where I do one thing and you do the other. We benefit by *not* working together. Perhaps an advocate against more AI cooperation and parity across countries, then, could draw an analogy between AI and the street race: in the specific context of AI development for military purposes, if a leading country unilaterally places restrictions (i.e. 'swerves') concerning the level of performance exhibited by AI models, for the fear that they could produce deleterious consequences if not aligned properly, some may worry that its geopolitical rival can march ahead by 'driving straight' – namely, in applying a brakes-off approach to AI development.

Should both countries cease innovating with their AI, however, they could deprive their populations of its potential opportunities. And should both countries fail to apply any restrictions to their AI development, we could witness substantial *interactive risks*, as discussed in Section 2.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Despite these ominous conclusions, there are reasons to think that it is possible to engineer AI governance dynamics in a way that renders it disanalogous to a Hawk-Dove game. We shall see why shortly.

4.3 The Case for Relationally Egalitarian Cooperation in AI

We now have the theoretical blocks to make the substantive case for cooperation over AI consistent with the principles of Global AI Relational Egalitarianism that we developed in Section 3 – taking the forms of reducing barriers to access to cutting-edge technologies for actors from other states, and empowering a much wider range of partners in setting agenda and influencing laws within multilateral institutions, as well as improving the availability and accessibility of AI to ordinary citizens. Let us begin by examining the problem of free-riding.

Free-Riding as a Problem?

In the case of nuclear disarmament, for example, the optimality of free-riding is obvious: if I can get you to eliminate your nuclear arsenal, while I keep mine, which I have already invested in, that creates a low-cost security buffer for me. I would rather keep the optionality of holding on to my nuclear arsenal 'just in case' – while you commit to reducing yours.

Yet it is not clear what free-riding amounts to in the context of interstate AI dynamics. If the choice between two states is to share or not share certain frontier models, technologies, or the components required for their development, then refusing to share while the other party shares is not necessarily optimal: you then lose the ability to disseminate your own platforms, standards, raw materials, and modes of development.

The underlying good – which may be construed as the frontier AI models and systems, for instance – is largely non-rivalrous (we can all use the same model, so long as it remains openly available (e.g. is made open-source, and the use of which is charged at a minimum or implemented locally). While the marginal user places a non-zero strain on the underlying data centres and energy resources used to power that model when they are not run locally, such additional cost could be shared yet it is excludable (I can choose to block your access to it). Indeed, we have seen many examples of exclusion; at the market level, OpenAI's decision to block access to its models from IP addresses originating in China (Altman 2024); and at the state level, the US Department of Commerce's decision to block exports of leading GPUs (BIS 2024) or Congress's decision to in effect force ByteDance's divestiture of the famous social media application, TikTok (138 Stat. 955, ByteDance 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

As such, this is the reverse of the Hawk-Dove game: from any given state's perspective, we have largely *non-rivalrous but excludable goods* instead of *rivalrous but non-excludable goods*. While US-China examples are plentiful, the Dutch government faced a similar choice as to ASML: they can exclude countries from purchasing EUV lithography equipment, while any given country's use of such equipment is not strictly inconsistent with another country's use of it (of course if one machine is in Germany, that same machine cannot be Italy – but the point is that Germany's ability to purchase ASML's EUV lithography system is not incompatible with Italy's ability to purchase one as well, unless it were unduly difficult for ASML to manufacture both (which is not the case). We have also seen examples of non-exclusion and collaboration: DeepSeek R1 is a largely open-access model which anyone can use, while recently Alibaba and Nvidia agreed to integrate Nvidia's robotics and embodied AI tools on Alibaba Cloud (Chow 2025). When we have goods that are excludable but largely non-rivalrous, it is advantageous for both parties to cooperate since they can share in the resource, yet it is likewise tempting to not cooperate since exclusion is a viable strategy.

In particular, if we have two states facing a choice between sharing an AI resource, such as a frontier model, or excluding the other party from that resource, then the best outcome occurs where they both choose to share, thereby capitalising on the benefits from cooperation and mutual gains from trade and joint governance. The worst outcome is where the target state retreats into isolation while its competitor reaps all the gains from cooperation – the reverse of this is second best, for the target state, trailing to the cooperative outcome. The second worst is where both parties descend into isolation.

That leads to the following game-theoretic setup (Figure 5).

This is in some sense the anti-Hawk-Dove game, and it is also similar to a PD game where defections are punished, as in Figure 2: there are two pure strategy Nash Equilibria, but they occur where we coordinate and take

	State B: Share	State B: Exclude
State A: Share	**4, 4**	1, 3
State A: Exclude	3, 1	**2, 2**

Figure 5 Cooperation. A typical Stag-Hunt or Assurance Game with two pure strategy Nash Equilibria highlighted in bold font.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

the same action instead of where we perform the opposite acts. Readers familiar with game theory will recognise this game as the 'Stag Hunt', also known as a trust game, assurance game, or more generally cooperation game. Coincidentally, the name comes from Jean-Jacques Rousseau, who discusses such a situation in his *A Discourse on Inequality* (1755). Rousseau uses the parable to explore pre-social humans organising a hunt; whereby they face a choice of hunting a bigger prey together (the stag) or chasing a smaller prey alone (the hare). The lesson for Rousseau is that cooperation is possible, but it is also fragile because defection is tempting. Embedded in this is a criticism of Thomas Hobbes' social contract tradition; for Hobbes, the state of nature is characterised by a war of all against all (Hobbes 1651).

This dual situation – the possibility of cooperation together with its fragility – is perfectly captured by this game as it contains two pure-strategy Nash Equilibria, one of which is a Pareto improvement over the other. If both parties choose not to share, they will indeed find themselves in Nash Equilibrium. This therefore poses a substantial problem; one that is closest to capturing the present world as of 2025 – where US tariffs and export controls lead to similar responses, and a tit-for-tat exchange of non-cooperative moves which leads the parties to get 'stuck' in a Pareto inferior outcome, yet because that outcome is still Nash, there is no incentive to change course. The vicious cycle then becomes self-reinforcing during extended repeated play. Yet if the parties can cooperate – they can likewise get 'stuck' in the Pareto superior Nash Equilibrium (top left), and in that case there is likewise no incentive to deviate, which can lead to a virtuous cycle that is likewise self-reinforcing, but encourages cooperative behaviour over repeated play. Indeed, the Nash Equilibrium on the top left is preferable to that in the bottom right.

The critical challenge concerns how we should get to the point where all parties in the global AI space move towards and remain in the top left. This will require, among other policy initiatives, compelling and credible international institutions to enforce and channel cooperation and commitments. In the concluding section, we examine ways in which this can be done.

It is worth highlighting that these are not the only Nash Equilibria in the cooperative game. We can identify a mixed strategy equilibrium given the framing in Figure 6. Let p represent the probability that State A *cooperates to some degree* (i.e. *shares* at least some AI technologies with others) and 1-p represent the probability that it *does not at all* (i.e. *shares* no AI technology with others). Similarly, let q represent the probability that State B cooperates and 1-q represent the probability it does not. It

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

	State B: Share (q)	State B: Exclude (1-q)
State A: Share (p)	**4, 4**	1, 3
State A: Exclude (1-p)	3, 1	**2, 2**

Figure 6 Cooperation. This is functionally the same game as in Figure 5, but with probabilities of each action expressed in terms of p and q.

is rational for each player to adopt a probability distribution over pure strategies when the expected payoffs of their pure strategies are equal.

Now we can ask: what is the mixture of $p/1-p$ *(and* $q/1-q$*)* that makes State B (and State A) indifferent in the sense that its expected payoff from sharing is equal to its expected payoff from not sharing? This will occur when $p = 0.5$ and $q = 0.5$. Therefore, if both countries act perfectly arbitrarily, that would be a Nash Equilibrium.

This too is something that we partially observe in real life: if one state's actions seem wholly erratic and unpredictable – i.e. approximately random – then other states will have an incentive to behave in a way that is equally unpredictable – i.e. to randomise or at least behave as if they are randomising. Yet even though behaving arbitrarily is a Nash Equilibrium strategy, it is Pareto inferior to cooperating. When $p = 0.5$, then the expected payoff is 2.5 (and when $q = 0.5$, then it is likewise 2.5). Hence, randomising is a Pareto improvement over both parties refusing to share, but it is still Pareto-inferior to both parties sharing. This nicely captures Rousseau's parable of the fragility of trust and the difficulty of sustained cooperation.

We even see evidence of something like this not at the state level, but rather at the industry level, and industry choices are increasingly important as large tech companies increasingly influence states' foreign policies. For example, OpenAI's Sam Altman argues that sharing with China would undermine *democratic* values (Altman 2024); hence, from his perspective, the absolute worst thing to do is to share. This is even more extreme than Figure 6. If we revise the preceding game with the assumption that sharing is worst, regardless of what the other party does, and excluding is best, regardless of what the other party does, then (from Player A's perspective) the situation looks like this:

In this case, the unique pure strategy Nash Equilibrium occurs when Player A does not share while Player B shares (Figure 7 bottom left). Unsurprisingly, this game has no mixed strategy Nash Equilibria. This captures OpenAI's argument in Sam Altman's Wall Street Journal op-ed;

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

	Player B: Share (q)	Player B: Exclude (1-q)
Player A: Share (p)	1, 4	1, 3
Player A: Exclude (1-p)	**4, 2**	4, 1

Figure 7 Stubbornness. This is the same game as in Figure 6, except now Player A most prefers not to share and least prefers to share.

the belief that upholding democratic values enjoins one to *never* share with an adversary means that one's preferred strategy is unresponsive to what the other player does.

Yet what if *mutual non-cooperation*, where both states retreat into isolation, *is the worst outcome* – as compared to unilateral cooperation?

Let us turn to the industry level again and take DeepSeek and OpenAI as an illustrative example. One could argue that it is ultimately better for DeepSeek, and in turn overall for China, to remain open than to retaliate against OpenAI, because their model, and in turn China's overall approach to AI governance, gains increased exposure and both DeepSeek and China's approach to AI governance gain the opportunity to become entrenched as the dominant model and framework through expanded access. Indeed, the US is one of the top five user countries of DeepSeek (Kumar 2025). Additionally, given the extensive discussion in Sections 1 and 2, we can see that the costs of non-cooperation are significant and will likely spiral over time through lock-in and path dependence (Hebbar and Hubinger 2022; M. Wong 2025).

We can generalise this suggestion – namely, that unilateral cooperation is better than mutual non-cooperation – to state-level dynamics more generally. The following is in effect a version of the classical argument for free trade. Consider a situation where China knows that the US will refuse to grant export licences for leading SMEs, and in return, China must decide whether to allow exports of critical minerals, in particular its large stock of rare earth elements (REEs). If China allows exports of REEs while the US refuses exports of SMEs, then China can continue to develop its own manufacturing expertise while building goodwill and critical relationships with trading partners. This could plausibly be better than punishing the US' non-cooperation. Similarly, regardless of whether China controls the export of REEs, there is a tremendous economic benefit to the US in allowing the exports of SMEs – indeed, its companies can thereby become global leaders in cutting-edge technology. In doing so, they can capture

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

more of the global market, develop the relevant know-how overseas, and position the US to lead on effective governance and regulation.

One may be tempted to respond that there is a national security threat to the US if it unilaterally exports SMEs. Indeed, this has been the argument throughout much of the legislative activity leading to export control restrictions on SMEs in the US, including in the Interim Final Rules of 2022 and 2023, as described in Section 1 (IFR 2022; 88 F.R. 73458 (BIS 2023)).

Dario Amodei articulates this argument in an essay published in January 2025, claiming that export controls are essential in slowing down China's development and ensure that the US wins the AI race (Amodei 2025). Yet the speciousness of that argument should be clear now. The entirety of this section has argued that inter-state AI Dynamics should not be viewed as a race; if it does not make sense to speak about racing to win, then it does not make sense to privilege slowing down an opponent at the expense of economic, multilateral, and other cooperative gains.

Had the US continued selling leading Nvidia chips to China, and indeed had they not pressured the Netherlands into refusing to sell EUV lithography equipment to China, would we be in a worse world than we are now? We do not think so. Indeed, we think it would have been a better world. The US could have reaped tremendous economic benefits, its high technology industry could have grown at a higher rate, and the market capitalisations of the big five US tech companies could be much higher than they are now, many American startups could have been able to benefit from a large overseas market, and US financial markets, as well as the many institutional funds tied to them, could have done better (while arguing over counterfactual predictions can be treacherous – such a possible world is not implausible, and is indeed likely). Hence, far from the doom and gloom scenario that Amodei tries to paint, we see a lost opportunity not only from mutual gains from cooperation, but even from unilateral non-cooperation. Co-operating can be persuasively seen as the dominant strategy regardless of what one's opponent does.

By contrast, when it comes to the space race or the race to develop a nuclear weapon, there were distinctly apparent upsides associated with one's victory. The US, for example, was able to secure a decisive victory against Japan in World War II in part due to its possession and willingness to use the atomic bomb (Alperovitz 1995; Maddox 2004). Yet this kind of unique gain from unilateral victory simply does not exist in inter-state AI dynamics, given the excludable yet largely non-rivalrous nature of the significant upsides.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

	State B: Share (q)	State B: Exclude (1-q)
State A: Share (p)	**4, 4**	3, 3
State A: Exclude (1-p)	3, 2	1, 1

Figure 8 Cooperation. This is the same game as in Figure 4, except now unilateral cooperation is preferred to mutual non-cooperation. The Nash Equilibrium is highlighted in bold.

One may criticise our discussion for reducing the policy debate into cooperation and non-cooperation. Of course, the real world is not binary. Instead of referring to two actions, we could consider a continuum of possibilities that the US can choose from between eliminating all export controls and refusing to export any SMEs. Yet the binary framing helps to clarify our intuitions without much loss of generality.

Let us consider further what the preceding discussion implies from a game-theoretic perspective. Under the assumption that unilateral cooperation is preferred to mutual non-cooperation, we have the following framing, where the two choices are to, for example, share/refuse to share technology/resources (Figure 8):

This game has one pure strategy Nash Equilibrium, which is to share, and it does not have any mixed strategy Nash Equilibria.

If we grant that mutual isolationism is least preferred, then indeed cooperation is the unique Nash Equilibrium, in either pure or mixed strategies. The core point is not technical, but rather a normative point about preference ordering, and we have tried to argue for it in the preceding paragraphs. As we said at the outset, the concept of Nash Equilibrium is itself normative: this is not a prediction about what states will do. It is a claim about what states should do. If we see states behaving otherwise, then either they are irrational or they have calculated the payoffs differently. Hence, as scholars, we can either (a) encourage them to act more rationally or (b) encourage them to revise their payoff matrix. For example, we may publicly advocate that in developing its export control rules in the Interim Final Rules, the US Department of Commerce has exaggerated the benefits of export controls while overlooking its full costs.

Section Conclusion

To summarise the core lessons of this section: we have argued that instead of seeing inter-state AI dynamics as a prisoner's dilemma, or as a race,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

we should instead view it through the lens of a stag-hunt: a cooperative game with two Nash Equilibria. This reframing allows us to understand how states can get caught in a non-cooperative state, while also illuminating the existence of a mutually beneficial, cooperative path. Game theory cannot on its own push players to choose one socially preferable equilibrium over another, less socially preferable one (by hypothesis, both are equilibria), but the Stag Hunt model does help us to understand what it takes to encourage states to take the cooperative path. In short, we have to somehow nudge actors in the direction of cooperating. This can happen either by directly changing their beliefs about what others will do, or by gradually encouraging such a belief change via repeated interaction and engagement. In the next section, we will focus more carefully on developing these two paths to cooperation and articulating policy recommendations that follow from them.

We thus complete our game-theoretic analysis. From a strategic perspective, states in fact have strong reasons to cooperate, to open up, and to share in AI development and resources. From a strategic perspective, inter-state AI dynamics are not a PD, and they are not a competition or a game of chicken; rather, they are often best captured by a cooperation game where we have significantly non-rivalrous but excludable public goods, and the choice is up to individual states (and actors within them) as to whether to cooperate and reap the benefits or retreat into isolationism settling for a Pareto inferior Nash Equilibrium.

5 Conclusion and Implications

The previous section shed light on the existence of multiple Nash Equilibria – a payoff-optimal ('better') collaborative one, and a risk-optimal ('worse'), individualistic one – in the context of AI and geopolitics modelled (as we argued it should be) in the form of a Stag Hunt. Our redescription of inter-state AI dynamics as a Stag Hunt and not a PD or race thus allows us to think through this cooperative game and consider how to improve the prospects of reaching the payoff-optimal Nash Equilibrium in the real world: how to steer both state and non-state actors towards the better, collaborative Nash Equilibrium.

5.1 From Hobbes and Hume to Global AI Governance

Brian Skyrms has extensively developed the implications of the Stag Hunt for the emergence of norms and the possibility of cooperation. Skyrms writes: the fundamental question in the Stag Hunt is "[h]ow can you get

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

from the noncooperative hare hunting equilibrium to the cooperative stag hunting equilibrium?" (Skyrms 2004). This is indeed the fundamental problem for inter-state AI dynamics.

Unlike in the PD, what is rational for each player in the Stag Hunt depends on their beliefs about what the other player will do: it is best to hunt the stag if the other player hunts the stag and to hunt a hare if the other player hunts a hare. In order to cooperate, the players need a measure of trust in each other (Skyrms 2001). For Skyrms, there are two approaches to achieving the better equilibrium: the Hobbesian and the Humean.

The Hobbesian approach is the standard rational choice answer – namely, that in order to move from a hare equilibrium to a stag equilibrium, the players must "change their beliefs about what the others will do" (Skyrms 2001, 35). Yet as he adds, "rational choice based game theory as usually conceived has nothing to say about how or why such a change of mind might take place" (Skyrms 2001, 35). Indeed, in the context of uncoordinated repeated play, Kandori et al. (1993) show that only the risk-dominant equilibrium is stable over time.

Nothing stops us, however, from stepping outside the bounds of rational choice-based game theory and finding ways to encourage the players to *change their beliefs*. Arló-Costa and Pedersen (2010) speak to the significant role played by social norms as mechanisms of belief change (cf. Sen 1996).

The Humean approach (Hume 1978) suggests that norms arise gradually and through a slow progression via repeated interactions. If individuals can adjust who they interact with over repeated play, then the Stag Hunt becomes a reinforcement learning problem, where interactions with other stag hunters are positively reinforced, i.e. each time the interaction ends in cooperation, a positive reward corresponding to the difference between a hare and the player's share of the stag is obtained. Recall, as we showed earlier, that in a one-shot Stag Hunt game, the Nash Equilibrium mixture is the uniform distribution (Figure 7). As the game evolves over time, reinforcement can have an effect by changing the mixture probability and thereby increasing the chance of cooperation. In the long run, as Skyrms and Pemantle (2000) show, cooperators learn to interact with other cooperators (coincidentally, defectors also learn to interact with other defectors).

From the perspective of a public policy practitioner, what kinds of norms with regard to inter-state AI dynamics should we adopt in order to cultivate the players to believe in the possibility of cooperation? Both

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

the Hobbesian and the Humean approaches suggest some directions for thinking about governance, construed here as a challenge for how to move from a noncooperative equilibrium to a cooperative equilibrium in the Stag Hunt game. From the perspective of the former, we ought to develop norms and institutions that foster cooperation by changing the players' beliefs about what the other players will do; from the perspective of the latter, we ought to develop norms and institutions that foster cooperation by discouraging isolationism and encouraging positive reinforcement through repeated interaction.

5.2 Central Guiding Principles

The preceding sections developed the substantive normative and game-theoretic frameworks that can help us make sense of the ethics of the geopolitics of AI. We will now develop several principles and tentative policy recommendations, drawing on the ethos of the Hobbesian and the Humean approaches. We begin with four guiding principles for inter-state AI dynamics.

1. **Broad State Representation and Effective Dispute Resolution**. Any structure for global AI coordination and decision-making, whether it be via formal agreements among several states or institutions jointly established between them, or informal norms and principles, must include effective, efficient, and credible dispute resolution mechanisms, which accord *fair weight and representation* to the interests of representatives of all states and other relevant stakeholders.
2. **Reduction of Existing Global Inequalities**. All actors with the capabilities to do so should strive to improve supply management and reduce global inequalities in the access, development, and use of leading AI technologies and the components required for their development.
3. **Diffusion of Blameless Responsibilities**. The forward-orientated, blameless forms of responsibilities for AI-generated injustices should be distributed among agents who are involved in reproducing the background conditions of such injustices. Actors should see to the enforcement of such blameless responsibilities through imposing *reasonable demands* and *proportionate costs*.
4. **The State as the Final Stop**. In the event that agents within their territorial jurisdiction refuse to take up their forward-looking normative responsibilities for AI-generated injustices, the state possesses the tentative responsibility to step in to partially ameliorate said injustices.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

The practical importance of the first principle rests in the fact that global dispute mechanisms can lose their credibility if they are perceived to unfairly favour one state or region. For example, the International Criminal Court (ICC) has been much maligned for both operating without major states as members (neither China nor the US are signatories to the Rome Statute) (International Criminal Court 2021), and for having prosecutorial priorities that are partisan, politically motivated, and shaped by the geopolitical interests of its member states (Wilmshurst 2019; Carney 2025; Askarisirchi 2025).

Meanwhile, the second and third principles channel much of the preceding discussion on relational inequality – the unequal development of AI risks exacerbating global inequality, and hence actors in inter-state AI dynamics shall aim for a flatter AI world. Likewise, forward-looking blameless responsibilities should be distributed broadly in a way that encourages repairing the background conditions of inequality.

Finally, the fourth principle echoes the centrality of sovereignty and the nation-state. While the nation-state is by no means the only normatively relevant actor in the process of responsibility assignment, it must serve as the guarantor of last resort, specifically for transgressions occurring *within* its borders.

5.3 Direct Policy Recommendations

In this section, we propose three concrete – albeit tentative – policy recommendations in line with the preceding four guiding principles.

Establishing Governance and Dispute Resolution Mechanisms (Principles #1 and #4)

If states believe that the commitments of other states on AI can be trusted, they are more likely to cooperate. This credibility is eroded with unpredictable and erratic actions, which have characterised the foreign policies of numerous international and regional powers, for example. One way to encourage credible commitment is by establishing effective and robust dispute resolution mechanisms. Such a recommendation is largely supported by our first two guiding principles; in particular, dispute resolution mechanisms are centrally important in collaborative games with multiple equilibria, perhaps even more so than they are in the work of Ostrom (1990), who places central importance on dispute resolution mechanisms for the management of common resources in PD-style games. When we have multiple equilibria, dispute resolution mechanisms can foster

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

cooperation and cultivate the trust required for players to believe that their counterparts will engage in cooperative behaviours; such a belief is in turn required for these individuals to likewise engage in cooperative behaviour. This is in the spirit of the Hobbesian approach to encouraging cooperation in the Stag Hunt: cultivating the institutions that are required to change players' beliefs about the likelihood of cooperation.

When effective, accessible and low-cost conflict resolution mechanisms are in place, they can shape parties' beliefs about the prospects of cooperation, by providing assurance that non-cooperative behaviour can be quickly addressed. This should not come as a surprise – for instance, among the very central functions of the WTO has been its staged process for resolving trade disputes. Complaints at the WTO are triaged accordingly: members begin with a complaint, which is then processed by the Dispute Settlement Board (DSB) and begins with consultations; where consultations fail to resolve the dispute, a panel is formed by the DSB. The panel produces a report which is then adopted by the DSB. Crucially, parties unsatisfied with the panel's report may appeal to the AB (World Trade Organization 2004). Unfortunately, in the case of the WTO, the AB has been non-functional since December 2019, due to the US' blockage of new appointments, which has resulted in a lack of quorum (Gao 2021).

Disputes around inter-state AI dynamics are inevitable, across a range of products and industries – including the mining of rare earth minerals, the manufacturing of semiconductors, the requisite equipment, the assembly of chips and GPUs, and the export/import restrictions of any of these items. The current regime does not reinforce cooperative behaviour. Instead, what we see is either unilateral trade barriers or tit-for-tat punitive measures. What is separately needed is a dispute resolution mechanism that stimulates confidence among states in the possibility of cooperation. There is no magic pill to solve this: the way to build such trust is gradual and it requires states to be willing to enter into and, more importantly, honour their agreements over time.

As we saw earlier, both China and the EU have advocated for the creation of an international AI governance institution, either independently or within the purview of the United Nations. For such an institution to be successful, it will be critical to include a dispute resolution mechanism, with a credible enforcement scheme (Hobbesian approach), or strongly institutionalised norms and rituals of cooperation and compromise (Humean approach). Indeed, more substantively, China has recently established the International Organization for Mediation (IOMed) (Government of the Hong Kong Special Administrative Region 2025).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

One way forward would be to build on this development and establish a mediation – or more broadly dispute adjudication – centre for geopolitical conflicts regarding inter-state AI dynamics, perhaps in a strategically neutral geographic location. Another way forward could be to repair the WTO AB and build within its purview a body dedicated to inter-state AI dynamics, though this is less comprehensive, as the WTO's mandate is limited to trade disputes.

Reforming Supply Chain Management (Principles #2 and #3)

The core of deep conflicts regarding inter-state AI dynamics involves access inequality and supply constraints. On grounds of both relational egalitarian principles and the interests of actors involved, managing supply in rare earth minerals, semiconductors, chips, GPUs, and related SME is essential for reinforcing cooperative behaviour. A number of policy developments in China have reflected a broader drive for global collaboration in supply management (e.g. MFA 2023 and MFA 2025b). These policy pronouncements could be reinforced with multilateral attempts to manage the supplies of AI-related technology, perhaps again through a multilateral institution, either within the auspices of the UN or independently. More recently, the US has also begun showing signs of dethawing of some export control restrictions on Nvidia GPUs (Mickle 2025); such positive policy changes, too, can be reinforced with more systematic multilateral supply management agreements.

When countries fear that their critical supplies will be cut off, this strongly incentivises reciprocal non-cooperative behaviour, which leads to a cycle of isolation. We grappled with this earlier, when the US export controls, described in Section 1, restricted, for example, Huawei's access to the leading Nvidia GPUs as well as restricting its market access in the United States. As a result, Huawei turned to domestic supply chains, and to cultivating a robust domestic market (Olcott 2025). While they managed to succeed in that market, a cooperative counterfactual where American consumers can freely purchase the latest Huawei products, and Huawei can get unfettered access to leading Nvidia GPUs, would have benefited the stakeholders of and consumers in both countries. Indeed, this is precisely the lesson of our argument in Section 4 – that mutual noncooperation ought to be the least preferred outcome and, by extension, that unilateral cooperation should be preferred to mutual noncooperation.

In the future, the role and agency of small and medium states – which serve as vital nodes in what Brantly Womack terms the 'multinodal' world (Womack 2024), or as critical chokepoints when it comes to AI supply

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

chains – cannot be overstated. These states not only possess the empirical room, but also the normative imperative, to maneuver across different, competing spheres of digital influence and AI standards. They should exercise their leverage against more powerful counterparts with care and precision, with the intention of pressing for more options, choices, and room for them to be non-aligned. In the face of false binaries, the rest of the world may seek an alternative.

Devising Global Mechanisms of Blameless Responsibility Uptake (Principle #3)

It is beyond our scope to discuss proposals aimed at bolstering mechanisms of legal accountability over AI failures and errors at length, though one of us has written extensively on the subject elsewhere (Babic et al. 2021). Granting that there exists no single overarching legal authority with the capacity of exacting and enforcing punishments and penalties for AI risks – especially across national borders – an alternative path may entail the establishment and implementation of policies, both within and in coordination with other nations, that would allow for responsibility uptake. Such responsibilities should not be rhetorically couched in terms of criminal liability, moral culpability, or political fault, for such labels, as noted by structural injustice theorists (e.g. Young 2011, Lu 2024, McKeown 2018, Nuti 2019), could often be invidious, exacerbate injustices, and undermine effective fact-finding.

Domestically, governments should explore legislation that binds actors across all stages of the AI training, development, and application process to duties to redress the risks associated with their decisions. A concrete example would be OpenAI, whose amassing of extreme levels of financial wherewithal, model opacity and disdain for open-source development, and repeated corporate restructuring efforts that have diluted its commitment to "ensure […] technology serves humanity" has drawn considerable consternation from experts in the AI community (A Right to Warn about Advanced Artificial Intelligence 2024; An Open Letter to OpenAI 2025). More specifically, in the geopolitical context, the US government should recognise that such concentration of resources within individual corporations – in the absence of accountability mechanisms – could well generate radical geopolitical uncertainties harmful not just to rival states, but also to the interests of its own people and itself. Legislation delineating limits to private corporate usages of AI in geopolitically sensitive contexts, without government or public authorisation, should be considered across jurisdictions in the world – democratic and non-democratic alike.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Conclusion

As promised, this Element has sought to advance three claims: (1) the Risk-Opportunity Claim, (2) the Justice Claim, and (3) the Interest Claim. While we cannot hope to exhaustively cover a field as vast and interdisciplinary as the geopolitics of AI, we hope to have put forward compelling arguments for these three claims and opened up a vast space for further research. As AI becomes ever entrenched in our daily lives, the case for inter-disciplinary research employing a wide, rich range of methods has never been stronger.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Glossary

Concept Drift A change in the relationship between input features and output labels across contexts or over time, causing a model's predictions to become less reliable.

Covariate Shift A change in the distribution of input features when a model is applied to a different population from the one on which it has been trained.

Global AI Relational Egalitarianism (GARE) A framework proposed by the authors applying relational egalitarian principles to the global governance of AI, encompassing its development, diffusion, and use, with particular attention to status hierarchies, decision-making power, and respect.

Hawk-Dove Game A game-theoretic model of conflict in which two players choose between aggressive (Hawk) and conciliatory (Dove) strategies, used to analyse the conditions under which escalation or compromise emerges.

Interactional Risks Risks arising from the compounding of AI-related dangers by geopolitical dimensions – such as conflict escalation, biased cross-border decision-making, or the siloing of AI development across adversarial states.

Liability Model (LM) An approach to moral and legal responsibility that identifies culpable agents by establishing intentional, causal connections between their actions and harmful outcomes, and holds them blameworthy where no valid excuse or exemption applies.

Luck Egalitarianism An approach to equality that posits that differences in welfare or resources stemming from circumstances beyond an individual's control, such as place of birth or nationality, are morally arbitrary and therefore unjustifiable.

Mixed Strategy Nash Equilibrium A Nash Equilibrium in which at least one player randomises across two or more strategies according to a probability distribution, such that no player can gain by deviating.

Nash Equilibrium A stable state in a game in which no player can improve their outcome by unilaterally changing their strategy, given the strategies chosen by all other players.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Pareto Dominance An outcome A is said to Pareto-dominate outcome B if at least one player is better off under A than under B, and no player is worse off.

Prisoner's Dilemma Game A game-theoretic scenario in which two individually rational players, unable to coordinate, each choose to defect, producing a collectively suboptimal outcome.

Pure Strategy Nash Equilibrium A Nash Equilibrium in which each player adopts a fixed strategy with certainty, rather than randomising across available options.

Relational Egalitarianism An approach to equality focused not primarily on resource distribution but on ensuring that individuals relate to one another without entrenched hierarchies of domination, inferiority, or disrespect.

Responsibility Gap Challenge The difficulty of assigning moral or legal responsibility for AI-induced harms, arising from algorithmic opacity, the distributed architecture of AI systems, and the diffuse nature of the resulting harms.

Social Connection Model (SCM) An alternative to the Liability Model, developed by Iris Marion Young and refined by the authors, which assigns forward-looking, non-blame-centric responsibilities to all actors whose conduct reproduces the structural conditions giving rise to an injustice.

Stag Hunt Game A game-theoretic scenario with multiple Nash Equilibria (including a cooperative one and a non-cooperative one) modelling the tension between social cooperation and individual security, in which the collectively optimal outcome requires mutual trust and coordination.

Status Hierarchies Entrenched social structures in which certain actors hold systematically superior positions of esteem, standing, or power over others – a form of inequality that relational egalitarianism, and GARE in particular, seeks to dismantle.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

References

Abizadeh, Arash. 2008. "Democratic Theory and Border Coercion: No Right to Unilaterally Control Your Own Borders." *Political Theory* 36 (1): 37–65.

Ahn, Michael, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, et al. 2022. "Do as I Can, Not as I Say: Grounding Language in Robotic Affordances." arXiv preprint arXiv:2204.01691.

Alperovitz, Gar. 1995. *The Decision to Use the Atomic Bomb, and the Architecture of an American Myth*. New York: Alfred A Knopf.

Altman, Sam. 2024. "Who Will Control the Future of AI?" *The Washington Post*, July 15.

Amodei, Dario. 2025. "On DeepSeek and Export Controls." *SuperIntelligence – Robotics – Safety & Alignment* 2 (1): 1–6.

Amoroso, Daniele, and Guglielmo Tamburrini. 2019. "Filling the Empty Box: A Principled Approach to Meaningful Human Control over Weapons Systems." *ESIL Reflections* 8 (11): 1–9.

Anderson, Elizabeth. 1999. "What Is the Point of Equality?" *Ethics* 109 (2): 287–337.

Anderson, Elizabeth. 2010. "The Fundamental Disagreement between Luck Egalitarians and Relational Egalitarians." *Canadian Journal of Philosophy Supplementary* 40: 1–23.

Anderson, Elizabeth. 2012. "Equality." In *The Oxford Handbook of Political Philosophy*, edited by David Estlund. Oxford: Oxford University Press.

Angwin, Julia, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23.

Anthropic. 2025. "Anthropic's Recommendations to OSTP for the U.S. AI Action Plan." *Anthropic*. March 6.

A Right to Warn About Advanced Artificial Intelligence. "A Right to Warn about Advanced Artificial Intelligence." 2024.

Arló-Costa, Horacio, and Arthur Paul Pedersen. 2010. "Social Norms, Rational Choice and Belief Change." In *Belief Revision Meets Philosophy of Science*, edited by Erik J. Olsson and Sebastian Enqvist. Dordrecht: Springer.

Aschenbrenner, Leopold. 2024. "Situational Awareness: The Decade Ahead." *Situational Awareness*, June.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Askarisirchi, Emma. 2025. "International Criminal Court (ICC) and the Pursuit of Global Justice: 'A Court Worth Having' – Predicaments to the Court's Enforcement and Frameworks for Improving Compliance." *Columbia Human Rights Law Review* 57 (1): 151–205.

Atalan, Yasir, Ian Reynolds, and Benjamin Jensen. 2025. "AI Biases in Critical Foreign Policy Decisions." Center for Strategic and International Studies, February 26.

Axelrod, Robert, and William D. Hamilton. 1981. "The Evolution of Cooperation." *Science* 211 (4489): 1390–1396.

Babic, Boris, and I. Glenn Cohen. 2023. "The Algorithmic Explainability 'Bait and Switch.'" *Minnesota Law Review* 108 (4): 857–909.

Babic, Boris, and Zoë Johnson King. 2025. "Algorithmic Fairness and Resentment." *Philosophical Studies* 182 (1): 87–119.

Babic, Boris, and Brian Wong. 2026. "Countries Must Move beyond seeing AI as a Race, where One Side Must Beat the Other." *Fortune*, January 3.

Babic, Boris, I. Glenn Cohen, Ariel Dora Stern, Yiwen Li, and Melissa Ouellet. 2025. "A General Framework for Governing Marketed AI/ML Medical Devices." *j Digital Medicine* 8 (1): 328.

Babic, Boris, I. Glenn Cohen, Theodoros Evgeniou, and Sara Gerke. 2021. "When Machine Learning Goes Off the Rails." *Harvard Business Review*, January.

Baschuk, Bryce. 2023. "How Trump Could Deal Another Blow to Already Hobbled WTO." *Bloomberg*, September 4.

Bay, Morten. 2023. "AI Ethics and Policymaking: Rawlsian Approaches to Democratic Participation, Transparency, Accountability, and Prediction." *Social Science Research Network*, May 31.

Beck, Ulrich. 1992. *Risk Society: Towards a New Modernity*. London: Sage Publications.

Beduschi, Ana. 2021. "International Migration Management in the Age of Artificial Intelligence." *Migration Studies* 9 (3): 576–596.

Beitz, Charles R. 1999. "Social and Cosmopolitan Liberalism." *International Affairs* 75 (3): 515–529.

Beitz, Charles R. 2009. *The Idea of Human Rights*. Oxford: Oxford University Press.

Bell, Daniel A., and Pei Wang. 2020. *Just Hierarchy: Why Social Hierarchies Matter in China and the Rest of the World*. Princeton: Princeton University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Margaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In Proceedings of the 2021

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

ACM Conference on Fairness, Accountability, and Transparency, 610–623. New York, NY: ACM.

Bengio, Yoshua, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, et al. 2025. "International AI Safety Report." DSIT 2025/001.

Bennett, James, and Stan Lanning. 2007. "The Netflix Prize." KDDCup'07, August 12.

Bidadanure, Juliana, and David Axelsen. 2025. "Egalitarianism." *Stanford Encyclopedia of Philosophy*.

Bindra, Jaspreet. 2025. "AI Warfare Is Here: How Intelligent Drones Harop and Heron Fronted India's Operation Sindoor." *The Economic Times*, May 17.

Bisht, Ankit, Lareina Yee, and Roger Roberts. 2025. "Open Source Technology in the Age of AI." McKinsey & Company, April 22.

Blackmon, Shauna. 2024. "Bridging the Education Divide: How AI Can Support Underserved Communities." Tech to the Rescue, November 19.

Blevins, Emily G., Alice B. Grossman, and Karen M. Sutter. 2023. *Frequently Asked Questions: CHIPS Act of 2022 Provisions and Implementation*. R47523. Congressional Research Service. April 25.

Bojić, Ljubiša, Irena Stojković, and Zorana Jolić Marjanović. 2024. "Signs of Consciousness in AI: Can GPT-3 Tell How Smart It Really Is?" *Humanities and Social Sciences Communications* 11 (1): 1631.

Bostrom, Nick. 2014. *Superintelligence*. Oxford: Oxford University Press.

Bown, Chad P., and Melina Kolb. 2018. "Trump's Trade War Timeline: An Up-to-Date Guide." Peterson Institute for International Economics. Last updated January 10, 2025.

Bradford, Anu. 2023. *Digital Empires: The Global Battle to Regulate Technology*. New York: Oxford University Press.

Braun, Matthias, Hannah Bleher, Eva Maria Hille, and Jenny Krutzinna. 2023. "Tackling Structural Injustices: On the Entanglement of Visibility and Justice in Emerging Technologies." *American Journal of Bioethics* 23 (7): 100–102.

Braun, Matthias, Patrik Hummel, Susanne Beck, and Peter Dabrock. 2020. "Primer on an Ethics of AI-Based Decision Support Systems in the Clinic." *Journal of Medical Ethics* 47 (12).

Brennan, Tim, and William Dieterich. 2018. "Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)." In *Handbook of Recidivism Risk/Needs Assessment Tools*, 49–75. Hoboken, NJ: John Wiley & Sons.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Browne, Jude. 2023. "AI and Structural Injustice: A Feminist Perspective." In *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, edited by Jude Browne, Stephen Cave, Eleanor Drage, and Kerry McInerney. Oxford: Oxford University Press.

Browne, Jude. 2025. *Political Responsibility and Tech Governance*. Cambridge: Cambridge University Press.

Bureau of Industry and Security. 2022. "Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification." *Federal Register* 87 (197): 62186–62246, October 13.

Bureau of Industry and Security. 2023. "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections." *Federal Register* 88 (205): 73458–73513, October 25.

Bureau of Industry and Security. 2024a. "Commerce Control List Additions and Revisions; Implementation of Controls on Advanced Technologies Consistent with Controls Implemented by International Partners." *Federal Register* 89 (173): 72926–72956, September 6.

Bureau of Industry and Security. 2024b. "Foreign-Produced Direct Product Rule Additions, and Refinements to Controls for Advanced Computing and Semiconductor Manufacturing Items." *Federal Register* 89 (234): 96790–96864, December 5.

Bureau of Industry and Security. 2025a. "Expansion of End-User Controls to Cover Affiliates of Certain Listed Entities." *Federal Register* 90 (187): 47201–47214, September 30.

Bureau of Industry and Security. 2025b. "Framework for Artificial Intelligence Diffusion." *Federal Register* 90 (9): 4544–4584, January 15.

Callaway, Ewen. 2020. "'It will Change Everything': DeepMind's AI Makes Gigantic Leap in Solving Protein Structures." *Nature*, November 30.

Callaway, Ewen. 2022. "How AlphaFold and Other AI Tools Could Help us Prepare for the Next Pandemic." *Nature*, October 11.

Candès, Emmanuel J., and Benjamin Recht. 2009. "Exact Matrix Completion via Convex Optimization." *Foundations of Computational Mathematics* 9: 717–772.

Caney, Simon. 2006. "Cosmopolitan Justice, Responsibility, and Global Climate Change." *Leiden Journal of International Law* 18 (4): 747–775.

Carens, Joseph H. 2013. *The Ethics of Immigration*. New York: Oxford University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Carlsmith, Joseph. 2022. "Is Power-Seeking AI an Existential Risk?" arXiv preprint arXiv:2206.13353.

Carney, Charlotte. 2025. "Abolishing the International Criminal Court." *The International Journal of Human Rights* 30 (1): 254-272.

Carr, Edward Hallett. 1940. *The Twenty Years' Crisis, 1919–1939: An Introduction to the Study of International Relations*. New York: Harper & Row.

Carroll, Micah, Alan Chan, Henry Ashton, and David Krueger. 2023. "Characterizing Manipulation from AI Systems." In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)* article 6, 1–13. New York, NY: Association for Computing Machinery.

Casey, Christopher A., and Paul K. Kerr. 2021. *The U.S. Export Control System and the Export Control Reform Act of 2018*. R46814. Congressional Research Service. June 7.

Center for Security and Emerging Technology. 2022. Translation of "Notice of the State Council on the Publication of 'Made in China 2025'." March 10. https://cset.georgetown.edu/publication/notice-of-the-state-council-on-the-publication-of-made-in-china-2025/.

CHIPS and Science Act of 2022, 136 Stat. 1366 (2022). www.congress.gov/117/plaws/publ167/PLAW-117publ167.pdf ('CHIPS Act 2022').

Chow, Andrew R. 2024. "Some Top AI Labs Have 'Very Weak' Risk Management, Study Finds." *TIME*, October 2.

Chow, Vincent. 2025. "Alibaba Integrates Nvidia's AI Robotics Tools on Cloud Platform." *South China Morning Post*, September 24.

Coeckelbergh, Mark. 2025. "Three Challenges for a Global AI Ethics: Towards a More Relational Normative Vision." *AI Ethics* 5: 5527–5533.

Cohen, G. A. 1989. "On the Currency of Egalitarian Justice." *Ethics* 99 (4): 906–944.

Davenport, Thomas H., and Rajeev Ronanki. 2018. "Artificial Intelligence for the Real World." *Harvard Business Review* 96 (1):108–116.

Davidson, Helen. 2025. "China Steps up Control of Rare-Earth Exports Citing 'National Security' Concerns." *The Guardian*, October 9.

Delcker, Janosch. 2020. "POLITICO AI: Decoded: How Cambridge Analytica Used AI – No, Google Didn't Call for a Ban on Face Recognition – Restricting AI Exports." *POLITICO*, January 28.

Dib, Daniela, Lam Le, and Yashraj Sharma. 2025. "These Countries Want to Be the Next Big Semiconductor Hubs." Rest of World, August 11.

Dignum, Virginia. 2020. "Responsibility and Artificial Intelligence." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford: Oxford University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Ding, Chao, Jing Ke, Mark Levine, and Nan Zhou. 2024. "Potential of Artificial Intelligence in Reducing Energy and Carbon Emissions of Commercial Buildings at Scale." *Nature Communications* 15 (1): 5916.

Diya, Sabhanaz Rashid. 2025. "Advancing Multi-Stakeholderism for Global Governance of the Internet and AI." Centre for International Governance Innovation, Policy Brief No. 201, March 26.

Doshi, Rush, and Chris McGuire. 2025. "Export Controls Failed? China Begs to Differ." *Wall Street Journal*, September 1.

Dwork, Cynthia. 2006. "Differential Privacy." In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Proceedings, Part II*, edited by Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener. *Lecture Notes in Computer Science*, vol 4052. Berlin: Springer.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography*, edited by Shai Halevi and Tal Rabin. Lecture Notes in Computer Science, vol 3876. Berlin: Springer.

École Polytechnique. 2025. "MGX, Bpifrance, Mistral AI, and NVIDIA Launch Joint Venture to Build Europe's Largest AI Campus in France." May 19.

Elford, Gideon. 2017. "Survey Article: Relational Equality and Distribution." *Journal of Political Philosophy* 25 (4): 80–99.

Elish, Madeleine Clare. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5: 40–60.

Esposito, Mark. 2025. "AI Geopolitics and Data Centres in the Age of Technological Rivalry." World Economic Forum, July 24.

European Commission. 2025. "AI Act." Digital Strategy EC Europa, Last Updated November 28, 2025. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Fang, Tianyu, and Tim Hwang. 2023. "The Rise of Techno-Nationalism". New America, November 9, 2023.

Fanon, Frantz. 1963. *The Wretched of the Earth*. New York: Grove Press.

Flint, Colin. 2021. *Introduction to Geopolitics*. London: Routledge.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Valena. 2018. "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28 (4): 689–707.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Foley, John. 2024. "Google and Peers Weigh an AI Prisoner's Dilemma." *The Financial Times*, October 30.

Frenkel, Sheera, and Natan Odenheimer. 2025. "How the War in Gaza Drove Israel's A.I. Experiments." *The New York Times*, April 25.

Frost & Sullivan. 2025. "AI Empowerment across Industries White Paper." *Frost & Sullivan*. August 25.

Fung, Pascale, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, et al. 2025. "Embodied AI Agents: Modeling the World." arXiv preprint arXiv: 2506.22355.

Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30 (3): 411–437.

Gabriel, Iason. 2022. "Toward a Theory of Justice for Artificial Intelligence." *Daedalus* 151 (2): 218–231.

Gao, Henry. 2021. "Finding a Rule-Based Solution to the Appellate Body Crisis: Looking Beyond the Multiparty Interim Appeal Arbitration Arrangement." *Journal of International Economic Law* 24 (3): 534–550.

Gass, Saul I., and Arjang A. Assad. 2005. *An Annotated Timeline of Operations Research: An Informal History*. New York: Springer.

Gerke, Sara, Boris Babic, Theodoros Evgeniou, and I. Glenn Cohen. 2020. "The Need for a System View to Regulate Artificial Intelligence/ Machine Learning-Based Software as Medical Device." *NPJ Digital Medicine* 3 (1): 53.

Glazier, Alexandra K., Mathew Moss, and Lindsay A. Martin. 2021. "Electronic Health Records Can Improve the Organ Donation Process." *Harvard Business Review*, December 13.

GlobeScan. 2016. "Global Citizenship a Growing Sentiment among Citizens of Emerging Economies: Global Poll." April 27.

Goh, Brenda, Eduardo Baptista, and Qiaoyi Li. 2025. "China's AI-Powered Humanoid Robots Aim to Transform Manufacturing." *Reuters*, May 13.

Government of Canada. 2025. "AI Sovereign Compute Infrastructure Program." Last updated May 5.

Government of the Hong Kong Special Administrative Region. 2025. "Speech by CE at Signing Ceremony of the Convention on the Establishment of the International Organization for Mediation." News Release. May 30.

Government of the Netherlands. 2025. "Klever: Export Controls on Advanced Semiconductor Manufacturing Equipment to be Tightened." January 15.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Gu, Hongfei. 2024. "Data, Big Tech, and the New Concept of Sovereignty." *Journal of Chinese Political Science* 29: 591–612.

Gyevnár, Bálint, and Atoosa Kasirzadeh. 2025. "AI Safety for Everyone." *Nature Machine Intelligence* 7 (4): 531–542.

Hagen, Lisa, Jingnan Huo, and Audrey Nguyen. 2025. "Elon Musk's AI Chatbot, Grok, Started Calling Itself 'MechaHitler.'" *NPR*, July 9.

Hao, Karen. 2025. *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. London: Penguin Random House.

Harithas, Barath. 2025. "The AI Diffusion Framework: Securing U.S. AI Leadership While Preempting Strategic Drift." Center for Strategic & International Studies.

Hebbar, Vivek, and Evan Hubinger. 2022. "Path Dependence in ML Inductive Biases." *Less Wrong*, AI Alignment Forum.

Heeks, Richard. 2022. "Digital Inequality beyond the Digital Divide: Conceptualizing Adverse Digital Incorporation in the Global South." *Information Technology for Development* 28 (4): 688–704.

Himmelreich, Johannes, and Désirée Lim. 2022. "AI and Structural Injustice: Foundations for Equity, Values, and Responsibility." In *The Oxford Handbook of AI Governance*, edited byJustin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang. Oxford: Oxford University Press.

Hobbes, Thomas. 1651. *Leviathan*, edited by W. G. Pogson Smith. Reprint edition, Oxford: Clarendon Press (1929).

Huang, Jensen. 2025. "A New Innovation Economy: A Conversation with NVIDIA CEO Jensen Huang." Presentation at the Milken Institute Global Conference, May 6, Los Angeles, CA.

Hume, David. 1978. A Treatise of Human Nature, reprinted from the original edition in three volumes and edited, with an analytical index, by L. A. Selby-Bigge and Second Edition revisions with variant readings by P. H. Nidditch. Oxford: Oxford University Press.

International Criminal Court. 2021. *Rome Statute of the International Criminal Court*.

Ip, Kevin K.W. 2016. *Egalitarianism and Global Justice: From a Relational Perspective*. London: Palgrave Macmillan.

Jensen, Benjamin, Ian Reynolds, and Yasir Atalan. 2025. "Critical Foreign Policy Decisions (CFPD)-Benchmark: Measuring Diplomatic Preferences in Large Language Models." arXiv preprint arXiv:2503.06263.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Jervis, R. 1978. "Cooperation under the Security Dilemma." *World Politics* 30 (2): 167–214.

Ji, Jiaming, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, et al. 2025. "AI Alignment: A Contemporary Survey." *ACM Computing Surveys* 58 (5): 1–38.

Jiang, Ben. 2024. "As OpenAI blocks China, developers scramble to keep GPT access through VPNs." *South China Morning Post*, July 10.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596: 583–589.

Kandori, Michihiro, George J. Mailath, and Rafael Rob. 1993. "Learning, Mutation, and Long Run Equilibria in Games." *Econometrica* 61 (1): 29–56.

Kiener, Maximilian. 2024. "Strict Moral Answerability." *Ethics* 134 (3): 360–386.

Kirk, Hannah R., Iason Gabriel, Christopher Summerfield, et al. 2025. "Why Human–AI Relationships Need Socioaffective Alignment." *Humanities and Social Sciences Communications* 12 (1): 728.

Kmentt, Alexander. 2025. "Geopolitics and the Regulation of Autonomous Weapons Systems." *Arms Control Association*, January/February.

Kuik, Cheng-Chwee. 2008. "The Essence of Hedging: Malaysia and Singapore's Response to a Rising China." *Contemporary Southeast Asia* 30 (2): 159–185.

Kuik, Cheng-Chwee. 2016. "How Do Weaker States Hedge? Unpacking ASEAN States' Alignment Behavior towards China." *Journal of Contemporary China* 25 (100): 500–514.

Kumar, Naveen. 2025. "DeepSeek AI Statistics 2025: Users & Revenue." Demand Sage.

Lazard Geopolitical Analysis. 2023. "The Geopolitics of Artificial Intelligence." Published November 17.

Learning Res. Inc. v. Trump, No. 24-1287 (consolidated with Trump v. V.O.S. Selections, Inc., No. 25-250), Transcript of Oral Argument (US Nov. 5, 2025).

Lee, Angie. 2024. "What Is Sovereign AI?" NVIDIA Blog. February 28.

Lee, Lizzie C. 2024. "China's Big Fund 3.0: Xi's Boldest Gamble Yet for Chip Supremacy." *The Diplomat*, June 6.

Lewis, David. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Lewis, David. 2000. "Causation as Influence." *The Journal of Philosophy* 97 (4): 182–197.

Li, Cathy. 2023. "Responsible AI Governance can be Achieved Through Multistakeholder Collaboration." World Economic Forum, November 14.

Li, Xiao Peng, Lei Huang, Hing Cheung So, and Bo Zhao. 2019. "A Survey on Matrix Completion: Perspective of Signal Processing." arXiv preprint arXiv:1901.10885.

Liu, Qianer. 2023. "How Huawei Surprised the US with a Cutting-Edge Chip Made in China." The Financial Times, November 30.

Lu, Catherine. 2024. "Responsibility, Structural Injustice, and Settler Colonialism." In *What is Structural Injustice?* edited by Jude Brown and Maeve McKeown. Oxford: Oxford University Press.

Lu, Chuanying. 2025. "EU AI Strategy and Prospects for China-EU Cooperation on AI Governance." *China Watch* 5 (35). China-CEE Institute. September.

Lushenko, Paul, and Keith Carter. 2024. "A New Military-Industrial Complex: How Tech Bros Are Hyping AI's Role in War." Bulletin of the Atomic Scientists, October 7.

Madaio, Michael, Su Lin Blodgett, Elijah Mayfield, and Ezekiel Dixon-Román. 2022. "Beyond "fairness": Structural (in)justice Lenses on AI for Education." In *The Ethics of Artificial Intelligence in Education*, edited by Wayne Holmes and Kaśka Porayska-Pomsta. London: Routledge.

Maddox, Robert James. 2004. *Weapons for Victory: The Hiroshima Decision*. Columbia: University of Missouri Press.

Mahindroo, Abhijit, Anupama Suryanarayanan, and Marc de Jong. 2025. "Silicon Squeeze: AI's Impact on the Semiconductor Industry." McKinsey & Company, April 30.

Maynard Smith, J., and G. A. Parker. 1973. "The Logic of Animal Conflicts." *Nature* 246: 15–18.

McCain, John S. National Defense Authorization Act for Fiscal Year 2019, 132 Stat. 1636 (2018). www.congress.gov/115/plaws/publ232/PLAW-115publ232.pdf ('ECRA 2018').

McKernan, Bethan, and Harry Davies. 2024. "'The Machine Did It Coldly': Israel Used AI to Identify 37,000 Hamas Targets." *The Guardian*, April 3.

McKeown, Maeve. 2018. "Iris Marion Young's 'Social Connection Model' of Responsibility: Clarifying the Meaning of Connection." *Journal of Social Philosophy* 49 (3): 484–502.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

McLaren, Margaret A. 2019. "Global Gender Justice: Human Rights and Political Responsibility." *Critical Horizons* 20 (2): 127–144.

Mearsheimer, John J. 2001. *The Tragedy of Great Power Politics*. New York: W. W. Norton & Company.

Medina, Ayman Falak. 2025. "Malaysia's Semiconductor Growth: Can It Move up the Value Chain?" ASEAN Briefing, March 17.

Miao, Fengchun, Wayne Holmes, Ronghuai Huang, and Hui Zhang. 2021. *AI and Education: A Guidance for Policy-Makers*. Paris: UNESCO.

Mickle, Tripp. 2025. "Nvidia Says U.S. Has Lifted Restrictions on A.I. Chip Sales to China." *The New York Times*, July 14.

Miller, Chris. 2025. "How US Export Controls Have (and Haven't) Curbed Chinese AI." *AI Frontiers*, July 8.

Miller, David. 1995. "Citizenship and Pluralism." *Political Studies* 43 (3): 432–450.

Miller, David. 2007. *National Responsibility and Global Justice*. Oxford: Oxford University Press.

Ministry of Foreign Affairs, People's Republic of China. 2022. "Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence." November 17.

Ministry of Foreign Affairs, People's Republic of China. 2023. "Global AI Governance Initiative." October 20.

Ministry of Foreign Affairs, People's Republic of China. 2024. "Shanghai Declaration on Global AI Governance." July 4.

Ministry of Foreign Affairs, People's Republic of China. 2025a. "Ding Xuexiang on EnablingAI and Other Scientific and Technological Innovation to Better Serve the Well-being of All Humanity." January 22.

Ministry of Foreign Affairs, People's Republic of China. 2025b. "Global AI Governance Action Plan." July 26.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2): 1–21.

Murrill, Brandon J. 2019. "The WTO's Appellate Body Loses Its Quorum: Is This the Beginning of the End for the "Rules-Based Trading System"?" Library of Congress.

Narayanan, Arvind, and Vitaly Shmatikov. 2007. "How to Break Anonymity of the Netflix Prize Dataset." arXiv preprint cs/0610105.

Nash, John. 1951. "Non-Cooperative Games." *Annals of Mathematics* 54 (2): 286–295.

Nath, Rekha. 2011. "Equal Standing in the Global Community." *The Monist* 94 (4): 593–614.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Nath, Rekha. 2015. "On the Scope and Grounds of Social Equality." In *Social Equality: On What It Means to be Equals*, edited by Carina Fourie, Fabian Schuppert, and Ivo Wallimann-Helmer. New York: Oxford University Press.

National Telecommunications and Information Administration. 2024. *Risks and Benefits of Dual-Use Foundation Models with Widely Available Model Weights*. Washington, DC: NTIA.

Nefsky, Julia. 2012. "Consequentialism and the Problem of Collective Harm: A Reply to Kagan." *Philosophy & Public Affairs* 39 (4): 364–395.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Nussbaum, Martha C. 2011. "Foreword." In *Responsibility for Justice*, edited by Iris Marion Young, Oxford: Oxford University Press, ix–xxv.

Nuti, Alasia. 2019. *Injustice and the Reproduction of History: Structural Inequalities, Gender and Redress*. Cambridge: Cambridge University Press.

NVIDIA. 2024. "Japan Cloud Leaders Build NVIDIA AI Infrastructure to Transform Industries for the Age of AI." November 12.

Nye, Joseph S. 2020. "Why Morals Matter in Foreign Policy." Project Syndicate, January 12.

OECD. 2025. "AI in Policy Evaluation: Governing with Artificial Intelligence." September 18. www.oecd.org/en/publications/governing-with-artificial-intelligence_795de142-en/fullreport/ai-in-policy-evaluation_c88cc2fd.html.

Okidegbe, Ngozi. 2022. "Discredited Data." *Cornell Law Review* 107: 2007–2065.

Olander, Eric, and Brian Wong. 2025. "China's Play for Global Governance Leadership." The China-Global South Project. Podcast.

Olcott, Eleanor. 2025. "Huawei Seeks to Grab Market Share in AI Chips from Nvidia in China." *The Financial Times*, January 21.

OpenAI. 2025. "OpenAI's Economic Blueprint." https://openai.com/global-affairs/openais-economic-blueprint/.

OpenAI Transparency. "An Open Letter to OpenAI." 2025.

Ostrom, Elinor. 1990. *Governing the Commons*. Cambridge: Cambridge University Press.

Padmanaban, Harish. 2023. "Navigating the Intricacies of Regulations: Leveraging AI/ML for Accurate Reporting." *Journal of Knowledge Learning and Science Technology* 2 (3): 401–412.

Parfit, Derek. 1984. "Five Mistakes in Moral Mathematics." In *Reasons and Persons*. Oxford, GB: Oxford Paperbacks: 67–86.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Park, Peter S., Simon Goldstein, Aidan O'Gara, Megan Chen, and Dan Hendrycks. 2024. "AI Deception: A Survey of Examples, Risks, and Potential Solutions." *Patterns* 5 (5): 1–16.

Pattison, James. 2015. "The Morality of Sanctions." *Social Philosophy and Policy* 32 (1): 192–215.

Payne, Kenneth. 2021. *I, Warbot: The Dawn of Artificially Intelligent Conflict*. Oxford: Oxford University Press.

Pejó, Balázs, and Damien Desfontaines. 2020. "SoK: Differential Privacies." *Proceedings on Privacy Enhancing Technologies Symposium* 2020.

Perez, Carlos E. 2025. "The Prisoner's Dilemma of AI Competition." *Intuition Machine*. Medium Post, Aug 4. https://medium.com/intuitionmachine/in-a-nondescript-conference-room-in-mountain-view-california-a-group-of-google-researchers-6b0859ba4541

Perrigo, Billy. 2023. "Exclusive: OpenAI Used Kenyan Workers on Less than $2 per Hour to Make ChatGPT Less Toxic." *TIME*, January 18.

Petric, Sebastian. 2025. "How to Enhance Geopolitical Risk Assessment using this Combined Strategy." World Economic Forum, February 21.

Pettit, Philip. 1999. *Republicanism: A Theory of Freedom and Government*. New York: Oxford University Press.

Pogge, Thomas. 2008. *World Poverty and Human Rights: Cosmopolitan Responsibilities and Reforms*. 2nd ed. Cambridge: Polity.

Praas, Robert, and Pierre-Alexandre Balland. 2025. "Europe's Hidden AI Champions Are Vital in Its Quest to Becoming a Global Tech Leader." *Centre for European Policy Studies*, February 10.

Press, Gil. 2024. "In 2024, Israel Became A Global Leader in Applied AI Innovation." *Forbes*, December 22.

Protecting Americans from Foreign Adversary Controlled Applications Act, 138 Stat. 955 (2024). www.congress.gov/118/plaws/publ50/PLAW-118publ50.pdf ('ByteDance 2024').

Race, Michael, and Tom Espiner. 2025. "How Is China Hitting Back Against US Tariffs?" *BBC News*. March 4.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. "AI and the Everything in the Whole Wide World Benchmark." Preprint, submitted November 30, 2021. arXiv preprint arXiv:2111.15366.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Rawls, John. 2001. *Justice as Fairness*, edited by Erin Kelly. Cambridge: Harvard University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Reed, Smith. 2024. "Navigating the Complexities of AI Regulation in China." Reed Smith In-depth, August 7.

Reiman, Jeffrey. 2012. "Review: The Structure of Structural Injustice: Thoughts on Iris Marion Young's "Responsibility for Justice."" *Social Theory and Practice* 38 (4): 738–751.

Riker, Walter. 2009. "The Democratic Peace Is Not Democratic: On Behalf of Rawls' Decent Societies." *Political Studies* 57 (3): 617–638.

Rivera, Juan-Pablo, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. "Escalation Risks from LLMs in Military and Diplomatic Contexts." Stanford University Human-Centered Artificial Intelligence, May 2.

Rousseau, Jean-Jacques. (1755) 1985. *Discourse on the Origin and Basis of Inequality Among Men*. New York: Penguin Classics.

Rubera, Gaia. 2023. "In the Beginning It Was Cambridge Analytica … ." Bocconi University. Published May 23.

Santoni de Sio, Filippo, and Giulio Mecacci. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." *Philosophy & Technology* 34 (4): 1057–1084.

Scheffler, Samuel. 2003. "What Is Egalitarianism?" *Philosophy & Public Affairs* 31 (1): 5–39.

Scheffler, Samuel. 2014. "The Practice of Equality." In *Social Equality: On What It Means to be Equals*, edited by Carina Fourie, Fabian Schuppert, and Ivo Wallimann-Helmer. Oxford: Oxford University Press.

Scheffler, Samuel. 2015. "Distributive Justice, the Basic Structure and the Place of Private Law." *Oxford Journal of Legal Studies* 35 (2): 213–235.

Schmid, Stefka, Daniel Lambach, Carlo Diehl, and Christian Reuter. 2025. "Arms Race or Innovation Race? Geopolitical AI Development." *Geopolitics* 30 (4): 1–30.

Sen, Amartya. 1996. "Is the Idea of Purely Internal Consistency of Choice Bizarre?" In *Language, World and Reality, a festschrift for Bernard Williams*, edited by J. E. J. Altham and T. R. Harrison. Cambridge: Cambridge University Press.

Sharp, Daniel. 2022. "Relational Equality and Immigration." *Ethics* 132 (3): 644–679.

Sheehan, Matt. 2023. "China's AI Regulations and How They Get Made." Carnegie Endowment for International Peace.

Shivakumar, Sujai, Charles Wessner, and Thomas Howell. 2024. "Balancing the Ledger: Export Controls on U.S. Chip Technology to China." Center for Strategic & International Studies.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Shivakumar, Sujai, Charles Wessner, and Thomas Howell. 2025. "The Limits of Chip Export Controls in Meeting the China Challenge." Center for Strategic & International Studies.

Simmons-Edler, Riley, Ryan Badman, Shayne Longpre, and Kanaka Rajan. 2024. "AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research." arXiv preprint arXiv:2405.01859.

Sipalan, Joseph. 2025. "Malaysia Tightens Chip Regulations amid US Scrutiny over Nvidia Diversions to China." *South China Morning Post*, March 26.

Skyrms, Brian. 2001. "The Stag Hunt." *Proceedings and Addresses of the American Philosophical Association* 75 (2): 31–41.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Skyrms, Brian, and Robin Pemantle. 2000. "A Dynamic Model of Social Network Formation." *Proceedings of the National Academy of Sciences of the United States of America* 97 (16): 9340–9346.

Sleat, Matt. 2010. "Bernard Williams and the Possibility of a Realist Political Theory." *European Journal of Political Theory* 9 (4): 485–503.

Smart, Andrew, and Atoosa Kasirzadeh. 2025. "Beyond Model Interpretability: Socio-Structural Explanations in Machine Learning." *AI & Society* 40: 2045–2053.

Spencer, Craig. 2025. "The Trump Administration Will Automate Health Inequities." *The Atlantic*, August 29.

State Council. 2025. "China Issues Guideline to Accelerate 'AI Plus' Integration across Key Sectors." The People's Republic of China, August 27.

Stephenson, Randall, Johanna Rinceanu, and Marc André Bovermann. 2025. "Regulating Political Advertising in the EU." *Eucrim* 20 (1): 90–97.

Strawson, Galen J. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75: 5–24.

Swanson, Ana, and Meaghan Tobin. 2025. "China's Rare Earth Restrictions Aim to Beat U.S. at Its Own Game." *The New York Times*, October 16.

The Guardian Staff and Agencies. 2024. "Chip War Ramps up with New US Semiconductor Restrictions on China." December 3.

The White House. 2025. *Winning the Race: America's AI Action Plan*. Washington, DC

Tran, Ha-Chi. 2025. "Bridging Liability Gaps in the Age of AI: The Case for No-Fault Compensation Schemes." *Proceedings of the AAAI/ACM Conference on AI Ethics and Society* 8 (3): 2930–2932.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

United Nations General Assembly. 2025. "Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance." A/RES/79/325.

United Nations Trade & Development (UNCTAD). 2021. "Technology and Innovation Report 2021: Catching Technological Waves – Innovation with Equity". *United Nations Publications*.

United Nations Trade & Development (UNCTAD). 2025. "AI Market Projected to hit $4.8 trillion by 2033, Emerging as Dominant Frontier Technology." April 7.

Villasenor, John. 2025. "How Overly Aggressive Bans on AI Chip Exports to China can Backfire." Brookings, August 15.

Voigt, Kristin. 2020. "Relational Egalitarianism." *Oxford Research Encyclopedia of Politics*.

Von Neumann, John, and Oskar Morgenstern. 1953. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Waltz, Kenneth. 1979. *Theory of International Politics*. New York: McGraw-Hill.

Wang, Jufang, Claire Milne, Jess Zichen Hu, and Furqan Khan. 2024. "Navigating Geopolitics in AI Governance." *Oxford Global Society Research Report*.

Weise, Karen, and Cade Metz. 2023. "When A.I. Chatbots Hallucinate." *The New York Times*, May 1.

Wellman, Christopher Heath. 2011. "The Egalitarian Case for Open Borders." In *Debating the Ethics of Immigration: Is There a Right to Exclude?* edited by Christopher Heath Wellman and Phillip Cole. Oxford: Oxford University Press.

Westerstrand, Salla. 2024. "Reconstructing AI Ethics Principles: Rawlsian Ethics of Artificial Intelligence." *Science and Engineering Ethics* 30 (5): 1–21.

Williams, Bernard. 1973. *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge: Cambridge University Press.

Williams, Bernard. 1981. *Moral Luck*. Cambridge: Cambridge University Press.

Williams, Bernard. 2005. *Realism and Moralism in Political Argument*, edited by Geoffrey Hawthorn. Princeton: Princeton University Press.

Wilmshurst, Elizabeth. 2019. "Strengthen the International Criminal Court." Chatham House, June 12.

Wolff, Jonathan. 1998. "Fairness, Respect, and the Egalitarian Ethos." *Philosophy & Public Affairs* 27 (2): 97–122.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Wolff, Jonathan. 2015. "Political Philosophy and the Real World of the Welfare State." *Journal of Applied Philosophy* 32 (4): 360–372.

Womack, Brantly. 2024. "Strategic Principles for Co-Existence in a Multinodal World." *China International Strategy Review* 6: 22–38.

Wong, Brian Yue Shun. 2025a. "Structural Unemployment and Structural Injustice." *Inquiry*: 1–34.

Wong, Brian Yue Shun. 2025b. *Reparative Justice in Authoritarian States*. London: Routledge.

Wong, Brian Yue Shun. 2025c. *Moral Debt: Defending a New Account of Reparative Justice*. London: Routledge.

Wong, Brian Yue Shun. 2026. "Towards A Confucian Ethics Governing Human-AI Relations." In *Dao Companion to Applied Confucian Ethics*, edited by Julia Tao, Yu Kam-por, and Rina Camus. Berlin: Springer.

Wong, Matteo. 2025. "The Great AI Lock-In Has Begun." *The Atlantic*, April 22.

World Trade Organization. 2004. "The Process – Stages in a Typical WTO Dispute Settlement Case." Dispute Settlement System Training Module.

Wu, Zijing. 2025. "China Seeks to Triple Output of AI Chips in Race with the US." *The Financial Times*, August 27.

Yang, Cheng, Xuemeng Yang, Licheng Wen, Daocheng Fu, Jianbiao Mei, Rong Wu, Pinlong Cai, et al. 2025. "Learning on the Job: An Experience-Driven Self-Evolving Agent for Long-Horizon Tasks." arXiv preprint arXiv:2510.08002.

Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University Press.

Young, Iris Marion. 2003. "Political Responsibility and Structural Injustice." The Lindley Lecture, The University of Kansas.

Young, Iris Marion. 2006. "Responsibility and Global Justice: A Social Connection Model." *Social Philosophy and Policy* 23 (1): 102–130.

Young, Iris Marion. 2011. *Responsibility for Justice*. Oxford: Oxford University Press.

Zarakol, Ayşe. 2017. "Why Hierarchy?" In *Hierarchies in World Politics*, edited by Ayşe Zarakol. Cambridge: Cambridge University Press.

Zhang, Angela Huyue. 2024. *High Wire: How China Regulates Big Tech and Governs Its Economy*. Oxford: Oxford University Press.

Zhang, Angela Huyue. 2025. "US Export Controls have Forced Chinese tech Companies to be More Innovative." *The Financial Times*, January 23.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Zhang, Juan, and Yawei Liu. 2025. "DeepSeek and U.S.-China AI Competition w/ Brian Wong." *The Monitor*.

Zhao, Xinlei. 2024. "Technological Hedging and Differentiated Responses of Southeast Asian Countries to U.S.–China Technological Competition: A Case Study on Artificial Intelligence (AI)." *The Pacific Review* 38 (3): 1–32.

Zhou, Jiawei, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. "Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* 1 (436): 1–20.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Acknowledgements

Both authors contributed equally to the preparation of this manuscript and should be credited as co-authors.

B.B. reports a grant from the Government of Hong Kong, SAR China, University Grants Committee, General Research Fund, for the project *The Legal and Ethical Foundations of Generative Models* (Project Code 17616324), and support from the HKU Musketeers Foundation Institute of Data Science and the HKU100 Initiative.

B.W. credits the HKU100 Initiative for its generous support. B.B. and B.W. would also like to thank those who provided feedback at various workshops, lectures, and talks at the Foreign Correspondents' Club of Thailand, Fudan University, the Institute of Security and International Studies (ISIS) at Chulalongkorn University, the Malaysian National Resilience College, Peking University, the Philosophy of AI Summer School in Paris, and Tsinghua University, for their insights and comments.

In addition, funding from The University of Hong Kong Faculty of Arts, Hong Kong Ethics Lab, and the AI & Humanity Lab made it possible for this Element to be published open access, making the digital version freely available for anyone to read and reuse under a Creative Commons licence.

Thanks also go to Anna Chan, Xuan Xiang, Ingrid Yeung, Louie Lang, and Zoey Zhou for their excellent research assistance.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Cambridge Elements

Philosophy and AI

Herman Cappelen

University of Hong Kong

Herman Cappelen is a Chair Professor at the University of Hong Kong and the founder and director of the AI&Humanity-Lab. He has worked in many areas of philosophy, including philosophy of language, conceptual engineering, meta-philosophy, and the philosophy of AI. His most influential works include *Fixing Language: An Essay on Conceptual Engineering* (OUP, 2018), *Philosophy without Intuitions* (OUP, 2012), *The Concept of Democracy* (OUP, 2023), *Making AI Intelligible* (OUP, 2021), *Relativism and Monadic Truth* (OUP, 2009, with John Hawthorne), *Insensitive Semantics* (Blackwell, 2005 with Ernie Lepore), *The Inessential Indexical* (OUP, 2013, with Josh Dever). Cappelen and Josh Dever have written a three-volume introduction to contemporary philosophy of language: *Context and Communication* (OUP, 2016), *Puzzles of Reference* (OUP, 2018), and *Bad Language* (OUP, 2019). Cappelen has previously been a Chair Professor at The University of St Andrews, an associate professor at Oxford University, and a professor at The University of Oslo.

About the Series

The burning question of our time is what it means to be human in the age of Artificial Intelligence. Philosophy plays a unique and crucial role in answering this question. Philosophers investigate whether AI can understand language, be creative, possess emotions, be conscious, have intentions, make plans, execute intentional actions, be moral or immoral, empathize, possess knowledge, and exhibit wisdom. The answers to these and many related questions not only are intellectually challenging but also have potential real-world implications for how AI is developed and used. Cambridge Elements in Philosophy and AI explores these questions and their ramifications. Each Element provides a survey of the literature, which will be a reliable resource for researchers and students, and also develops new ideas and arguments from the author's viewpoint. The authors of the Elements include some of the most prominent senior figures and up-and-coming junior scholars in the field.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904

Cambridge Elements

Philosophy and AI

Elements in the Series

Emerging Questions in AI Welfare
Geoff Keeling and Winnie Street

AI and Consciousness: A Skeptical Overview
Eric Schwitzgebel

Geopolitics of Artificial Intelligence
Boris Babic and Brian Wong Yue Shun

A full series listing is available at: www.cambridge.org/EPAI

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 21:16:10, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009731904