Asian Journal of Law and Society (2026), 1–20
doi:10.1017/als.2026.10053

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

AI Sovereignty, Data Bias, and Ethical Challenges in International Law

Noble Po-kan Lo[1,2]

[1]College of Professional and Continuing Education, The Hong Kong Polytechnic University, Hong Kong and [2]Lancaster University, Lancaster, Lancashire, UK
Email: noble.lo@cpce-polyu.edu.hk

Abstract

As artificial intelligence (AI) increasingly shapes global governance, the concept of AI sovereignty raises significant legal and ethical concerns. A central challenge lies in systemic bias within dominant AI models, which are predominantly trained on Western data, languages, and cultural perspectives. Such bias influences content generation, legal interpretation, and historical representation, potentially marginalising non-Western knowledge systems. This paper examines the implications of biased training data for AI sovereignty in Asian legal and geopolitical contexts, focusing on linguistic inequality, restricted access to regional datasets, and AI-driven content moderation. It analyses how these factors may generate information asymmetries and constrain national regulatory autonomy. By comparing Western-led governance frameworks with alternative state-centric regulatory models, the study evaluates whether localised AI development or international regulatory cooperation better addresses algorithmic bias, digital censorship, and human rights protections in an increasingly fragmented global AI order.

Keywords: AI sovereignty; data bias; algorithmic governance; human rights in AI; geopolitics of AI

1. Introduction

1.1. Background and context

Artificial intelligence (AI) can be defined as any machine-based system that, for a given set of human-defined objectives, makes predictions, recommendations, or decisions that influence real or virtual environments. Whilst there are many different technical definitions of AI which can be used, it is most commonly accepted that the definition requires a machine to be capable of solving a problem which a human brain or mind might struggle with, or which solves a problem by acting in a rational way in a manner similar to which a mind appears to work through (Arroyo, Schöttle, and Christensen, 2022, p. 69). Although these systems are not human, their outputs are often indistinguishable from human-generated content, complicating our traditional understanding of agency. These machines learn by analysing vast datasets which a single human could never hope to synthesise or process, and as they become more and more sophisticated, they begin to have ever greater potential use cases. Already, AI is being used widely by private industry and technology companies in marketing and in research and development (R&D) around the world. In the public sector meanwhile, AI has the potential to revolutionise the

© The Author(s), 2026. Published by Cambridge University Press in association with Shanghai Jiao Tong University. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

performance and efficiency of the state and of its governance. AI might potentially be used to help draft laws, interpret legislation, support policing, and assist with the drafting of policy in many different parts of public administration. It might be used to assist doctors in screening, diagnosing, and treating patients. In some areas, such as the use of AI facial recognition technology by the police, this is already controversial. In the future, risk assessments carried out by AI agents might be able to anticipate potential crime or criminals before they ever commit an offence, creating suggestions of "pre-crime" (Simmler and Canova, 2025). From this, it is clear that AI has the potential to play a very important, but also potentially intrusive role in people's lives whether they want it to or not in the future.

A number of problems immediately become apparent from this. The first is that whilst AI systems are designed to be able to solve problems which a human mind might solve, the processes and parameters which a given machine might use to solve that puzzle are largely incapable of being interpreted by a human (Gignac and Szodorai, 2024). Therefore, while an AI system generates a response, it does so by mimicking patterns rather than through the conscious intent typical of human cognition. As the machine in question learns from other machines and potentially billions of data points, it grows ever more complex, and its workings become more and more esoteric for the human seeking to understand it. This is the so-called "black box" problem (Papadouli, 2023, p. 41) and refers to the inherent lack of transparency complex AI system display wherein the developers are unable to review or trace the internal logic by which any decision is made. Because these systems operate through multi-layered neural networks that identify patterns across potentially millions or billions of data points, the decision-making process remains opaque, making it difficult to audit the system for errors or hidden biases before they result in real-world harm. Second, we know that human cognition results from decisions being reached by the processing of millions or potentially billions of pieces of data, by the brain, against a matrix of experience, learning, and schemas which the individual has developed over the course of their life (Gignac and Szodorai, 2024). This matrix operates as a filter and guide for decisions to be made in, and operates to help prioritise information so that a decision can be made. The filter, however, also acts as a form of prejudice or bias (Woods, 2023, p. 123). Some cognitive scientists suggest that we may make a decision almost immediately on the basis of information presented to us as a result of our inherent biases or cognitive shortcuts (Woods, 2023, p. 123) before the thinker is even consciously aware that they have made a decision, and their entire thought process subsequently acts as a justification for it. The risks of discrimination being exacerbated here are not merely theoretical moreover but have already been documented in the performance disparities of widely used systems. For example, a landmark study on facial recognition technology revealed that while error rates for lighter-skinned males were as low as 0.8%, they spiked to nearly 34.7% for darker-skinned females (Buolamwini and Gebru, 2018). These statistical failures show that when AI is trained on non-representative datasets, it does not merely reflect human prejudice, but also helps to scale it, worsening systemic barriers and prejudice.

Whilst AI is not "coded" in the information it learns, human beings did perform the initial coding and creation of the machinery which allows the AI (or at least their first generation) to learn, creating a risk of biases being passed on. Human creators and trainers of AI and the selection of these trainers may also introduce bias into the training process. Even the specific languages used in the coding and training may be influential here. Finally, human beings often act to fine-tune or reinforce machine learning, and there are therefore several avenues where human bias may be either introduced or reinforced in an AI program. When the consequences of AI might be so significant as being important for a person's health and liberty as may one day be the case, it is essential that these bias risks are minimised, so that injustices are avoided, but the black box nature of AI makes this difficult.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

This all means that the global architecture of AI is being laid at present upon a highly fractured foundation, filled with potential bias. When the global inequality of the race to develop and deploy AI systems is added to this, the risks of AI potentially exacerbating and perpetuating racial, ethnic, national, religious, or cultural injustices and biases in many parts of the world become a serious concern. After all, whilst there have been notable Asian success stories related to AI development (notably China's "DeepSeek" LLM), the vast majority of funding and research into AI has historically been concentrated in a relatively small geographical area in the United States where it has been developed by corporations. The dominance of Western-centric datasets in foundational AI models creates a critical sovereignty gap for non-Western states. When a nation relies on foreign AI to assist in public administration, law enforcement, or judicial decision-making, it effectively "outsources" its normative standards to a system that may not recognise its unique legal traditions, histories, or social values. Consequently, the pursuit of AI sovereignty is not merely a matter of technological or economic competition, but a fundamental requirement for states to guarantee that their governance domestically remains aligned with their own cultural and legal settling, rather than being overwritten by a global digital hegemony.

Most AI programmes currently in use or under development have, therefore, been trained in English, using predominantly Western data sources, within a corporate and a Western cultural context. As such, the paper will examine how linguistic disparities, access to datasets (or restrictions arising from lack of access to data under data transfer agreements or treaties), and AI-driven content moderation can all contribute to information asymmetries which disadvantage those nations which lack control over AI data training. This paper will examine differences in approaches to the governance of AI in the United States, the United Kingdom, the European Union (EU), and China, where efforts to secure sovereignty over home-grown AI has received prioritisation in recent years (Chang, Arcesati and Hmaidi, 2025). The United States and United Kingdom seem so far to be intent on allowing the corporate world (a corporate model) and market to develop AI models, whilst the EU seems intent on regulation (regulatory model), seeking to ensure digital sovereignty through requiring foreign AI models to comply with their own standards before being used in the EU. China is, by contrast, being built with sovereignty in mind (a sovereignty model).

This paper will analyse the convergence of technological power and systemic bias in order to identify the risks it poses around the world. The corporate-dominated, Western-centric dominance over the early foundational AI models on which future AI iterations are likely to be trained creates risks in at least three areas: national sovereignty, human rights, and the ongoing development of the Global South (i.e., the developing world).

This paper will assess, in light of this, whether domestic AI sovereignty is desirable, and whether nations or culturally, politically, and linguistically close regions (such as the EU perhaps) ought to seek to develop localised AI models to mitigate bias challenges. This paper will also consider the alternative to this, which is to seek to ensure international cooperation in AI development which aims to provide international equality in terms of access to datasets, diversity in AI training, and whether this can ever practically be achieved. How global AI regulations can effectively implement protections against the risks to human rights posed by the use of AI in surveillance and censorship will also be examined, as how different cultural attitudes and governance models may politically make the harmonisation of AI data training under a cooperative model more difficult will also be examined.

1.2. Research aims and objectives

The main research aim of this paper is to assess whether AI sovereignty is an important tool for countries to help protect their own national sovereignty, human rights within

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

their borders and elsewhere, and the ongoing pursuit of international equality and economic development in the Global South.

In answering this, this paper is focussed around the following research questions:

1. Is there a Western-centric bias inbuilt into existing AI models resulting from the Western and corporate dominance of these platforms?
2. How does this bias challenge existing legal, cultural, and political sovereignty of states outside the Western world, particularly in Asia?
3. In what ways do competing AI governance models—in the form of the Western corporate-led approach on the one hand and China's sovereignty-focussed model on the other—impact the protection of sovereignty, human rights, and development?
4. Can prioritising AI sovereignty help minimise the risk of Western-centric AI data biases and their consequences for cultural hegemony, human rights threats, and threats to economic development.
5. Can cooperation between nations effectively mitigate AI algorithmic bias and ensure equitable AI development, or is the pursuit of national AI sovereignty more likely to provide this?

1.3. Methodology

The methodological approach to be used here is a combination of both doctrinal legal methodology along with qualitative theoretical research and a comparative legal and policy-based analysis of the approaches taken to the regulation of AI development in the West and in China.

The theoretical nature of the research questions posed by this paper precludes the use of a purely doctrinal analysis, which would look only at the nature of the law and legal frameworks around AI regulation. It will, instead, be necessary here to engage in a comparative and functionalist analysis of the governance models relating to AI development and use in the chosen regions of study, namely the United States and China.

From the theoretical perspective, therefore, the methodology is informed by Third World Approaches to International Law (TWAIL) which consider the way in which international law helps to exacerbate and perpetuate existing inequalities and historic injustices (Anghie, 2023, p. 48). This is highly relevant in explaining the sovereignty, development, and human right risks which arise from a developing state being denied AI sovereignty and being required to accept the imposition of a Western model of AI instead.

2. AI sovereignty conceptualising control in a fractured technological landscape

2.1. What is AI sovereignty?

AI sovereignty may be regarded as the "ability to independently control the infrastructure, data, models, and policies" that power an AI system (Vohra, 2024). The desire to control the operation of AI and to run it along the lines of national interest and sovereignty are part of a growing trend in Asia and other parts of the world for enhanced levels of digital sovereignty more widely (Woods, 2023, p. 123).

For states in the developing world, the fact that most commercially available AI products and programmes have been developed in the Western world is one such issue (Roberts, 2024). The same is true for governments of countries outside the Western world,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

such as those in Asia, particularly as one of the most important facets of machine learning that underpins the training of new AI programmes is that these models learn from datasets already established and contributed to by earlier generations of AI technologies developed elsewhere in the world (Franklin et al., 2024). If there is bias built into these early generations of AI models, or in the models currently available and used widely in either government or the private sector around the world, then there is a risk that this will self-perpetuate in the future as a result of the way AI systems learn. The risk of self-perpetuating bias extends far beyond narrow government applications, creating a feedback loop across the entire socio-economic landscape. If, for example, foundational AI models used in the private sector for healthcare diagnostics, credit scoring, or recruitment are built on non-representative datasets, they could generate biased real-world outcomes that are then captured as "new" data. This can be seen by way of a real world example. In a study by Obermeyer and others (Obermeyer et al., 2019), an AI-driven health system, for example, consistently deprioritised care for certain demographics based on biased historical spending data creates a new dataset where those demographics appear "healthier" or "less in need," thus validating and reinforcing the original bias. This was built off an original bias which resulted from these groups being provided with less resources in the first place leading to a lower degree of engagement with healthcare in the first place and lower levels of spending being provided. This results in an automated systemic inequality where the private sector's technological efficiency inadvertently hard-codes historical discrimination into the future infrastructure of societal well-being, making it increasingly difficult for later sovereign regulations to unpick these already embedded prejudices.

Perceptions of these bias problems help drive a desire in some states for AI sovereignty. For many developing states, however, the high capital and technical barriers to creating truly autonomous systems often preclude the pursuit of absolute national AI sovereignty. In the absence of such capacity, these nations frequently find themselves managing the expansion of Western AI through a position of technological dependency where they are forced to adopt foreign models that may be ill-suited to their local linguistic and cultural contexts.

Ensuring that AI sovereignty is retained by a state has, therefore, become an important issue. In the United Kingdom, the UK Government has recently established its own "Sovereign AI Unit" intended to promote the development of sovereign UK AI tools and programmes, to encourage investment in the development of a UK-based AI sector, and to identify and establish use cases for AI in government and governance of the country. The UK Government has even suggested that AI sovereignty represents a question of fundamental national security, given the importance of this concept (HM Government, 2025). Britain is not alone in the pursuit of AI sovereignty; China's policy towards AI is to develop an autonomous and strategically independent and sovereign AI sector, at every level of the technology, from the manufacture of computer microchips required to the development of its own Large Language Model (LLM) models (Chang, Arcesati and Hmaidi, 2025). China's stated goal of such a policy is to establish an "independent" model of AI which it controls fully.

It is this which helps to explain what we mean when discussing the concept of AI sovereignty. AI sovereignty refers broadly to the control which a given state has over the development, training, deployment, and uses of an AI programme (Woods, 2023, p. 125). The current movement towards AI sovereignty is best understood as a reaction against the historical trajectory of the internet. To understand why states are now asserting territorial control over algorithms, one must first examine how the borderless nature of the early web established the very cultural and legal asymmetries that AI now threatens to entrench.

The reality is that the open nature of the internet is a reflection of the culturally hegemonic nature of the Western world and its culture, priorities, and objectives at the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

time at which the internet became widely adopted globally. Much of the programming of what would become the "world wide web" was performed by computer scientists educated in a Western cultural milieux; As Woods notes, however, the promotion of the "free" internet was also aligned with the United States' (and other Western countries such as the United Kingdom's) own understandings of open democratic governance and individual liberty (Woods, 2023, p. 123). Promoting the internet as being an entirely free domain, removed from and suspended above the geographically bound laws of any given nation state, was regarded as being perfectly consistent with US political ambitions and was even an element of the state's propaganda and push for global cultural influence.

In short, the internet posed a real threat to China's stability, its development, and its governance model due to its open nature resulting from its development primarily in the West. China had to dedicate significant levels of resources to effectively modify and retrofit the internet to better match its own cultural and political understandings and settlement through its initiatives such as the Great Firewall (Gravett, 2020). This retrofitting of Chinese governance models onto a Western technology allowed China to exercise geographic digital sovereignty, and to carve out of cyberspace a geographic China-shaped hole in the way in which the world wide web otherwise operated. This came with significant difficulties, however, not the least of which was seeking to convince its own population that these restrictions were necessary once the global culture of the internet had been configured to a default setting of openness (Gravett, 2020).

The experience may well have left Chinese policymakers all too aware of the risks which a lack of digital sovereignty creates for a nation, and of the importance of influencing the genesis of a system, rather than being required to adapt it to their needs later. This history is critical because it explains the centrality of territoriality in modern AI sovereignty. States like China, having experienced the "threat" of an open internet to their internal stability, now seek to avoid a repeat of this algorithmic colonisation (Appleton, 2026). By insisting that AI infrastructure, data, and logic must reside within and obey the laws of a specific geographic jurisdiction, these states are attempting to "retrofit" territorial boundaries onto a technology that would otherwise default to Western norms. Thus, AI sovereignty represents a shift from the "open" internet model to a "managed" model where technological logic is legally tethered to the state's territory. This can be seen in China's approach to AI sovereignty, set out in its "Global AI Governance Initiative" (Embassy of the People's Republic of China in Grenada, 2023). These principles emphasise China's own understanding of what AI ought to be used for, not just in China, but globally, with the priority being stated to be to "increasing the wellbeing of humanity . . . so that AI develops in a way that is beneficial to human civilization." In addition to this, however, China also emphasises the need to ensure sustainable development and environmental pretection, as well as ensuring that AI respects "national sovereignty and strictly abides . . . by their laws when providing them with AI products or services," whilst the potential use of AI for the purposes of "spreading disinformation," "manipulating public opinion," or "intervening in other countries' internal affairs" is officially opposed by China under this initiative (Embassy of the People's Republic of China in Grenada, 2023).

Two things, in particular, can be seen from this approach. The first is that, on a practical level, China's initiative clearly aligns with its own national priorities and its concerns; maintenance of internal stability and avoiding internal manipulation by foreign powers or agents is prioritised, as is the insistence upon respect for state sovereignty and the application of national law to AI within established geographic legal jurisdictions. Second, however, it is clear from this that China wishes to assert its own understandings in the framing of AI from as early a stage as possible, so that the technology does not develop globally without the country's own input being relevant as initially happened with the internet in the past. This then is an assertion of national sovereignty in the development and use of AI. It has within it, however, another goal which is to help shape and funnel the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

international development of AI along these lines. In this way, an apparent dichotomy between the idea of national sovereignty and an opposing idea of international cooperation, or pooling of sovereignty, is effectively accepted as being an element of the same national ambition.

China's understanding of AI sovereignty is only one view here. There are, as Woods suggests (Woods, 2023, p. 123), many different approaches which could be taken to the exercise of AI sovereignty by states. Whilst some might seek to assert their own national interests through establishing their own, wholly autonomous domestic AI industry, others might determine that it is likely to be more effective to seek to shape the international framework of AI development at an early stage so as to ensure that they do not later become subject to a system they had no say in developing. In contrast to China's state-led approach, the United States exercises a form of "de facto" AI sovereignty through market dominance. Rather than centralising control within government agencies, the US model relies on the global reach of its private sector, specifically "Big Tech" firms, to establish American technical standards, ethical frameworks, and linguistic norms as the global default (Perisic, 2024). For the United States, sovereignty is maintained by ensuring that the foundational infrastructure of global AI remains rooted in the American corporate ecosystem, thereby projecting influence without the need for the rigid territorial "firewalls" seen in other jurisdictions.

2.2. *National sovereignty versus pooled sovereignty: the conceptual divide*

Whilst some states might see their AI sovereignty through the prism of absolute autonomy, it is also possible to see AI sovereignty as being capable of being pursued by the pooling of sovereignty between different states, through collaboration and sharing of both risks and knowledge.

Pooled sovereignty refers to the systemic attempt to strengthen a nation's exercise of sovereign power by sharing this and resources with other states. This ought to be distinguished from mere international cooperation, because pooling of sovereignty often means delegating or conferring power to make some decisions on a common institution. By contrast, cooperating internationally means working together with other states in pursuit of shared goals, through treaties, but without any loss of individual sovereign power or autonomy. Sovereignty pooling is often regarded as a reaction to the reality of modern, complex problems which have global or regional contributory causes and which cannot be effectively resolved alone, by any single country. With regard to the establishment of AI systems, pooling sovereignty might be seen as a rational response to the practical challenges posed by AI systems, particularly in respect of the entry barriers and costs associated with the development and training of AI models (Cottier et al., 2024).

For countries without the financial might of a superpower such as China or the United States, the pooling of sovereignty and resources might, therefore, provide an essential means of ensuring "leadership" of AI and in establishing a degree of sovereignty which can help them to avoid the worst consequences of bias, as the Tony Blair Institute for Global Change (Tony Blair Institute for Global Change, 2025) suggests. Pooling resources and sovereignty into creating and training a shared AI model allows the contributors to provide the AI with the initial programming language and infrastructure which they regard as desirable. It also allows these contributors to retain control over the datasets which are used to train these programmes. In other words, there is a chance that by working in collaboration with other states, even medium-sized states can ensure that they remain sovereign in respect of AI systems which they use or develop, at least to some degree.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

Whilst this pooling of sovereignty is one means by which countries today might seek to obtain some control over the development of AI in a cost-effective manner, it is important to note that there are likely to be drawbacks to engaging in a shared form of AI sovereignty over obtaining absolute, sole control over such a system. In the first place, not all of the contributors in question are likely to agree precisely on the parameters involved, states might disagree over the governance model to be used, and there may be disagreements over language used for training or financial or other work sharing commitments or contributions. Even here, however, some such as Ishkhanyan (Ishkhanyan, 2025) suggest that there are tools to mitigate these problems. For Ishkhanyan, collaboration over AI sovereignty can be most effective if the contributors agree to adopt what the author terms a "digital federalism" model of governance, built on principles of subsidiarity and multi-tiered sovereignty (Ishkhanyan, 2025). The EU's might be expected to be the most prominent such governance model for AI which establishes a base set of minimum standards and seeks to harmonise legal and regulatory ambitions and boundaries for example at the highest level of governance, but which then allows its own Member States to provide the precise legislative framework regarded as being desirable within those limits.

2.3. *Legal autonomy versus interdependence*

Deciding whether a state might wish to retain absolute sovereignty over AI, or to engage in a cooperative model, is sometimes referred to as being the "sovereignty-internationalism paradox," implying an inherent tension between sovereignty on the one hand and collaboration on the other (Ishkhanyan, 2025). The reality is, however, that the practical and legal realities of AI are more complex than such a neat paradigm allows for. As Srivastava suggests, the structural complexity of AI and machine learning means that it is functionally impossible to isolate an AI model from global learning in some ways, and thus, it is best to reconceptualise sovereignty not as being merely a way of exerting absolute, sole control over a given programme, but rather as being a hybrid form of authority resulting from the various interactions which the state has with others, and with corporations (Srivastava, 2022).

This is fair particularly when considering the prevailing "Western" model of AI governance which has been adopted in the United States for example, where legislators and regulators have sought to facilitate and legalise the use of training of AI programmes by corporations or natural or legal persons rather than to seek to ensure absolute state control over the technology. In the United States, there is as of yet no formal federal legislation governing the use or development of AI. Instead, an Executive Order from the office of the President issued in 2023 established a "Framework for Risk Management" to be issued by the US National Institute of Standards and Technology, whilst various US states are themselves entitled to make their own legislation on the development and use or training of AI. This approach has been criticised as providing a fragmentary approach which creates risks of misuse whilst also imposing between "$2 and $6 million compliance costs per firm" as they seek to navigate this regulatory labyrinth (White, 2025). The Trump administration, however, has been clear in its own "action plan" for AI regulation that the executive branch wishes to remove as much regulation as possible to allow the technology to develop at pace according to commercial incentives (The White House, 2025). The objective of this is to ensure that the United States achieves effective AI sovereignty not through state control, but through dominance culturally, with the language, programmers, testers, and headquarters of global AI efforts being based in the United States. For that reason, the Trump administration has actually proposed a federal moratorium of ten years on any regulation or legislation which might interfere with or retard the development or deployment of AI in the country (O'Donoghue, 2025). When seen in this way, the United States' approach to "sovereignty" is not one which appears monolithic. It is instead closer

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

to this sort of hybrid form of sovereignty described by Srivastava, which identifies sovereignty as being hybridised authority, coming from a combination of factors such as relationships and corporate partnerships rather than from absolute control over every element in the chain (Srivastava, 2022).

The EU meanwhile has taken a different approach, with their model representing more closely the federalised or decentralised model of governance suggested by Ishkhanyan as being one of a supranational, federal model (Ishkhanyan, 2025). This is centred around the EU's "AI Act" (Regulation (EU) 2024/1689) which harmonises the top layer of rules applicable to AI in the EU on a risk-based approach. Of the highest level of concern under these rule include AI models which deploy subliminal techniques beyond a person's conscious awareness or which purposefully seek to manipulate or use deceptive techniques with the objective of materially distorting the behaviour of a person or group (Article 5(1) Regulation (EU) 2024/1689) or which seek to exploit the vulnerability of a person or group due to their age, disability, or social or economic status for example, or which result in the biometric identification and categorisation of people (Article 5(2)(3) Regulation (EU) 2024/1689). These types of highest risk AI programmes are prohibited from being placed on or operating within the EU. Beyond this, a group of high-risk AI programmes such as those which perform profiling of natural persons (and which might therefore be widely used or developed for use by the state and its law enforcement agencies for example) or those which are used inside products falling under the EU's product safety legislation are regarded as being subject to certification and continued assessment through the life cycle of the programme after its approval for entry onto the market is allowed (Annex III Regulation (EU) 2024/1689). Beyond these rules, however, the EU's own Member States are free to determine their own approach to AI regulation in areas not covered by the Regulation, as long as their own rules do not come into conflict with these (under normal principles and doctrines of EU law such as the doctrine of the supremacy of EU law (*Costa v ENEL Case* C-6/64 [1964] ECR 585). The EU's governance model is then to allow the development of AI in the EU, whilst setting agreed upon minimum standards before which entry to the EU's own internal single market can take place. This is reflective of a model which seeks to promote AI sovereignty through encouraging investment and practical autonomy. Actual positive efforts to develop a fully "sovereign" EU version of AI, however, is not built around either collaborating with other partners in the creation of top-down, centralised, and controlled system of AI, but in encouraging the development of facilities through fostering legal certainty and clarity. The hope here is that this will encourage investment in the EU, and the single-market's rules will then encourage the establishment of supply chains such as microchips and data centres within the EU, helping practical autonomy.

All of this shows that there are competing understandings of what AI sovereignty actually means. Whilst it would undoubtedly be preferable for the EU to ensure total control and autonomy of the entire AI supply chain, this is in practice very difficult or impossible to achieve, without adopting a top-down high level of state control over the industry such as that which China has sought to achieve (Ishkhanyan, 2025). There are drawbacks to such a model as that seen in China too, including the risk of regulatory chill and frustrating investment and the potential "railroading" of technological development before it has a chance to organically develop. Furthermore, the reality is that even a country like China might not be able to guarantee full effective practical autonomy over every link in the chain needed to guarantee full AI sovereignty.

Finally, there is a third possible path for ensuring sovereignty here, in the form of joint collaborative efforts being made by regions or neighbouring allied countries to develop an interoperable AI system. Whilst the EU's supranational legislative effort is one such model, it still relies at its base on this hybrid form of sovereignty, where it must work in partnership with private capital, and can never guarantee (without further

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

legislation at least) that certain essential aspects of functional autonomy, such as microchip supply, data centre locations, or training literature, are controlled by these states. Whether or not any of these paths are necessarily more likely to guarantee sovereignty over AI in the future remains to be seen.

Whilst it is likely that China's attempts to autonomously control the various elements of the AI supply chain may be successful, it is not clear whether this will defeat entirely the bias challenges posed by AI training and machine learning, nor whether this model can be replicated in other Asian states which might likewise be concerned by the Western dominance of initial AI programmes on the market. These nations cannot guarantee practical autonomy for AI supply chains and lack a regional legal framework as deep as the EU's. The EU's treaties, after all, allow for real harmonisation within the domestic legal systems of its Member States on a level which other regional trade blocs like the ASEAN have not yet replicated (Chirathivat, Kunnamas, and Welfens, 2020, p. 556). For these countries then, it may well be preferable to seek to obtain some degree of AI sovereignty in another manner, through a treaty-based system of regional collaboration, technology sharing, and investment pooling. This is the suggestion which the Tony Blair Foundation suggests that other countries (such as the United Kingdom, which is also outside of the EU and its integrated single market) ought to consider adopting, to spread costs and ensure sovereignty through agreement with others (Tony Blair Institute for Global Change, 2025). Asian nations, such as Japan, Singapore, Thailand, India, and Malaysia, could, for example, consider engaging in a regional collaboration to develop and use a regional AI system and to supply its supply chains, training, and scope of pursued and prohibited applications. Without such cooperation, the gap between their legal autonomy and their ability to exercise meaningful control will remain vast, leaving them vulnerable to the biases and strategic interests embedded in foreign AI systems. The following chapter will delve into the precise nature of these biases and the concrete risks they pose.

3. Data bias and AI learning

States have a desire to ensure AI sovereignty for a number of reasons. One might be that other states wish to develop AI for other uses than they themselves have in mind for the technology; other states might have more concerns over the implications for human rights over potential use cases which a state wishes to consider. Having sovereignty over AI precludes such a problem. AI is also expected to be the basis of a very important and growing new digital economy in the coming years. Ensuring sovereignty helps lock in investment and jobs in the state itself (Garcia, 2019, p. 83). One of the most important reasons that drive a perceived need to ensure AI sovereignty in the developing world and in Asia is the concerns over bias stemming from the Western domination of early iterations of the technology. Seen in this way, the pursuit of AI sovereignty is explained best as being a response to a perceived threat: that of a foreign nation's values and norms becoming inextricably embedded in a core technology, just as the internet was. In fact, seen from the perspective of TWAIL, ensuring AI sovereignty for the Global South represents an essential tool in resting cultural imperialism and dominance (de Sousa Santos, 2024, p. 636). By obtaining AI sovereignty, and retaining that, the Global South can help to preserve its own cultural heritage, whilst taking advantage of the economic and developmental gains which might come from having an autonomous, sovereign AI system established in their own territory.

The issue for this paper is to explain the reality of these concerns, by examining how bias in AI systems arises, and why it is so potentially harmful to those in Asia and in the Global South more generally.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

3.1. The basis for bias: how bias enters systems

Bias is a human cognitive function which arises when a person makes a decision (Berthet, 2022). The effect of bias is that the decision-maker makes their decision contrary to a norm or rational outcome, but there are various causes to this. The first is the idea of "heuristics," or that the brain seeks to use "simplified information processing strategies" or heuristic shortcuts when making decisions which tend to lead to repeated, recurring types of biases (Berthet, 2022). Some of these include things such as confirmation bias (a person is more likely to select as convincing evidence of a belief they already hold over a new, more challenging belief), hindsight bias (to perceive events as being more predictable once it has occurred), or overconfidence bias, where a person might tend to overestimate their own judgment over that of others (Berthet, 2022). In addition to this, however, humans have a vast range of other biases, many of them unconscious and unrecognised, which occur as a result of that individuals cultural background, profession, or even their language (Shah and Bohlen, 2023). These types of biases may be referred to as implicit biases. A person born and raised in a traditional Western cultural environment for example is apt to interpret information they are presented with through their own culture, and assume almost automatically that the same conclusions they reach would be reached in respect of that information by others (Moser, Steiglechner, and Schlueter, 2022). These biases in turn can lead to ethno-centrism (by which the system sees and evaluates information through the lens of a particular culture, language whose values are regarded as an inherent automatic standard) and in turn to a belief in superiority which can become prejudicial against other cultures, beliefs, or races. This form of prejudice is only result which can stem from an implicit, cultural bias; entire societies and belief systems can become infected with them, leading their repetition, as other heuristic biases such as confirmation bias operate to reinforce them.

3.2. How western-centric bias shapes AI systems

Given that bias, either of a heuristic or implicit type, is a fundamentally human failing, it might be surprising to think that a computer system or model could likewise suffer from such irrationality. AI, it must be remembered, does not replicate human thought and is intended only to reach conclusions to problems which human brains might see as being capable of solving, and to present the solution in a manner intelligible to a human observer (Webster, 2025). Nevertheless, bias is a concern for AI, because the way in which AI is trained is on the basis of human interaction and on training data itself produced by humans. If the data used to train an AI system contain bias, as may occur when the majority of training data derive from a single cultural, regional, or linguistic context, the system's foundational assumptions will reflect that bias from the outset. The problem for the Global South here is that most of the major available AI programmes and LLMs used internationally today were trained on data which suffer a critical lack of diversity in origin (Tao et al., 2024). This has subsequently been reinforced by trainers, who would again be likely to be responding to the AI in English, and who would themselves be likely to be biased against the LLM or AI model displaying traits, answers, or reasoning which was unfamiliar to their own cultural or linguistic bias (Xu et al., 2025). It has been found that over 90% of LLMs available in the world have been trained predominantly on English language datasets, by English-speaking trainers (Xu et al., 2025).

AI systems have been shown repeatedly to be susceptible to many of the types of implicit and heuristic biases that humans show, and it may well be the case that this is the result of having been trained and selected by these trainers for displaying such biases. For example, as Navigli, Conia, and Ross write, trainers of early AI systems use data sets which were available, intelligible, and which reflected the things they wanted their LLMs or AI

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

programmes to be able to respond to in a manner which replicated human responses (Navigli, Conia, and Ross, 2023, p. 4). Because the vast majority of these programmes originated in the United States or Western Europe, data selection has favoured English language sources available. Trainers from a Western background would be more likely to assess a model as being a highly performing one if it showed the same biases which they themselves shared; an answer on Shakespeare for example which displayed a Western style of prose, or rationality built on a Hegelian dialectic, would be regarded by that trainer as being of higher quality than it might if it displayed a Confucian interpretation of the answer, or an African-Islamic understanding of the text for example (Tao et al., 2024).

3.3. Consequences of AI algorithmic bias on legal systems and culture

In turn, this can result in algorithmic biases becoming embedded in the code of a machine, leading it to make irrational assumptions, or to lead it to engage in stereotyping, racial, or cultural bias for example. Hofmann et al., for example, found that AI itself was likely to show both selection bias and an implicit cultural or racist bias towards favouring the use of what it had been trained to see as correct use of the English language (Hofmann et al., 2024, p. 147). In the United States, where many different ethnic minority groups use their own form of dialect or language such as African American English, LLMs encountering users of this language format were prone to reaching covert, but highly negative racist stereotypes about the users of this language (Hofmann et al., 2024, p. 147). The researchers suggested that the machines displayed results far more negative than any human stereotypes of African Americans which have ever been even experimentally recorded, with LLMs suggesting that these individuals ought to work less prestigious jobs, be convicted of crimes, or be sentenced to death for crimes more often than white users on the same pattern of facts (Hofmann et al., 2024, p. 147). For any governmental organisation seeking to utilise AI in the justice system, or in law enforcement, in healthcare, or in almost any other area, these results must be of extreme concern. For the Global South, whose users are predominantly not white Americans or Western-centric users, reliance on such a system is fraught with danger and reinforces the need for AI sovereignty. There are suggestions, for example, that states might use AI models to help assist judges in the sentencing of criminals, by providing estimates of recidivism or their likelihood to respond well to certain types of sentences (Cavus et al., 2025). Similarly, AI might be used to assist probation officers in their review of prisoner's rehabilitation prospects (Cavus et al., 2025). Any risk of racial or other bias here which results from the non-sovereign nature of the AI in question creates a real risk of injustice in the state.

AI models also have the potential to shape and influence culture and development themselves. Many LLM models now have millions of daily users (Heersmink et al., 2024, p. 41). There is growing concern at the levels of trust many users place in these AI models, even when there remain known problems in respect of the accuracy of these systems. Lawyers around the world, for example, have been warned over reliance on the use of LLMs in their own case preparation as more than once an advocate has now been censured by a court for seeking to rely on authorities which simply did not exist, and which the LLM had "hallucinated" (Bohannon, 2023). In the United Kingdom, for example, in *R (Ayinde) v London Borough of Haringey* ([2025] EWHC 1040 (QB), the court made a wasted costs order against lawyers who had misled the court on the basis of an AI hallucination and reiterated that doing so was a breach of advocates' professional duties not only to the court but also to their own clients. It is not yet clear why LLMs are prone to hallucinating, and some have suggested that total eradication of it may be impossible and that this represents an inherent part of the way these models work (Sun et al., 2024). The potential for injustice to be done should states rely too heavily on these closed black-box systems, however, particularly if they are developed in the cultural and legal context of another country and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

culture, is yet another reason why states in the Global South ought to be cautious about utilising AI models developed in another country or culture. Of course, it not the case that AI sovereignty alone is able to cure such a failing. If it is the case as Sun et al. suggest that hallucinations are an inherent part of the operation of LLMs or other AI models, then this will surely be replicated by the native development and training of these programmes. However, the effects of misinformation or fabrications imported from a foreign culture may be more damaging than those from a nation with its own sovereign AI, where mistakes can be identified more readily.

In addition to this is the cultural impact of using non-sovereign AI in a legal setting. AI models trained in the United Kingdom or United States are built in the cultural context of a common law system and legal language. These legal systems incorporate doctrines and concepts such as equity into them, recognise a range of common rights such as natural justice for example, and are built on a model of *stare decisis*, or binding precedence (Lewis, 2021, p. 873). These systems simply do not translate to other legal systems such as those built on civil law, but an AI system which hallucinates answers is likely to not only provide the wrong answers but also bring into a legal system influences and understandings of the law which do not belong within it. For a lawyer using such a system as part of their case preparation in an Asian country for example, Western-centric LLMs may influence their understanding of the law, and in turn, this can, over time, influence the interpretation of the law. These situations again highlight the risks faced by countries in the Global South resulting from the use of non-sovereign AI and, more than anything, the risk of the perpetuation of Western-centric dominance over the Global South.

The risks posed by non-sovereign AI use in digital surveillance and censorship and their risks for the protection of human rights in the developing world will now be considered further.

4. AI in digital surveillance, censorship, and human rights

4.1. The tension between sovereignty and human rights

Whilst AI systems may certainly have the power to improve governance for states and governments, this same technology may also be misused, to assist a state in the control and oppression of its own people and through the infringement of their human rights. It is now accepted under various instruments of public international law that human beings have a range of inalienable rights by virtue of their personhood, such as found under the Universal Declaration on Human Rights (UDHR 1948) although this is not a binding instrument of international law. Many states around the world have subsequently adopted binding multilateral conventions protecting human rights such as in the International Covenant on Civil and Political Rights (ICCPR) 1966 and European Convention on Human Rights 1950 (ECHR). There is, therefore, an imperative on states to ensure that the development, training, and use of AI take place in a manner which is consistent with human rights.

Unfortunately, as will now be seen, there is a risk that not all states will seek to use AI to uphold human rights. This paper will now set out some of the key concerns which arise from the potential use of AI as a tool of state surveillance and oppression, and why this risk is heightened by a world in which every state pursues its own version of AI sovereignty. This paper will suggest that it is only by engaging in multilateralism and cooperative development of AI and through the pooling of sovereignty as seen in the EU's approach to AI, that human rights can be properly protected.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

4.2. Why AI poses risks to human rights

AI systems have the capacity to analyse and synthesise huge datasets in a relatively short period of time. This makes some AI models useful for the purposes of processing information about individuals and making predictions as to their likely behaviour in the future. In China, the widespread use of facial recognition technology as part of the state's routine and mass surveillance efforts over its own people now uses predictive AI to identify risks from the perspective of crime, disorder, or other undesirable activities before they necessarily occur (Brussee, 2022). An AI model might find the presence of a given individual in a particular place suspicious, out of place, or inconsistent with other information known about the individual, leading to a police intervention or human surveillance team being instructed to follow the individual, or investigation begun for example (Richardson, Schultz and Crawford, 2019). Whilst the Chinese authorities suggest that the model is an exemplar of efficient policing and helps ensure state security, the reality is that this type of ongoing blanket surveillance of individuals and the screening of individuals without cause or reasonable objective grounds for suspicion of wrongdoing constitutes a severe infringement of human rights as they have traditionally been understood. Under the ICCPR for example, every person is entitled to the right to privacy, family life, home, and correspondence (Article 17(1) International Covenant on Civil and Political Rights (ICCPR), 1966). Whilst a surveillance state such as China might suggest that the state has a legitimate interest in using closed-circuit television cameras (CCTV) or other forms of camera technology to ensure law and order in real time, or to assist law enforcement, international courts and tribunals have repeatedly reiterated that targeted surveillance of any people is required to be pursued only on the basis of proportionality (*S and Marper v United Kingdom* [2008] Applications 30562/04 and 30566/04). Engaging in blanket surveillance of individuals by screening or by effectively spying on them, their correspondence, or their associations as they go about their daily business, is likely to contravene this right to privacy, as well as the right to freedom of assembly and the right to a presumption of innocence typically found as part of the right to a fair trial (Galetta, 2013). AI tools which facilitate this are being used to help infringe fundamental human rights, therefore.

4.3. AI, surveillance, and social credit

The risk to human rights posed by AI is even more likely to interfere with human rights when the state goes beyond using it as merely part of a surveillance system, but uses it to help justify arbitrary detention or punishment without trial, and without any criminal offence being committed. It is a fundamental human right to be entitled to freedom from arbitrary detention, punishment, confiscation, or forfeiture by the state without due process of law (Ferstman, 2024, p. 21). In the Western world, such a principle can be seen at least as far back as Magna Carta (2024) in England in 1215. This principle is reiterated in the Universal Declaration of Human Rights (UDHR) at Article 11(2), and in the ICCPR for example at Article 9. The rule of law typically requires that all are considered equal under the law, and that no one can be denied rights or privileges offered to others, or have their own rights curtailed arbitrarily without due process of law (in effect, by being found guilty of a criminal, penal offence). Despite this, some states such as China have utilised their surveillance technology in order to create a "social credit" score for each of their citizens (Brussee, 2022). While often popularised in Western discourse as a monolithic social credit score, the reality of China's Social Credit System is more accurately described as a fragmented regulatory framework focusing heavily on corporate compliance and judicial enforcement. Rather than there being a single magic algorithm assigning a numerical rank to every citizen, the system primarily operates through interlocking blacklists, such as the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

Supreme People's Court list for "untrustworthy" debtors, which can restrict access to high-speed rail, luxury hotels, or civil service positions (Brussee, 2022). A low social credit score, in turn, might result the individual being denied access to benefits or privileges allowed for others, effectively upending the principle of equality before the law which the rule of law requires, and constituting a punishment for that individual without due process of law. Importantly, the state in question might be able to arbitrarily change what constitutes social credits or demerits as it wishes, moving the goalposts from time to time. Again, this is likely to constitute an interference of human rights and the rule of law, which requires that the law is promulgated properly, clearly, and is understandable and predictable (Article 9(1) ICCPR 1966; Bingham, 2007, p. 67). AI may be used to play a central role in the establishment, monitoring, and enforcement of such a social credit system. Indeed, given AI's capacity for analysing vast amounts of data, it is perfectly suited to such a role (Agrawal, 2022, p. 13124). AI might be able to review CCTV footage, for example, and identify retrospectively incidents which require a social credit point to be demerited against a person autonomously. It might be able to review social media in real time, identify those who perform some act or show a pattern of behaviour deemed undesirable, and so deny social credit on that basis. AI might, in short, be used in both the surveillance and punishment of people on their behaviour even when it is not ever proved before a court of law at a fair trial that the individual has committed any criminal offence.

4.4. Algorithmic adjudication and the right to a fair trial

Even where there is a trial of a person, the use of AI in evidence or to act as a so-called "robot judge," which might be sought by the state in order to create efficiency in the justice system, might be regarded as a threat to human rights. AI is vulnerable to biases and may even harbour covert racism which leads it to prejudice against certain individuals due to its training biases being activated (Hofmann et al., 2024, p. 147). It may even "hallucinate" facts which the prosecution or defence have not adduced in evidence in its determination of guilt or in sentencing against the defendant (Sun et al., 2024). Because the system is a virtual "black box" however, it is impossible to tell if this has even occurred. This is yet another way AI poses a future risk to the protection of human rights.

Potentially even more concerning than this is the suggestion that AI might be used to predict criminal events taking place. The UK Government has launched an AI initiative for example intended to try and anticipate where and when certain criminal offences might be likely to take place (HM Government, 2025). As AI technology progresses and more data over individuals garnered from social credit systems, digital ID initiatives (Open Rights Group, 2025), there is a risk that this information might be used to train AI to identify criminals based on a calculated probability of them committing certain types of offences. A person's own cognitive patterns might be mapped from their social media and known behaviours, spending, socio-economic status, and so on, and this used to model a likelihood of their potential offending or status as a "good citizen" in line with the idea of "social credit" (Schuilenburg, 2021, p. 47). This "precrime," if it subjected the individual to any sort of punishment outside of the law—such as arbitrary detention, surveillance, monitoring, or house arrest—would very likely pose a threat to human rights too. The UK's courts have already made clear that imposing orders on individuals who are only suspected of being likely to be involved of certain offences (such as terrorism offences) is incompatible with the right to private and family life, right to liberty and security of the person, and the presumption of innocence and right to fair trial under Articles 8, 5, and 6 of the ECHR (*JJ v Secretary of State for the Home Department* [2007] UKHL 45; *A v Secretary of State for the Home Department* [2004] UKHL 56). The imposition of a similar scheme allowing detention or punishment for "pre-crime" would similarly be likely to constitute a breach

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

of fundamental rights elsewhere should a State seek to develop an AI model capable of operating in this way.

4.5. The implications for AI sovereignty

This chapter has identified a wide range of ways in which AI has the potential to interfere with individual's human rights almost anywhere in the world. Whilst AI may improve people's lives through increasing economic development or by assisting them in their daily lives and work, these risks to human rights are real. There is genuine scope for a totalitarian state to seek to use AI to control, oppress, or punish their people, and to discourage them from resisting via a number of tools. The imposition of social credit systems, the risk of "robot" judges covertly prejudicing minorities, or the use of AI in imposing punishment or restrictions on people for "pre crime" all suggest a dystopian future path which states might take with their R&D of AI.

This has implications for the debate on AI sovereignty. This paper has sought to answer whether it is necessary and desirable for states, particularly those in the developing world and in Asia to seek to obtain some degree of AI sovereignty, either through ensuring absolute autonomy over AI and its supply chain as China has sought to do, or through distributed, pooled sovereignty. In the main, it has been found that the risks of bias and the risk of cultural hegemony of the West are good reasons to consider some means of security AI sovereignty for these developing states.

However, the risk that states might decide to design, build, and implement their own autonomous AI models which might contravene human rights law or international law obligations forces this desire to obtain AI sovereignty to be curtailed. Whilst sovereignty is desirable, there is a caveat. AI sovereignty is no longer as desirable if the state engages on the development of AI systems which are designed to assist in the interference of human rights in their own states, or to facilitate the oppression or abuse of people contrary to law. In such circumstances, collaboration with other states, under agreements which require the parties to respect human rights, is much preferable. The EU's rules here in the EU AI Act (Regulation (EU) 2024/1689) legislation, for example, are a good example of such an approach; the Regulation prohibits the use of AI systems which are regarded as being dangerous for human rights for a number of reasons set out in Article 5(1) or (2) for example (Article 5(1)(2) Regulation (EU) 2024/1689). States in Asia who seek some degree of AI sovereignty could learn from this approach and seek to ensure that the use of AI is carried out in their own country, and in others, with respect for human rights while continuing to pool their resources and sovereignty with other like-minded states.

Whilst it is true that a multilateral legal framework does not automatically guarantee compliance, the pooling of sovereignty does provide a structural and technical capacity for human rights protection that individual states often lack when operating alone. By collaboratively governing the datasets, training parameters, and infrastructure, participating states can ensure that AI models are built on non-biased foundations from the outset, rather than attempting to retrofit human rights safeguards onto opaque, "black box" systems developed elsewhere. This cooperative model functions as a form of "digital federalism," where harmonised minimum standards create a collective barrier against the technological dependency that otherwise leaves states vulnerable to the ethical biases embedded in foreign AI.

Of course, this model still leaves the world vulnerable to the risk that states might decide that the best use of collaboration would be to help enhance their own surveillance of their own citizens. The question is what can be done to ensure that the future proliferation of these AI models is one which helps to respect freedoms and human rights. Here, the reality is that the importance of AI sovereignty is a high-stakes race (Mügge, 2024, p. 2200). Those models which are developed and deployed first are likely to form the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

basis of future models in turn, with supply chains, training, and economies of scale all meaning that those states first succeeding here will be most likely to dominate in the future. States which develop an integrated regional understanding on the importance of ensuring protection of human rights in their training of AI, their selection of datasets, or through the use of human oversight (through a human-in-the-loop model, at least until such time as the AI comes to oversee itself more autonomously in a manner which can be trusted to respect rights) ought to see the importance of AI sovereignty as being a key research priority. Should states which collaborate with one another to build AI models which are designed expressly to infringe human rights—through the building of social credit monitoring tools, for example—win this race, then the reality will be that there is little that can be done to prevent the use of AI for these ends. This is why ensuring investment in AI technology for the purposes of AI data-sovereignty is now so important.

5. Recommendations and conclusion

This paper has examined a number of facets of AI in order to identify the importance for AI sovereignty, and the risks which absolute sovereignty over AI might entail. It has been found that the Western dominance of early AI programmes and the risks of algorithmic bias being integrated into machine learning systems from cognitive, implicit, or heuristic biases of human trainers and non-diverse English language dataset dominance creates a strong case for AI sovereignty. This is particularly so for developing states in the Global South given the threat to cultural heritages and the risk of the perpetuation of Western hegemonic ideals upon these countries. The threat that absolute AI sovereignty creates for the protection of human rights, and the potential abuse and misuse of AI by totalitarian tyrannical states, however, suggests that some form of external control over a state's development and use of AI might be preferable.

In practice, as has been seen in this paper, the reality is that whilst a state might declare its own autonomy and sovereignty over AI, it is very difficult, in practice, for absolute autonomy to be achieved; a State would require closed datasets for training, control of microchips, rare earth metals, data centres and supply chains throughout to do so, and whilst some powerful states such as China might be able to do this, the vast majority of states will simply not be able to.

In order to obtain sovereignty, therefore, they must seek to share that sovereignty with others. Whilst the EU has a model of achieving this sort of pooled, decentralised sovereignty through the various instruments and bodies of EU law, and through its AI Act, this organisation is something of an outlier in international law. No other regional trade bloc has such a deep degree of integration allowing its own central courts to enforce the law and to ensure harmonious application of it domestically for example. As such, a slightly different model is required. For states in the Asian region, AI sovereignty which helps resist Western dominance might best be achieved through collaboration, through a multilateral treaty-based organisation where resources can be pooled, and AI developed in a mutually beneficial manner. The protection of human rights might be protected by treaty in this approach, with states agreeing not to allow research or use of dangerous forms of AI. Treaty protections could include sanctions, such as the ejection of a state from the research or imposition of fines. This model of cooperation would enable both sovereignty and a protection of human rights, with the practical benefits which cooperation bring such as through the pooling of resources and scientific knowledge, and this paper concludes by recommending the development of closer ties between nations in the Global South with an intent to create such agreement.

Acknowledgements. An earlier version of this paper was presented at the Workshop on AI Sovereignty and International Law, organised by the Centre for International Law, National University of Singapore, held on

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

Friday, 12 December 2025, at the Orchard Rendezvous Hotel, Singapore. The presentation was delivered by Noble Lo, with Ada Yuen participating in the workshop presentation online.

The presentation was based on the manuscript submitted to the *Asian Journal of Law and Society* special issue entitled "AI Sovereignty and International Law: Asian Regional Challenges and Global Implications." The author is grateful to the NUS Centre for International Law for its full institutional and financial support, including travel funding, accommodation, and a per diem allowance. The author also sincerely thanks the workshop participants for their valuable comments and constructive feedback, which informed subsequent revisions. Any remaining errors are the author's sole responsibility. Finally, the author gratefully acknowledges Lancaster University for its financial support for this publication.

Cases. Av Secretary of State for the Home Department [2004] UKHL 56.
Costa v ENEL Case C-6/64 [1964] ECR 585.
R (Ayinde) v London Borough of Haringey [2025] EWHC 1040 (QB).
JJ v Secretary of State for the Home Department [2007] UKHL 45.
S and Marper v United Kingdom [2008] Applications 30562/04 and 30566/04 Strasbourg 4 December 2008.

Legislation. European Convention on Human Rights (ECHR) 1950.
International Covenant on Civil and Political Rights (ICCPR) 1966.
Magna Carta 1215.
Regulation (EU) 2024/1689.
Universal Declaration of Human Rights (UDHR) 1948.

References

Agrawal, V. (2022). 'Demystifying the Chinese social credit system: A case study on AI-powered control systems in China', Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), pp. 13124–13125. Available at: https://doi.org/10.1609/aaai.v36i11.21698 (Accessed: 3 October 2025).

Anghie, A. (2023). 'Rethinking international law: A TWAIL retrospective', *European Journal of International Law*, 34(1), pp. 7–112.

Appleton, B. (2026). 'Algorithmic empire and the new digital colonialism: The legal struggle for technological self-determination in the age of AI'. Available at: https://doi.org/10.2139/ssrn.5389292 (Accessed: 13 April 2026).

Arroyo, P., Schöttle, A. and Christensen, R. (2022). 'A Shared Responsibility: Ethical and social dilemmas of using AI in the AEC Industry', in González, V.A., Hamzeh, F. and Alarcón, L.F. (eds) *Lean Construction 4.0: Driving a digital revolution of production management in the AEC industry*. London: Routledge, pp. 68–81.

Berthet, V. (2022). 'The impact of cognitive biases on professionals' decision-making: A review of four occupational areas', *Frontiers in Psychology*, 12, article no. 802439.

Bingham, T. (2007). 'The rule of law', *Cambridge Law Journal*, 66(1), pp. 67–85.

Bohannon, M. (2023). 'Lawyer used chatgpt in court—and cited fake cases. A judge is considering sanctions', Forbes, 8 June. Available at: https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/ (Accessed: 3 October 2025).

Brussee, V. (2022). 'China's social credit score—untangling myth from reality'. Available at: https://merics.org/en/comment/chinas-social-credit-score-untangling-myth-reality (Accessed: 3 October 2025).

Buolamwini, J. and Gebru, T. (2018). 'Gender shades: Intersectional accuracy disparities in commercial gender classification', *Proceedings of Machine Learning Research*, 81, pp. 1–15.

Cavus, M., Benli, M. N., Altuntas, U., Sari, M., Ayan, H. and Ugurluoglu, Y. F. (2025). 'Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice', *Applied Soft Computing*, 176, article no. 113160.

Chang, W., Arcesati, R. and Hmaidi, A. (2025). China's drive toward self-reliance in artificial intelligence: From chips to large language models. Available at: https://merics.org/en/report/chinas-drive-toward-self-reliance-artificial-intelligence-chips-large-language-models (Accessed: 3 October 2025).

Chirathivat, S., Kunnamas, N. and Welfens, P. J. J. (2020). 'Regional integration in the EU and ASEAN in the period of declining multilateralism and corona shocks', *International Economics and Economic Policy*, 17, pp. 555–561.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N. and Owen, D. (2024). How much does it cost to train frontier AI models? Available at: https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models (Accessed: 4 January 2026).

Embassy of the People's Republic of China in Grenada (2023). Global AI Governance Initiative. Available at: https://gd.china-embassy.gov.cn/eng/zxhd_1/202310/t20231024_11167412.htm (Accessed: 3 October 2025).

Ferstman, C. (2024). *Conceptualising arbitrary detention: Power, punishment and control*. Bristol: Bristol University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

Franklin, G., Stephens, R., Piracha, M., Tiosano, S., Lehouillier, F., Koppel, R. and Elkin, P. L. (2024). 'The sociodemographic biases in machine learning algorithms: A biomedical informatics perspective', *Life*, 14(6), article no. 652.

Galetta, A. (2013). 'The changing nature of the presumption of innocence in today's surveillance societies: Rewrite human rights or regulate the use of surveillance technologies?' *European Journal of Law and Technology*, 4(2).https://ejlt.org/index.php/ejlt/article/view/221/377 and https://ejlt.org/index.php/ejlt/article/view/221.

Garcia, A. R. (2019). 'AI, IoT, big data, and technologies in digital economy with blockchain at sustainable work satisfaction to smart mankind: Access to 6th dimension of human rights', in Lopes, N. V. M. (ed) *Smart governance for cities: Perspectives and experiences*. Cham: Springer, pp. 83–131.

Gignac, G. E. and Szodorai, E. T. (2024). 'Defining intelligence: Bridging the gap between human and artificial perspectives', *Intelligence*, 104, article no. 101832.

Gravett, W. (2020). 'Digital neo-colonialism: The Chinese model of internet sovereignty in Africa', *African Human Rights Law Journal*, 20(1), pp. 125–146.

Heersmink, R., de Rooij, B., Clavel Vázquez, M. J. and Colombo, M. (2024). 'A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness', *Ethics and Information Technology*, 26, article no. 41.

HM Government (2025). AI to help police catch criminals before they strike. Available at: https://www.gov.uk/government/news/ai-to-help-police-catch-criminals-before-they-strike (Accessed: 3 October 2025).

Hofmann, V., Kalluri, P. R., Jurafsky, D. and King, S. (2024). 'AI generates covertly racist decisions about people based on their dialect', *Nature*, 633, pp. 147–154.

Ishkhanyan, A. (2025). 'The sovereignty-internationalism paradox in AI governance: Digital federalism and global algorithmic control', *Discover Artificial Intelligence*, 5, article no. 123.

Lewis, S. (2021). 'Precedent and the rule of law', *Oxford Journal of Legal Studies*, 41(4), pp. 873–898.

Moser, D., Steiglechner, P. and Schlueter, A. (2022). 'Facing global environmental change: The role of culturally embedded cognitive biases', *Environmental Development*, 44, article no. 100735.

Mügge, D. (2024). 'EU AI sovereignty: For whom, to what end, and to whose benefit?' *Journal of European Public Policy*, 31(8), pp. 2200–2225.

Navigli, R., Conia, S. and Ross, B. (2023) 'Biases in large language models: Origins, inventory, and discussion', *ACM Journal of Data and Information Quality*, 15(2), article no. 10.

O'Donoghue, K. (2025). 'A patchwork of state AI regulation is bad. A moratorium is worse', AI Frontiers, 25 June. Available at: https://ai-frontiers.org/articles/congress-might-block-states-from-regulating-ai (Accessed: 3 October 2025).

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019). 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, 366 (6464), pp. 447–453.

Open Rights Group (2025). 'ID cards: UK risks sleepwalking into a pre-crime state', Open Rights Group, 26 September. Available at: https://www.openrightsgroup.org/press-releases/id-cards-uk-risks-sleeping-walking-into-pre-crime-state/ (Accessed: 3 October 2025).

Papadouli, V. (2023). 'Artificial intelligence's black box: Posing new ethical and legal challenges on modern societies', in Kornilakis, A., Nouskalis, G., Pergantis, V. and Tzimas, T. (eds) *Artificial Intelligence and normative challenges*. Cham: Springer, pp. 39–62.

Perisic, J.C. (2024). 'Digital transformation: Regulation of artificial intelligence in the European Union', in Kellerhals, A., Baumgartner, T. and Ademi, F. (eds) *European integration in times of security challenges*. Zürich: EIZ Publishing, pp. 147–164.

Richardson, R., Schultz, J. M. and Crawford, K. (2019). 'Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice', *New York University Law Review Online*, 94, pp. 192–228.

Roberts, H. (2024). 'Digital sovereignty and artificial intelligence: A normative approach', *Ethics and Information Technology*, 26, article no. 70.

de Sousa Santos, B. (2024). 'AI and the epistemologies of the south', *Journal of World-Systems Research*, 30(2), pp. 635–645.

Schuilenburg, M. (2021). 'The security society: On power, surveillance, and punishments', in Arrigo, B. and Sellers, B. (eds) *The pre-crime society: Crime, culture and control in the Ultramodern Age*. Bristol: Bristol University Press, pp. 43–62.

Shah, H. S. and Bohlen, J. (2023). *Implicit bias*. Treasure Island (FL): StatPearls Publishing.

Simmler, M. and Canova, G. (2025). 'Facial recognition technology in law enforcement: Regulating data analysis of another kind', *Computer Law & Security Review*, 56, article no. 106092.

Srivastava, S. (2022). *Hybrid sovereignty in world politics*. 1st edn. Cambridge: Cambridge University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053

Sun, Y., Sheng, D., Zhou, Z. and Wu, Y. (2024). 'AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content', *Humanities and Social Sciences Communications*, 11, article no. 1278.

Tao, Y., Viberg, O., Baker, R. S. and Kizilcec, R. F. (2024). 'Cultural bias and cultural alignment of large language models', *PNAS Nexus*, 3(9), pp. 346–355.

The White House (2025). Winning the race: America's AI action plan. Available at: https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf (Accessed: 3 October 2025).

Tony Blair Institute for Global Change (2025). Sovereignty, security, scale: A UK strategy for AI infrastructure. Available at: https://institute.global/insights/tech-and-digitalisation/sovereignty-security-scale-a-uk-strategy-for-ai-infrastructure (Accessed: 3 October 2025).

Vohra, D. K. (2024). 'Understanding AI sovereignty: Importance for nations and enterprises', NexGen Cloud Blog, 1 October. Available at: https://www.nexgencloud.com/blog/thought-leadership/what-is-ai-sovereignty-why-it-matters-for-national-and-enterprise-ai-strategy (Accessed: 3 October 2025).

Webster, C. S. (2025). 'Natural and artificial intelligence—The psychotechnical agenda of the 21st century', *Journal of Psychology and AI*, 1(1), article no. 2491445.

White, T. (2025). The AI regulatory maze: Why tech vendors are being set up to fail. Omdia, 27 November. Available at: https://omdia.tech.informa.com (Accessed: 4 January 2026).

Woods, A.K. (2023). 'Digital sovereignty + artificial intelligence', in Chander, A. and Sun, H. (eds) *Data sovereignty: From the digital silk road to the return of the state*. Oxford: Oxford University Press, pp. 115–136.

Xu, W., Agrawal, S., Zouhar, V., Freitag, M. and Deutsch, D. (2025). Deconstructing self-bias in LLM-generated translation benchmarks. Available at: https://arxiv.org/abs/2509.26600 (Accessed: 2 May 2026).

Noble Po-kan Lo completed his master's degrees at the University of Hong Kong and the University of Oxford. He received his Juris Doctor from the Chinese University of Hong Kong. He has actively pursued specialised training that bridges law and technology, completing a Certificate in Generative AI for the Legal Profession at UC Berkeley and a Certificate in Computer Science for Lawyers at Harvard Law School. Professionally, he completed mini-pupillages at Des Voeux Chambers and Liberty Chambers, as well as internships at Clifford Chance and Linklaters. These experiences have provided him with practical advocacy training and insight into the transformative role of technology in global legal practice. He is currently teaching at the College of Professional and Continuing Education, Hong Kong Polytechnic University.

Cite this article: Lo NP (2026). AI Sovereignty, Data Bias, and Ethical Challenges in International Law. *Asian Journal of Law and Society*. https://doi.org/10.1017/als.2026.10053

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 18 Nov 2026 at 07:27:43, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/als.2026.10053