

# Should Artificial Intelligence Governance be Centralised? Design Lessons from History

Peter Cihon\*

Centre for the Governance of AI,  
Future of Humanity Institute,  
University of Oxford  
petercihon@gmail.com

Matthijs M. Maas\*

CILCC, Faculty of Law,  
University of Copenhagen &  
Centre for the Governance of AI,  
Future of Humanity Institute,  
University of Oxford  
matthijs.maas@jur.ku.dk

Luke Kemp\*

Centre for the Study of Existential  
Risk, University of Cambridge  
ltk27@cam.ac.uk

## ABSTRACT

Can effective international governance for artificial intelligence remain fragmented, or is there a need for a centralised international organisation for AI? We draw on the history of other international regimes to identify advantages and disadvantages in centralising AI governance. Some considerations, such as efficiency and political power, speak in favour of centralisation. Conversely, the risk of creating a slow and brittle institution speaks against it, as does the difficulty in securing participation while creating stringent rules. Other considerations depend on the specific design of a centralised institution. A well-designed body may be able to deter forum shopping and ensure policy coordination. However, forum shopping can be beneficial and a fragmented landscape of institutions can be self-organising. Centralisation entails trade-offs and the details matter. We conclude with two core recommendations. First, the outcome will depend on the exact design of a central institution. A well-designed centralised regime covering a set of coherent issues could be beneficial. But locking-in an inadequate structure may pose a fate worse than fragmentation. Second, for now fragmentation will likely persist. This should be closely monitored to see if it is self-organising or simply inadequate.

## KEYWORDS

artificial intelligence policy, global governance, institutional design, governance fragmentation, centralization, regime complex

### ACM Reference Format:

Peter Cihon, Matthijs M. Maas, and Luke Kemp. 2020. Should Artificial Intelligence Governance be Centralised? Design Lessons from History. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, February 7–8, 2020, New York, NY, USA. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3375627.3375857>

\*Equal contribution, order selected at random.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

AIES '20, February 7–8, 2020, New York, NY, USA

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7110-0/20/02...\$15.00

<https://doi.org/10.1145/3375627.3375857>

## 1 INTRODUCTION

In 2018, Canada and France proposed the International Panel on Artificial Intelligence (IPAI). After being rejected at the G7 in 2019, negotiations shifted to the OECD and are presently ongoing. As the field of AI continues to mature and spark public interest and legislative concern [41], the priority of such governance initiatives reflects the growing appreciation that AI has the potential to dramatically change the world for both good and ill [9]. Research into AI governance needs to keep pace with policy-making and technological change. Choices made today may have long-lasting impacts on policymakers' ability to address numerous AI policy problems [7]. Effective governance can promote safety, accountability, and responsible behaviour in the research, development, and deployment of AI systems.

AI governance research to date has predominantly focused at the national and sub-national levels [6, 16, 44]. Research into AI *global* governance remains relatively nascent (though see [5]). Kemp et al. [24] have called for specialised, centralised intergovernmental agencies to coordinate policy responses globally, and others have called for a centralised 'International Artificial Intelligence Organisation' [14]. Others favour more decentralised arrangements based around 'Governance Coordinating Committees', global standards, or existing international law instruments [8, 28, 47].

No one has taken a step back to inquire: what would the history of multilateralism suggest, given the state and trajectory of AI? Should AI governance be centralised or decentralised? 'Centralisation', in this case, refers to the degree to which the coordination, oversight and/or regulation of a set of AI policy issues or technologies are housed under a single (global) institution. This is not a binary choice; it exists across a spectrum. Trade is highly (but not entirely) centralised under the umbrella of the WTO. In contrast, environmental multilateralism is much more decentralised.

In this paper, we seek to help the community of researchers, policymakers, and other stakeholders in AI governance understand the advantages and disadvantages of centralisation. This may help set terms and catalyse a much-needed debate to inform governance design decisions. We first outline the international governance challenges of AI, and review early proposed global responses. We then draw on existing literatures on regime fragmentation [3] and 'regime complexes' [36] to assess considerations in centralising the international governance of AI. We draw on the history of

other international regimes<sup>1</sup> to identify considerations that speak in favour or against designing a centralised regime complex for AI. We conclude with two recommendations. First, many trade-offs are contingent on how well-designed a central body would be. An adaptable, powerful institution with a manageable mandate would be beneficial, but a poorly designed body could prove a fate worse than fragmentation. Second, for now there should be structured monitoring of existing efforts to see whether they are self-organising or insufficient.

## 2 THE STATE OF AI GOVERNANCE

There is debate as to whether AI is a single policy area or a diverse series of issues. Some claim that AI cannot be cohesively regulated as it is a collection of disparate technologies, with different risk profiles across different applications and industries [45]. This is an important but not entirely convincing objection. The technical field has no settled definition for ‘AI’,<sup>2</sup> so it should be no surprise that defining a manageable scope for AI governance will be difficult. Yet this challenge is not unique to AI: definitional issues abound in areas such as environment and energy, but have not figured prominently in debates over centralisation. Indeed, energy and environment ministries are common at the domestic level, despite problems in setting the boundaries of natural systems and resources.

We contend that there are numerous ways in which a centralised body could be designed for AI governance. For example, a centralised approach could carve out a subset of interlinked AI issues to cover. This could involve focusing on the potentially high-risk applications of AI systems, such as AI-enabled cyberwarfare, lethal autonomous weapons (LAWS), other advanced military applications, or high-level machine intelligence (HLMI).<sup>3</sup> Another approach could govern underlying hardware resources (e.g. large-scale compute resources) or software libraries. We are agnostic on the specifics of how centralisation could or should be implemented, and instead focus on the costs and benefits of centralisation in the abstract. The exact advantages and disadvantages of centralisation are likely to vary depending on the institutional design. This is an important area of further study, particularly once more specific proposals are put forward. However, such work must be grounded in a higher-level investigation of trade-offs in centralising AI governance. It is this foundational analysis which we seek to offer.

Numerous AI issues could benefit from international cooperation. These include the potentially catastrophic applications mentioned above. It also encompasses more quotidian uses, such as AI-enabled cybercrime; human health applications; safety and regulation of autonomous vehicles and drones; surveillance, privacy and data-use; and labour automation. Multilateral coordination could also use AI to tackle other global problems such as climate change [43], or help meet the Sustainable Development Goals [46]. This is an illustrative but not exhaustive list of international AI policy issues.

Global regulation across these issues is currently nascent, fragmented, yet evolving. A wide range of UN institutions have begun to undertake some activities on AI [20]. The bodies covering AI policy issues range across existing organisations including the International Labour Organisation (ILO), International Telecommunication Union (ITU), and UNESCO. This is complemented by budding regulations and working groups across the International Organisation for Standardisation (ISO), International Maritime Organisation (IMO), International Civil Aviation Organisation (ICAO), and other bodies, as well as treaty amendments, such as the updating of the Vienna Convention on Road Traffic to encompass autonomous vehicles [28], or the ongoing negotiations at the Convention on Certain Conventional Weapons (CCW) on LAWS. The UN System Chief Executives Board (CEB) for Coordination through the High-Level Committee on Programmes has been empowered to draft a system-wide AI capacity building strategy. The High-level Panel on Digital Cooperation has also sought to gather together common principles and ideas for AI relevant areas [19]. Whether these initiatives bear fruit, however, remains questionable, as many of the involved international organisations have fragmented membership, were not originally created to address AI issues and lack effective enforcement or compliance mechanisms [32, p.2].

The trajectory of these initiatives matters. How governance is initially organised can be central to its success. Debates over centralisation and fragmentation are long-lasting and prominent with good reason. How we structure international cooperation can be critical to its success, and most other debates often implicitly hinge on structural debates. Fragmentation and centralisation exist across a spectrum. In a world lacking a global government, some fragmentation will always prevail. But the degree to which it prevails is crucial. We define ‘fragmentation’ as a patchwork of international organisations and institutions which focus on a particular issue area, but differ in scope, membership and often rules [3, p.16]. We define centralisation as an arrangement in which governance of a particular issue lies under the authority of a single umbrella body. A regime complex is a network of three or more international regimes on a common issue area. These should have overlapping membership and cause potentially problematic interactions [36, p.29]. These definitions and terms are by nature normatively loaded. For example, some may find ‘decentralisation’ to be a positive framing, while others may see ‘fragmentation’ to possess negative connotations. Recognising this, we seek to use these terms in a primarily analytical manner. We will use findings from each of these theoretical areas to inform our discussion of the history of multilateral fragmentation and its implications for AI governance.

## 3 CENTRALISATION CRITERIA: HISTORY OF GOVERNANCE TRADE-OFFS

In the following discussion, we explore a series of considerations for AI governance. Political power and efficient participation support centralisation. The breadth vs. depth dilemma, as well as slowness and brittleness support decentralisation. Policy coordination and forum shopping considerations can cut both ways.

<sup>1</sup>A regime is a set of ‘implicit or explicit principles, norms, rules and decision-making procedures around which actors’ expectations converge in a given area of international relations’ [27, p.186].

<sup>2</sup>We define ‘AI’ as any machine system capable of functioning ‘appropriately and with foresight in its environment’ [34, p.13]; see too [9, p.5].

<sup>3</sup>‘High-level machine intelligence’ has been defined as ‘unaided machines [that] can accomplish every task better and more cheaply than human workers’ [17, p.1].

### 3.1 Political Power

Regimes embody power in their authority over rules, norms, and knowledge beyond states' exclusive control. A more centralised regime will see this power concentrated among fewer institutions. A centralised, powerful architecture is likely to be more influential against competing international organisations and with constituent states [36, pp.36-7].

An absence of centralised authority to manage regime complexes has presented challenges in the past. Across the proliferation of Multilateral Environmental Agreements (MEAs) there is no requirement to cede responsibility to the UN Environmental Programme in the case of overlap or competition. This has led to turf wars, inefficiencies and even contradictory policies [3]. One of the most notable examples is that of hydrofluorocarbons (HFCs). HFCs are potent greenhouse gases, and yet their use has been encouraged by the Montreal Protocol since 1987 as a replacement for ozone-depleting substances. This has only recently been resolved via the 2015 Kigali Amendment to the Montreal Protocol, which itself has a prolonged implementation period. Similarly, the internet governance regime complex is diffuse. Multiple venues and norms govern technical standards, cyber crime, human rights, and warfare [35]. Although the UN Internet Governance Forum (IGF) discusses several cross-cutting issues, it does not have a mandate to consolidate even principles, let alone negotiate new formal agreements [33].

In contrast, other centralised regimes have supported effective management. For example, under the umbrella of the WTO, norms such as the most-favoured-nation principle (equally treating all WTO member states) principle have become the bedrock of international trade. The power and track-record of the WTO is so formidable that it has created a *chilling* effect: the fear of colliding with WTO norms and rules has led environmental treaties to self-censor and actively avoid discussing or deploying trade-related measures [12]. Both the chilling effect and the remarkably powerful application of common trade rules were not a marker of international trade until the establishment of the WTO. The power of these centralised body has stretched beyond influencing states in the domain of trade, to moulding related issues.

Political power offers further benefits in governing emerging technologies that are inherently uncertain in both substance and policy impact. Uncertainty in technology and preferences has been associated with some increased centralisation in regimes [25]. There may also be benefits to housing a foresight capacity within the regime complex, to allow for accelerated or even proactive efforts [39]. Centralised AI governance would enable an empowered organisation to more effectively use foresight analyses to inform policy responses across the regime complex.

### 3.2 Supporting Efficiency & Participation

Decentralised AI governance may undermine efficiency and inhibit participation. States often create centralised regimes to reduce costs, for instance by eliminating duplicate efforts, yielding economies of scale within secretariats, and simplifying participation [15]. Conversely, fragmented regimes may force states to spread resources and funding over many distinct institutions, particularly limiting the ability of less well-resourced states or parties to participate fully [32, p.2].

Historically, decentralised regimes have presented cost and related participation concerns. Hundreds of related and sometimes overlapping international environmental agreements can create 'treaty congestion' [1]. This complicates participation and implementation for both developed and developing nations [15]. This includes costs associated with travel to different forums, monitoring and reporting for a range of different bodies, and duplication of effort by different secretariats (*ibid.*).

Similar challenges are already being witnessed in AI governance. Simultaneous and globally distributed meetings pose burdensome participation costs for civil society. Fragmented organisations must duplicatively invest in high-demand machine learning subject matter experts to inform their activities. Centralisation would support institutional efficiency and participation.

### 3.3 Slowness & Brittleness of Centralised Regimes

One potential problem of centralisation lies in the relatively slow process of establishing centralised institutions, which may often be outpaced by the rate of technological change. Another challenge lies in centralised institutions' brittleness after they are established, i.e., their vulnerability to regulatory capture, or failure to react to changes in the problem landscape.

Establishing new international institutions is often a slow process. For example, the Kyoto Protocol took three years of negotiations to create and then another eight to enter into force. This becomes even more onerous with higher participation and stakes. Under the GATT, negotiations for a 26% cut in tariffs between 19 countries took 8 months in 1947. The Uruguay round, beginning in 1986, took 91 months to achieve a tariff reduction of 38% between 125 parties [31]. International law has been quick to respond to technological changes in some cases, and delayed in others [42, p.184]. Decentralised efforts may prove quicker to respond to complex, 'transversal' issues, if they rely more on informal institutions with a smaller but like-minded membership [32, pp.2-3]. Centralised AI governance may be particularly vulnerable to sparking lengthy negotiations, because progress on centralised regimes for new technologies tends to be hard if a few states hold clearly unequal stakes in the technology, or if there are significant differences in information and expertise among states or between states and private industry [42, pp.187-94]. Both these conditions closely match the context of AI technology. Moreover, because AI technology develops rapidly, such slow implementation of rules and principles could lead to certain actors taking advantage by setting *de facto* arrangements or extant state practice.

Even after its creation, a centralised regime can be *brittle*; the very qualities that provide it with political power may exacerbate the adverse effects of regulatory capture; the features that ensure institutional stability, may also mean that the institution cannot adapt quickly to unanticipated outside stressors outside its established mission. The regime might break before it bends. The first potential risk is regulatory capture. Given the high profile of AI issue areas, political independence is paramount. However, as illustrated by numerous cases, including undue corporate influence in the WHO during the 2009 H1N1 pandemic [11], no institution is fully immune to regime capture, and centralisation may reduce

the costs of lobbying, making capture easier by providing a single locus of influence. On the other hand, a regime complex comprising many parallel institutions could find itself vulnerable to capture by powerful actors, who are better positioned than smaller parties to send representatives to every forum.

Moreover, centralised regimes entail higher stakes. Many issues are in a single basket and thus failure is more likely to be severe if it does occur. International institutions can be notoriously path-dependent and thus fail to adjust to changing circumstances, as seen with the ILO's considerable difficulties in reforming its participation and rulemaking processes in the 1990s [2]. The public failure of a flagship global AI institution or governance effort could have lasting political repercussions. It could strangle subsequent, more well-conceived proposals in the crib, by undermining confidence in multilateral governance generally or capable governance on AI issues specifically. By contrast, for a decentralized regime complex to similarly fail, all of its component institutions would need to simultaneously 'break' or fail to innovate at once.<sup>4</sup> A centralised institution that does not outright collapse, but which remains ineffective, may become a blockade against better efforts.

Ultimately, brittleness is not an inherent weakness of centralisation—and indeed depends far more on institutional design details. There may be strategies to 'innovation-proof' [29] governance regimes. Periodic renegotiation, modular expansion, 'principles based regulation', or sunset clauses can also support ongoing reform [30, pp.29-30]. Such approaches have often proved successful historically, due partially to decentralisation but, importantly, also to particular designs.

### 3.4 The Breadth vs. Depth Dilemma

Pursuing centralisation may create an overly high threshold that limits participation. All multilateral agreements face a trade-off between having higher participation ('breadth') or stricter rules and greater ambition of commitments ('depth'). The dilemma is particularly evident for centralised institutions that are intended to be powerful and require strong commitments from states.

However, the opposite dynamics of sacrificing depth for breadth can also pose risks. The 2015 Paris Agreement on Climate Change was significantly watered down to allow for the legal participation of the US. Anticipated difficulties in ratification through the Senate led to negotiators opting for a 'pledge and review' structure with few legal obligations. Thus, the US could join simply through the approval of the executive [23]. In this case, inclusion of the US (which at any rate proved temporary) came at the cost of significant cutbacks on the demands which the regime sought to make of all parties.

In contrast, decentralisation could allow for major powers to engage in relevant regulatory efforts where they would be deterred from signing up to a more comprehensive package. This has precedence in the history of climate governance. Some claim that the US-led Asia-Pacific Partnership on Clean Development and Climate helped, rather than hindered climate governance, as it bypassed UNFCCC deadlock and secured non-binding commitments from actors not bound by the Kyoto Protocol [49, pp.259-60].

This matters, as buy-in may prove a thorny issue for AI governance. The actors who lead in AI development include powerful states that are potentially most adverse to global regulation in this area. They have thus far proved recalcitrant in the global governance of security issues such as anti-personnel mines or cyberwarfare. In response, some have already recommended a critical-mass governance approach to the military uses of AI. Rather than seeking a comprehensive agreement, devolving and spinning off certain components into separate treaties (e.g. for LAWS testing standards; liability and responsibility; and limits to operational usage) could instead allow for the powerful to ratify and move forward at least a few of those options [48].

The breadth vs. depth dilemma is a trade-off in multilateralism generally. However, it is a particularly pertinent challenge for centralisation. The key benefit of a centralised body would be to be a powerful anchor that ensures policy coordination and coherence, without suffering fragmentation in membership. This dilemma suggests it is unlikely to have both. It will likely need to restrict membership to have teeth, or lose its teeth to have wide participation. A critical mass approach may be able to deliver the best of both worlds. Nonetheless this dilemma poses a difficult knot for centralisation to unravel.

### 3.5 Forum Shopping

Forum shopping may help or hinder AI governance, depending on the particular circumstances. Fragmentation enables actors to choose where and how to engage. Such 'forum shopping' may take one of several forms: moving venues, abandoning one organisation, creating new venues, and working across multiple organisations to sew competition between them [4]. Even when there is a natural venue for an issue, actors have reasons to forum-shop. For instance, states may look to maximise their influence, appease domestic pressure [40] and placate constituents by shifting to a toothless forum [18].

The ability to successfully forum-shop depends on an actor's power. Most successful examples of forum-shifting have been led by the US [4]. Intellectual property rights in trade, for example, was subject to prolonged, contentious forum shopping. Developed states resisted attempts of the UN Conference on Trade and Development (UNCTAD) to address intellectual property rights in trade by trying to push them onto the World Intellectual Property Organization (WIPO) (*ibid.*, 566) and then subsequently to the WTO [18], overruling protests from developing states. Outcomes often reflect power, but weak states and non-state actors can also pursue forum shopping strategies in order to challenge the status-quo [22].

Forum shopping may help or hurt governance. This is evident in current efforts to regulate LAWS. While the Group of Governmental Experts has made some progress, on the whole the CCW has taken slow deliberations on LAWS. In response, frustrated activists have threatened to shift to another forum, as happened with the Ottawa Treaty that banned landmines [10]. This strategy could catalyse progress, but also brings risks of further forum shopping and weak or unimplemented agreements. Forum shopping may similarly delay, stall, or weaken regulation of time-sensitive AI policy issues, including potential future HLMI development. It is plausible that leading AI firms also have sway when they elect to participate in

<sup>4</sup>We thank Nicolas Moës for this observation.

some venues but not others. The OECD Expert Group on AI included representatives from leading firms, whereas engagement at UN efforts, including the Internet Governance Forum (IGF), do not appear to be similarly prioritised. A decentralised regime will enable forum shopping, though further work is needed to determine whether this will help or hurt governance outcomes on the whole.

### 3.6 Policy Coordination

There are good reasons to believe that either centralisation or fragmentation could enhance coordination. A centralised regime can enable easier coordination both across and within policy issues, acting as a focal point for states. Others argue that this is not always the case, and that fragmentation can mutually supportive and even more creative institutions.

Centralisation reduces the occurrence of conflicting mandates and enables communication. These are the ingredients for policy coherence. As noted previously, the WTO has been remarkably successful in ensuring coherent policy and principles across the realm of trade, and even into other areas such as the environment.

However, fragmented regimes can often act as complex adaptive systems. Political requests and communication between secretariats often ensures bottom-up coordination even in the absence of centralisation. Multiple organisations have sought to reduce greenhouse gas emissions within their respective remits, often at the behest of the UNFCCC Conference of Parties. When effective, bottom-up coordination can slowly evolve into centralisation. Indeed, this was the case for the GATT and numerous regional, bilateral and sectoral trade treaties, which all coalesced together into the WTO. While this organic self-organisation has occurred, it has taken decades, with forum shopping and inaction prevailing for many years.

Indeed, some have argued that decentralisation does not just deliver ‘good enough’ global governance [38] that reflects a demand for diverse principles in a multipolar world. Instead, they argue ‘polycentric’ governance approaches [37] may be more creative and legitimate than centrally coordinated regimes. Arguments in favour of polycentricity include the notion that it enables governance initiatives to begin having impacts at diverse scales, and that it enables experimentation with diverse policies and approaches, learning from experience and best practices (*ibid.*, 552). Consequently, these scholars assume “that the invisible hand of a market of institutions leads to a better distribution of functions and effects” [50, p.7].

It is unclear if the different bodies covering AI issues will self-organise or collide. Many of the issues are interdependent and will need to be addressed in tandem. Some particular policy-levers, such as regulating computing power or data, will impact almost all use areas, given that AI progress and use is closely tied to such inputs. Numerous initiatives on AI and robotics are displaying loose coordination [28], but it remains uncertain whether the virtues of a free market of governance will prevail here. Great powers can exercise monopsony-like influence in forum shopping, and the supply of both computing power and machine learning expertise are highly concentrated. In sum, centralisation can reduce competition and enhance coordination, but it may suffocate the creative self-organisation of more fragmented arrangements over time.

## 4 DISCUSSION: WHAT WOULD HISTORY SUGGEST?

### 4.1 A Summary of Considerations

The multilateral track record and peculiarities of AI yield suggestions and warnings for the future. A centralised regime could lower costs, support participation, and act as a powerful new linchpin within the international system. Yet centralisation presents risks for AI governance. It could simply produce a brittle dinosaur, of symbolic value but with little meaningful impact on underlying political or technological issues. A poorly executed attempt could lock-in a poorly designed centralised body: a fate worse than fragmentation. Accordingly, ongoing efforts at the UN, OECD, and elsewhere could benefit from addressing the considerations presented in this paper.

### 4.2 The Limitations of ‘Centralisation vs. Decentralisation’ Debates

Structure is not a panacea. Specific provisions such as agendas and decision-making procedures matter greatly, as do the surrounding politics. Underlying political will may be impacted by framing or connecting policy issues [26, pp.770-1]. The success of a regime is not just a result of fragmentation, but of design details.

Moreover, institutions can be dynamic and broaden over time by taking in new members, or deepen in strengthening commitments. Successful multilateral efforts, such as trade and ozone depletion, tend to do both. We are in the early days of global AI governance. Decisions taken early on will constrain and partially determine the future path. This dependency can even take place across regimes. The Kyoto Protocol was largely shaped by the targets and timetables approach of the Montreal Protocol, which in turn drew from the Convention on Long-range Transboundary Air Pollution. This targets and timetables approach continues today in the way that most countries frame their climate pledges to the Paris Agreement. The choices we make on governing short-term AI challenges will likely shape the management of other policy issues in the long term [7].

On the other hand, committing to centralisation, even if successful, may amount to solving the wrong problem. The problem may not be structural, but geopolitical. Centralisation could even exacerbate the problem by diluting scarce political attention, incurring heavy transaction costs, and shifting discussions away from bodies which have accumulated experience and practice [21]. For example, the Bretton Woods Institutions of the IMF and World Bank, joined later by the WTO, are centralised regimes that engender power. However, those institutions had the express support of the US and may have simply manifested state power in institutional form. Efforts to ban LAWS and create a cyberwarfare convention have been broadly opposed by states with an established technological superiority in these areas [13]. A centralised regime may not unpick these power struggles, but just add a layer of complexity.

## 5 LESSONS AND CONCLUSIONS

Our framework provides a tool for policy-makers to inform their decisions of whether to join, create, or forgo new institutions that tackle AI policy problems. For instance, the recent choice of whether

**Table 1: Regime complex monitoring suggestions**

Key theme	Question	Methods
Conflict	To what extent are regimes’ principles and outputs in opposition over time?	Expert and practitioner survey
Coordination	Are regimes taking steps to complement each other?	Network analysis (e.g. citation network clustering and centrality)
Catalyst	Is the regime complex self-organizing to proactively fill governance gaps?	Natural Language Processing (e.g., entailment and fact checking)

to support the creation of an independent IPAI involved these considerations. Following the US veto, ongoing negotiations for its replacement at the OECD may similarly benefit from their consideration. For now, it is worth closely monitoring the current landscape of AI governance to see if it exhibits enough policy coordination and political power to effectively deal with mounting AI policy problems. While there are promising initial signs [28] there are also already growing governance failures in LAWS, cyberwarfare, and elsewhere.

We outline a suggested monitoring method in Table 1. There are three key areas to monitor: conflict, coordination, and catalyst. First, *conflict* should measure the extent to which principles, rules, regulations and other outcomes from different bodies in the AI regime complex undermine or contradict each other or are in tension either in their principles or goals. Second, *coordination* seeks to measure the proactive steps that AI-related regimes take to work with each other. This includes liaison relationships, joint initiatives, as well as the extent to which their rules, outputs and principles tend to reinforce one another. Third, *catalyst* raises the important question of governance gaps: is the regime complex self-organising to proactively address international AI policy problems? Numerous AI policy problems currently have no clear coverage under international law, including AI-enabled cyber warfare and HLMI. Whether this changes is of vital importance.

These areas require investigation through multiple methods. Qualitative surveys of relevant organisations and actors can yield data on expert perceptions of these questions. Surveys can be augmented with quantitative methods, including network analyses of the regime complex relations [36, p.32]. Natural language processing could be used to examine contradictions and similarities between different regime outputs, e.g., statements, meeting minutes, and more. Monitoring the outcomes of fragmentation can help to determine whether centralisation is needed. One way forward would be to empower the OECD AI Policy Observatory or the UN CEB to regularly review the monitoring outcomes. This could inform a democratic discussion and decision of whether to centralise AI governance further.

Our framework and discussion may also be useful for non-state actors. Researchers and leading AI firms can play an important role in sharing technical expertise and informing forecasts of new policy problems on the horizon. The considerations may benefit their decisions of where to engage. Civil society has a key role as participants, watch-dogs, and catalysts. For example, the Campaign to Stop Killer Robots has sought to boost engagement and support for a LAWS ban within the CCW. Given prolonged delays and a pessimistic outlook, some have articulated a strategy of creating an entirely new forum for the ban, inspired by the Ottawa Treaty which outlawed landmines. Our framework can help reveal the potential virtues

(allowing for progress while avoiding high-threshold deadlocks) and vices (enabling forum shopping) of such an approach. It could even help inform the structure of a future international institution, such as allowing for a modular, flexible structure with ‘critical mass’ agreements. One cross-cutting consideration is clear: a fractured regime sees higher participation costs that may threaten to exclude many civil society organisations altogether.

The international governance of AI is nascent and fragmented. Centralisation under a well-designed, modular, ‘innovation-proof’ framework organisation may be a desirable solution. However, such a move must be approached with caution. How to define its scope and mandate is one problem. Ensuring a politically-acceptable and well-designed body is perhaps a more daunting one. It risks cementing in place a fate worse than fragmentation. Monitoring conflict and coordination in the current AI regime complex, and whether governance gaps are filled, is a prudent way of knowing whether the existing structure can suffice. For now we should closely watch the trajectory of both AI technology and its governance initiatives to determine whether centralisation is worth the risk.

**ACKNOWLEDGMENTS**

The authors would like to express thanks to Seth Baum, Haydn Belfield, Jessica Cussins-Newman, Martina Kunz, Jade Leung, Nicolas Moës, Robert de Neufville, and Nicolas Zahn for valuable comments. Any remaining errors are our own. No conflict of interest is identified.

**REFERENCES**

- [1] Don Anton. 2012. ‘Treaty Congestion’ in International Environmental Law. In *Routledge Handbook of International Environmental Law*, Shawkat Alam, Jahid Hossain Bhuiyan, Tareq M.R. Chowdhury, and Erika J. Techera (Eds.). Routledge. <https://www.taylorfrancis.com/books/9780203093474>
- [2] Lucio Baccaro and Valentina Mele. 2012. Pathology of Path Dependency? The ILO and the Challenge of New Governance. *ILR Review* 65, 2 (April 2012), 195–224. <https://doi.org/10.1177/001979391206500201>
- [3] Frank Biermann, Philipp Pattberg, Harro van Asselt, and Fariborz Zelli. 2009. The Fragmentation of Global Governance Architectures: A Framework for Analysis. *Global Environmental Politics* 9, 4 (Oct. 2009), 14–40. <https://doi.org/10.1162/glep.2009.9.4.14>
- [4] John Braithwaite and Peter Drahos. 2000. *Global Business Regulation*. Cambridge University Press. Google-Books-ID: DcEEW5OGWLcC.
- [5] James Butcher and Irakli Beridze. 2019. What is the State of Artificial Intelligence Governance Globally? *The RUSI Journal* 164, 5-6 (2019), 88–96. <https://doi.org/10.1080/03071847.2019.1694260> arXiv:<https://doi.org/10.1080/03071847.2019.1694260>
- [6] Ryan Calo. 2017. Artificial Intelligence Policy: A Primer and Roadmap. *UC Davis Law Review* 51 (2017), 37. [https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2\\_calo.pdf](https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_calo.pdf)
- [7] Stephen Cave and Seán S. ÓhÉigeartaigh. 2019. Bridging near- and long-term concerns about AI. *Nature Machine Intelligence* 1, 1 (Jan. 2019), 5. <https://doi.org/10.1038/s42256-018-0003-2>
- [8] Peter Cihon. 2019. *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Technical Report. Center for the Governance of AI, Future of Humanity Institute, University of Oxford, Oxford. <https://www.fhi.ox.ac.uk/wp-content/uploads/Standards.FHI-Technical>

- Report.pdf
- [9] Allan Dafoe. 2018. *AI Governance: A Research Agenda*. Technical Report. Center for the Governance of AI, Future of Humanity Institute, Oxford. 52 pages. <https://www.fhi.ox.ac.uk/govaiagenda/>
- [10] Janosch Delcker. 2019. How killer robots overran the UN. *POLITICO* (Feb. 2019). <https://www.politico.eu/article/killer-robots-overran-united-nations-lethal-autonomous-weapons-systems/>
- [11] Abigail C. Deshman. 2011. Horizontal Review between International Organizations: Why, How, and Who Cares about Corporate Regulatory Capture. *European Journal of International Law* 22, 4 (Nov. 2011), 1089–1113. <https://doi.org/10.1093/ejil/chr093>
- [12] Robyn Eckersley. 2004. The Big Chill: The WTO and Multilateral Environmental Agreements. *Global Environmental Politics* 4, 2 (May 2004), 24–50. <https://doi.org/10.1162/152638004323074183>
- [13] Mette Eilstrup-Sangiovanni. 2018. Why the World Needs an International Cyberwar Convention. *Philosophy & Technology* 31, 3 (Sept. 2018), 379–407. <https://doi.org/10.1007/s13347-017-0271-5>
- [14] Olivia J Erdelyi and Judy Goldsmith. 2018. Regulating Artificial Intelligence: Proposal for a Global Solution. In *Proceedings of the 2018 AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*. 7. [http://www.aies-conference.com/wp-content/papers/main/AIES2018\\_aper13.pdf](http://www.aies-conference.com/wp-content/papers/main/AIES2018_aper13.pdf)
- [15] Daniel C Esty and Maria H Ivanova. 2002. Revitalizing Global Environmental Governance: A Function-Driven Approach. In *Global Environmental Governance: Options & Opportunities*, Daniel C Esty and Maria H Ivanova (Eds.). Yale School of Forestry and Environmental Studies. <https://environment.yale.edu/publication-series/documents/downloads/a-g/esty-ivanova.pdf>
- [16] Urs Gasser and Virgilio A.F. Almeida. 2017. A Layered Model for AI Governance. *IEEE Internet Computing* 21, 6 (Nov. 2017), 58–62. <https://doi.org/10.1109/MIC.2017.4180835>
- [17] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research* 62 (2018), 729–754.
- [18] Laurence Helfer. 2004. Regime Shifting: The TRIPs Agreement and New Dynamics of International Intellectual Property Lawmaking. *Yale Journal of International Law* 29 (Jan. 2004), 1–83. [https://scholarship.law.duke.edu/faculty\\_scholarship/2014](https://scholarship.law.duke.edu/faculty_scholarship/2014)
- [19] on Digital Cooperation High-Level Panel. 2019. The Age of Digital Interdependence Report. *UN Secretary General* (2019).
- [20] ITU. 2019. *United Nations Activities on Artificial Intelligence (AI) 2019*. Technical Report. ITU. 88 pages. [https://www.itu.int/dms\\_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf](https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf)
- [21] Calestous Juma. 2000. Commentary: The Perils of Centralizing Global Environmental Governance. *Environment: Science and Policy for Sustainable Development* 42, 9 (Nov. 2000), 44–45. <https://doi.org/10.1080/00139150009605765>
- [22] Joseph Jupille, Walter Mattli, and Duncan Snidal. 2013. *Institutional Choice and Global Commerce*. Cambridge University Press, Cambridge. OCLC: 900490808.
- [23] Luke Kemp. 2017. US-proofing the Paris Climate Agreement. *Climate Policy* 17, 1 (Jan. 2017), 86–101. <https://doi.org/10.1080/14693062.2016.1176007>
- [24] Luke Kemp, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 2019. UN High-level Panel on Digital Cooperation: A Proposal for International AI Governance. <https://digitalcooperation.org/wp-content/uploads/2019/02/LukeKempSubmission-to-the-UN-High-Level-Panel-on-Digital-Cooperation-2019-Kemp-et-al.pdf>
- [25] Barbara Koremenos, Charles Lipson, and Duncan Snidal. 2001. Rational Design: Looking Back to Move Forward. *International Organization* 55, 4 (2001), 1051–1082. <https://doi.org/10.1162/002081801317193691>
- [26] Barbara Koremenos, Charles Lipson, and Duncan Snidal. 2001. The Rational Design of International Institutions. *International Organization* 55, 4 (2001), 761–799. <https://doi.org/10.1162/002081801317193592>
- [27] Stephen D. Krasner. 1982. Structural Causes and Regime Consequences: Regimes as Intervening Variables. *International Organization* 36, 2 (1982), 185–205. <https://doi.org/10.1017/S0020818300018920>
- [28] Martina Kunz and Seán Ó hÉigeartaigh. 2020. Artificial Intelligence and Robotization. In *Oxford Handbook on the International Law of Global Security*, Robin Geiss and Nils Melzer (Eds.). Oxford University Press. <https://papers.ssrn.com/abstract=3310421>
- [29] Matthijs M. Maas. 2019. Innovation-Proof Governance for Military AI? how I learned to stop worrying and love the bot. *Journal of International Humanitarian Legal Studies* 10, 1 (2019), 129–157. <https://doi.org/10.1163/18781527-01001006>
- [30] Gary E Marchant, Braden R Allenby, and Joseph R Herkert. 2011. *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*. Vol. 7. Springer Science & Business Media.
- [31] Will Martin and Patrick Messerlin. 2007. Why is it so difficult? Trade liberalization under the Doha Agenda. *Oxford Review of Economic Policy* 23, 3 (2007), 347–366.
- [32] Jean-Frédéric Morin, Hugo Dobson, Claire Peacock, Miriam Prys-Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietsch, Judit Fabian, John Kirton, Raffaele Marchetti, Simone Romano, Miranda Schreurs, Arthur Silve, and Elisabeth Vallet. 2019. How Informality Can Address Emerging Issues: Making the Most of the G7. *Global Policy* 10, 2 (May 2019), 267–273. <https://doi.org/10.1111/1758-5899.12668>
- [33] Milton Mueller, John Mathiason, and Hans Klein. 2007. The Internet and Global Governance: Principles and Norms for a New Regime. *Global Governance* 2 (2007), 237–254. <https://heinonline.org/HOL/P?h=hein.journals/glogo13&i=245>
- [34] Nils J. Nilsson. 2009. *The Quest for Artificial Intelligence* (1 edition ed.). Cambridge University Press, Cambridge ; New York.
- [35] Joseph S. Nye. 2014. *The Regime Complex for Managing Global Cyber Activities*. Technical Report 1. Global Commission on Internet Governance. <https://dash.harvard.edu/bitstream/handle/1/12308565/Nye-GlobalCommission.pdf>
- [36] Amandine Orsini, Jean-Frédéric Morin, and Oran Young. 2013. Regime Complexes: A Buzz, a Boom, or a Boost for Global Governance? *Global Governance: A Review of Multilateralism and International Organizations* 19, 1 (Aug. 2013), 27–39. <https://doi.org/10.1163/19426720-01901003>
- [37] Elinor Ostrom. 2010. Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change* 20, 4 (Oct. 2010), 550–557. <https://doi.org/10.1016/j.gloenvcha.2010.07.004>
- [38] Stewart Patrick. 2014. The Unruly World: The Case for Good Enough Global Governance. *Foreign Affairs* 93, 1 (2014), 58–73.
- [39] Eleonore Pauwels. 2019. *The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI*. Technical Report. United Nations University - Centre for Policy Research. 83 pages. <https://i.unu.edu/media/cpr.unu.edu/attachment/3472/PauwelsAIgeopolitics.pdf>
- [40] Saadia M. Pekkanen, Mireya Solís, and Saori N. Katada. 2007. Trading Gains for Control: International Trade Forums and Japanese Economic Diplomacy. *International Studies Quarterly* 51, 4 (2007), 945–970. <https://www.jstor.org/stable/4621750>
- [41] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. 2019. *The AI Index 2019 Annual Report*. Technical Report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA. [https://hai.stanford.edu/sites/g/files/tbjybj10986/f/ai\\_index\\_019\\_eport.pdf](https://hai.stanford.edu/sites/g/files/tbjybj10986/f/ai_index_019_eport.pdf)
- [42] Colin B. Picker. 2001. A View from 40,000 Feet: International Law and the Invisible Hand of Technology. *Cardozo Law Review* 23 (2001), 151–219. <https://papers.ssrn.com/abstract=987524>
- [43] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2019. Tackling Climate Change with Machine Learning. *arXiv:1906.05433 [cs, stat]* (June 2019). <http://arxiv.org/abs/1906.05433> arXiv: 1906.05433.
- [44] Matthew U. Scherer. 2016. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology* 2 (2016). <http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf>
- [45] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnneLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. 2016. *Artificial Intelligence and Life in 2030*. Technical Report. Stanford University, Stanford, CA. <http://ai100.stanford.edu/2016-report>
- [46] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Langhans, Max Tegmark, and Francesco Fusco Nerini. 2019. The role of artificial intelligence in achieving the Sustainable Development Goals. *arXiv:1905.00501 [cs]* (April 2019). <http://arxiv.org/abs/1905.00501> arXiv: 1905.00501.
- [47] Wendell Wallach and Gary E Marchant. 2018. An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics. In *Proceedings of the 2018 AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*. 7. [https://www.aies-conference.com/2018/contents/papers/main/AIES2018\\_aper7.7.pdf](https://www.aies-conference.com/2018/contents/papers/main/AIES2018_aper7.7.pdf)
- [48] John Frank Weaver. 2014. Autonomous Weapons and International Law: We Need These Three International Treaties to Govern “Killer Robots”. *Slate Magazine* (Dec. 2014). <https://slate.com/technology/2014/12/autonomous-weapons-and-international-law-we-need-these-three-treaties-to-govern-killer-robots.html>
- [49] Fariborz Zelli. 2011. The fragmentation of the global climate governance architecture. *Wiley Interdisciplinary Reviews: Climate Change* 2, 2 (2011), 255–270. <https://doi.org/10.1002/wcc.104>
- [50] Fariborz Zelli and Harro van Asselt. 2013. Introduction: The Institutional Fragmentation of Global Environmental Governance: Causes, Consequences, and Responses. *Global Environmental Politics* 13, 3 (July 2013), 1–13. [https://doi.org/10.1162/GLEP\\_a0180](https://doi.org/10.1162/GLEP_a0180)