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

Peter Cihon,[1*] **Matthijs M. Maas,**[1,2*] **Luke Kemp**[3*]
[1]Centre for the Governance of AI, Future of Humanity Institute, University of Oxford
[2]Centre for International Law, Conflict, and Crisis, Faculty of Law, University of Copenhagen
[3]Centre for the Study of Existential Risk, University of Cambridge
petercihon@gmail.com; matthijs.maas@jur.ku.dk; 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.

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 (Perrault et al. 2019), the priority of governance initiatives reflects the growing appreciation that AI has the potential to dramatically change the world for both good and ill (Dafoe 2018). 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 (Cave and ÓhÉigeartaigh 2019). 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 (Scherer 2016; Calo 2017; Gasser and Almeida 2017). Research into AI *global* governance remains relatively nascent (though see Butcher and Beridze 2019). Kemp et al. (2019) have called for specialised, centralised intergovernmental agencies to coordinate policy responses globally, and others have called for a centralised 'International Artificial Intelligence Organisation' (Erdelyi and Goldsmith 2018). Others favour more decentralised arrangements based around 'Governance Coordinating Committees', global standards, or existing international law instruments (Wallach and Marchant 2018; Cihon 2019; Kunz and Ó hÉigeartaigh 2020).

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 (Biermann et al. 2009) and 'regime complexes' (Orsini, Morin, and Young 2013) to assess considerations in centralising the international governance of AI. We draw on the history of other international regimes[1] 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 de-

[*]Equal contribution, order selected at random.
Working paper, last updated 15 December, 2019.

[1]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'(Krasner 1982, 186).

signed body could prove a fate worse than fragmentation. Second, for now there should be structured monitoring of existing efforts to see whether they are they are self-organising or insufficient.

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 (Stone et al. 2016). This is an important but not entirely convincing objection. The technical field has no settled definition for 'AI',[2] 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).[3] 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 (see Rolnick et al. 2019), or help meet the Sustainable Development Goals (see Vinuesa et al. 2019). 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 (ITU 2019). 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 (Kunz and Ó hÉigeartaigh 2020), 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 (High-Level Panel 2019). 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 (see Morin et al. 2019, 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. Our other definitions for key terms are provided below in Table 1. 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.

Centralisation Criteria: A 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.

[2] We define 'AI' as any machine system capable of functioning 'appropriately and with foresight in its environment' (Nilsson 2009, 13; see too Dafoe 2018, 5).

[3] 'High-level machine intelligence' has been defined as 'unaided machines [that] can accomplish every task better and more cheaply than human workers' (Grace et al. 2018, 1).

Table 1: Definition of Key Governance Terms

Term	Definition
Fragmentation or Decentralization	A patchwork of international organisations and institutions which focus on a particular issue area but differ in scope, membership and often rules (Biermann et al. 2009, 16).
Centralisation	An arrangement in which governance of a particular issue lies under the authority of a single umbrella body. This is a spectrum from highly centralised (the role of the WTO in trade) to decentralised (the plethora of multilateral environmental agreements).
Regime Complex	A network of three or more international regimes on a common issue area. These should have overlapping membership and cause potentially problematic interactions (Orsini, Morin, and Young 2013, 29).

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 (see Orsini, Morin, and Young 2013, 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 (Biermann et al. 2009). 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 (Nye 2014). 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 (Mueller, Mathiason, and Klein 2007).

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 (Eckersley 2004). 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 (Koremenos, Lipson, and Snidal 2001a). There may also be benefits to housing a foresight capacity within the regime complex, to allow for accelerated or even proactive efforts (Pauwels 2019). Centralised AI governance would enable an empowered organisation to more effectively use foresight analyses to inform policy responses across the regime complex.

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 (Esty and Ivanova 2002). 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 (Morin et al. 2019, 2).

Historically, decentralised regimes have presented cost and related participation concerns. Hundreds of related and sometimes overlapping international environmental agreements can create 'treaty congestion' (Anton 2012). This complicates participation and implementation for both developed and developing nations (Esty and Ivanova 2002). 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. 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 (Martin and Messerlin 2007). International law has been quick to respond to technological changes in some cases, and delayed in others (Picker 2001, 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 (Morin et al. 2019, 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 (Picker 2001, 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 (Deshman 2011), 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 (Baccaro and Mele 2012). 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.[4] 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'(Maas 2019) governance regimes. Periodic renegotiation, modular expansion, 'principles based regulation', or sunset clauses can also support ongoing reform (see generally Marchant, Allenby, and Herkert 2011, 29-30). Such approaches have often proved successful historically, due partially to decentralisation but, importantly, also to particular designs.

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 (Kemp 2017). 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 (Zelli 2011, 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)

[4] We thank Nicolas Moës for this observation.

could instead allow for the powerful to ratify and move forward at least a few of those options (Weaver 2014).

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 these dilemma poses a difficult knot for centralisation to unravel.

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 (Braithwaite and Drahos 2000). 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 (Pekkanen, Solís, and Katada 2007) and placate constituents by shifting to a toothless forum (Helfer 2004).

The ability to successfully forum-shop depends on an actor's power. Most successful examples of forum-shifting have been led by the US (Braithwaite and Drahos 2000). 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 (Helfer 2004), 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 (Jupille, Mattli, and Snidal 2013).

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 (Delcker 2019). 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 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.

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 (Patrick 2014) that reflects a demand for diverse principles in a multipolar world. Instead, they argue 'polycentric' governance approaches (Ostrom 2010) 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" (Zelli and van Asselt 2013, 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 (Kunz and Ó hÉigeartaigh 2020), 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.

Discussion: What Would History Suggest?

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, a summary of which is presented in Table 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 (Koremenos, Lipson, and Snidal 2001b, 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. The choices we make on governing short-term AI challenges will likely shape the management of other policy issues in the long term (Cave and ÓhÉigeartaigh 2019).

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 (Juma 2000). 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 (Eilstrup-Sangiovanni 2018). A centralised regime may not unpick these power struggles, but just add a layer of complexity.

HLMI: An Illustrative Example

The promise and peril of centralisation may differ by policy issue. HLMI is one issue that is markedly unique: HLMI is distinct in its risk profile, uncertainty and linkage to other AI policy issues, which can make it an interesting case through which to explore the tradeoffs of a centralised AI governance regime in a fresh context. While timelines are uncertain, the creation of HLMI systems is the express goal of various present-day projects (Baum 2017), and the future development of an unaligned HLMI could have catastrophic consequences (GCF 2018). Creation of a controlled HLMI by a subset of private or public actors could lead to grotesque power imbalances. It could also exacerbate other AI policy problems, such as labour automation and advanced military applications (by providing a coordinating platform, strategic advisor, or in nuclear command and control). Addressing many shorter-term issues, such as cyberwarfare, and improving global governance more broadly will have significant impacts on HLMI development and deployment (Kunz and Ó hÉigeartaigh 2020). There is also marked uncertainty about whether HLMI can be created in a single system, what it would look like, and significant disagreement as to how long this would take (Grace et al. 2018).

Below in Table 3 we provide a brief application of our framework to HLMI. It shows that centralisation of governance is particularly promising for HLMI. This is due to its neglect, stakes, scope, and need for informed, preemptive policy. Many other issues, such as the advanced military applications of AI systems, may similarly be more productively or safely subjected to cooperation, i.e., centralised. These cases may prove a broad rule about the value of centralisation, or they may be outliers. Rather than any AI governance blueprint, our trade-offs framework provides one way of thinking through the costs and benefits of centralising governance either on or across specific AI issues. Identifying areas which are more easily defined and garner the benefits of centralised regulation provides an organic approach to thinking through what subset of topics an AI umbrella body could cover. HLMI, appears to one of the most appealing candidates.

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 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 (Kunz and Ó hÉigeartaigh 2020) there are also already growing governance failures in LAWS, cyberwarfare, and elsewhere.

We outline a suggested monitoring method in Table 4. 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

Table 2: Summary of Considerations

Consideration	**Implications for Centralisation**	**Historical Example**	**AI Policy Issue Example**
Political Power	Pro	*Shaping other regimes:* WTO has created a chilling effect, where the fear of conflicting with WTO norms and rules has led environmental treaties to self-censor to avoid addressing trade-related measures.	Empowered regime using foresight on AI systems development can address policy problems more quickly.
Efficiency & Participation	Pro	*Decentralisation raises inefficiencies and barriers:* The proliferation of multilateral environmental agreements poses costs and barriers to participation in negotiation, implementation, and monitoring.	AI companies engage and share expertise, but if not checked by adversarial civil society, there is a greater concern of regulatory capture; increased costs undermine civil society participation.
Slowness & Brittleness	Con	*Slowness:* Under the GATT, 1947 tariff negotiations among 19 countries took 8 months. The Uruguay round, beginning in 1986, took 91 months for 125 parties to agree on reductions. *Regulatory capture:* WHO accused offor undue corporate influence in response to 2009 H1N1 pandemic. *Pathology of path-dependence:* Failed ILO reform attempts.	Process of centralised regime can not keep pace with high speed of AI progress and deployment, may miss the window of opportunity. Advanced AI issues (especially HLMI) may rapidly shift the risk landscape or problem portfolio of AI, beyond the narrow scope of an older institutional mandate
Breadth vs. Depth Dilemma	Con	*Watering down:* 2015 Paris Agreement suggest attempts to 'get all parties on board' to centralized regime may result in significant watering down.	Attempts to effectively govern the military uses of AI have been resisted by the most powerful states. Attempted to create an IPAI have been resisted by the US and shifted to a smaller forum (the OECD).
Forum Shopping	Depends on design	*Power predicts outcomes:* Intellectual property in trade shifted from UNCTAD to WIPO to WTO, with developed countries getting their way. *Accelerates progress:* NGOs and some states shifted discussions of anti-personnel mines ban away from CCW, ultimately resulting in the Ottawa Treaty.	Governance of military AI systems is fractured across CCW, multiple GGEs. This strategy may catalyze progress, but brings risks of fracture.
Policy Coordination	Depends on design	*Strong, but delayed convergence:* Diverse regimes can coalesce into centralized regime, as seen with GATT and numerous trade treaties coalescing into the WTO, but doing so may take many decades.	Numerous AI governance initiatives display loose coordination, but it is unclear if these initiatives can respond to policy developments in a timely manner.

Table 3: An Application of the Framework to High-Level Machine Intelligence (HLMI)

Consideration	**HLMI**
Political Power	Uncertainty around HLMI development makes credible forecasting particularly important. Understanding which inputs drive AI progress, and when and by whom HLMI could be created, is paramount to ensuring safe development (see Dafoe 2018). This will require a coordinated effort to track and forecast HLMI project efforts (see Baum 2017), as well as a politically empowered organisation to quickly act upon this information. The potentially catastrophic risks make the increased political power of a centralised institution desirable. The creation of HLMI, if it can be done by a well-resourced actor, is a 'free-driver' issue. An effective response needs to have the teeth to deter major players from acting unilaterally.
Efficiency & Participation	Centralisation would support economies of scale in expertise to support efficient governance. Given the significant financial resources and infrastructure likely needed for such a project, a joint global effort could be an efficient way to govern HLMI research.
Slowness & Brittleness	If short timelines (less than 10-15 years) are expected, the lengthy period to negotiate and create such a body would be a critical weakness. If longer timelines are more likely, there should be sufficient time to develop a centralised anchor institution. Institutional capture is a concern given the abundance of wealthy corporate actors involved in creating HLMI (Google, OpenAI, Microsoft). However, it is unclear if this would be more likely under a centralised body.
Depth vs. Breadth Dilemma	The limited scope of actors makes centralisation more feasible. Costs and requisite tacit knowledge may restrict the development of HLMI to a few powerful players. The breadth vs. depth dilemma could be avoided through a 'minilateral' or critical mass approach that initially involves only the few countries that are capable of developing it, although there would be benefits to broadening membership, such as legitimacy and fairness.
Forum Shopping	A centralised body is well placed to prevent forum-shopping as there is currently no coverage of HLMI development and deployment under international law.
Policy Coordination	Coordination is key for HLMI. It has close connections to issues such as labour automation and automated cyberwarfare. The creation or use of HLMI is not directly regulated by any treaties or legal instruments. This makes the creation of a new, dedicated institution to address cover it easier and less unlikely to trigger turf wars. It also makes it less likely that the existing tapestry of international law can quickly self-organise to cover HLMI.

Table 4: Regime Complex Monitoring Suggestions

Key Theme	**Questions**	**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)

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 (Orsini, Morin, and Young 2013, 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, watchdogs, 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.

Acknowledgements

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

Anton, D. 2012. 'Treaty Congestion' in International Environmental Law. In Alam, S.; Bhuiyan, J. H.; Chowdhury, T. M.; and Techera, E. J., eds., *Routledge Handbook of International Environmental Law*. Routledge.

Baccaro, L., and Mele, V. 2012. Pathology of Path Dependency? The ILO and the Challenge of New Governance. *ILR Review* 65(2):195–224.

Baum, S. 2017. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. SSRN Scholarly Paper ID 3070741, Social Science Research Network, Rochester, NY.

Biermann, F.; Pattberg, P.; van Asselt, H.; and Zelli, F. 2009. The Fragmentation of Global Governance Architectures: A Framework for Analysis. *Global Environmental Politics* 9(4):14–40.

Braithwaite, J., and Drahos, P. 2000. *Global Business Regulation*. Cambridge University Press. Google-Books-ID: DcEEW5OGWLcC.

Butcher, J., and Beridze, I. 2019. What is the state of artificial intelligence governance globally? *The RUSI Journal* 164(5-6):88–96.

Calo, R. 2017. Artificial Intelligence Policy: A Primer and Roadmap. *UC Davis Law Review* 51:37.

Cave, S., and ÓhÉigeartaigh, S. S. 2019. Bridging near- and long-term concerns about AI. *Nature Machine Intelligence* 1(1):5.

Cihon, P. 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, Oxford.

Dafoe, A. 2018. AI Governance: A Research Agenda. Technical report, Center for the Governance of AI, Future of Humanity Institute, Oxford.

Delcker, J. 2019. How killer robots overran the UN. *POLITICO*.

Deshman, A. C. 2011. Horizontal Review between International Organizations: Why, How, and Who Cares about Corporate Regulatory Capture. *European Journal of International Law* 22(4):1089–1113.

Eckersley, R. 2004. The Big Chill: The WTO and Multilateral Environmental Agreements. *Global Environmental Politics* 4(2):24–50.

Eilstrup-Sangiovanni, M. 2018. Why the World Needs an International Cyberwar Convention. *Philosophy & Technology* 31(3):379–407.

Erdelyi, O. J., and Goldsmith, J. 2018. Regulating Artificial Intelligence: Proposal for a Global Solution. In *Proceedings of the 2018 AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*, 7.

Esty, D. C., and Ivanova, M. H. 2002. Revitalizing Global Environmental Governance: A Function-Driven Approach. In Esty, D. C., and Ivanova, M. H., eds., *Global Environmental Governance: Options & Opportunities*. Yale School of Forestry and Environmental Studies.

Gasser, U., and Almeida, V. A. 2017. A Layered Model for AI Governance. *IEEE Internet Computing* 21(6):58–62.

GCF. 2018. Global Catastrophic Risks 2018. Technical report, Global Challenges Foundation.

Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; and Evans, O. 2018. When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research* 62:729–754.

Helfer, L. 2004. Regime Shifting: The TRIPs Agreement and New Dynamics of International Intellectual Property Lawmaking. *Yale Journal of International Law* 29:1–83.

High-Level Panel, o. D. C. 2019. The age of digital interdependence report. *UN Secretary General*.

ITU. 2019. United Nations Activities on Artificial Intelligence (AI) 2019. Technical report, ITU.

Juma, C. 2000. Commentary: The Perils of Centralizing Global Environmental Governance. *Environment: Science and Policy for Sustainable Development* 42(9):44–45.

Jupille, J.; Mattli, W.; and Snidal, D. 2013. *Institutional Choice and Global Commerce*. Cambridge: Cambridge University Press. OCLC: 900490808.

Kemp, L.; Cihon, P.; Maas, M. M.; Belfield, H.; Cremer, Z.; Leung, J.; and Ó hÉigeartaigh, S. 2019. UN High-level Panel on Digital Cooperation: A Proposal for International AI Governance.

Kemp, L. 2017. US-proofing the Paris Climate Agreement. *Climate Policy* 17(1):86–101.

Koremenos, B.; Lipson, C.; and Snidal, D. 2001a. Rational Design: Looking Back to Move Forward. *International Organization* 55(4):1051–1082.

Koremenos, B.; Lipson, C.; and Snidal, D. 2001b. The Rational Design of International Institutions. *International Organization* 55(4):761–799.

Krasner, S. D. 1982. Structural Causes and Regime Consequences: Regimes as Intervening Variables. *International Organization* 36(2):185–205.

Kunz, M., and Ó hÉigeartaigh, S. 2020. Artificial Intelligence and Robotization. In Geiss, R., and Melzer, N., eds., *Oxford Handbook on the International Law of Global Security*. Oxford University Press.

Maas, M. M. 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):129–157.

Marchant, G. E.; Allenby, B. R.; and Herkert, J. R. 2011. *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*, volume 7. Springer Science & Business Media.

Martin, W., and Messerlin, P. 2007. Why is it so difficult? trade liberalization under the doha agenda. *Oxford Review of Economic Policy* 23(3):347–366.

Morin, J.; Dobson, H.; Peacock, C.; Prys-Hansen, M.; Anne, A.; Bélanger, L.; Dietsch, P.; Fabian, J.; Kirton, J.; Marchetti, R.; Romano, S.; Schreurs, M.; Silve, A.; and Vallet, E. 2019. How Informality Can Address Emerging Issues: Making the Most of the G7. *Global Policy* 10(2):267–273.

Mueller, M.; Mathiason, J.; and Klein, H. 2007. The Internet and Global Governance: Principles and Norms for a New Regime. *Global Governance* (2):237–254.

Nilsson, N. J. 2009. *The Quest for Artificial Intelligence*. Cambridge ; New York: Cambridge University Press, 1 edition edition.

Nye, J. S. 2014. The Regime Complex for Managing Global Cyber Activities. Technical Report 1, Global Commission on Internet Governance.

Orsini, A.; Morin, J.-F.; and Young, O. 2013. Regime Complexes: A Buzz, a Boom, or a Boost for Global Governance? *Global Governance: A Review of Multilateralism and International Organizations* 19(1):27–39.

Ostrom, E. 2010. Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change* 20(4):550–557.

Patrick, S. 2014. The Unruled World: The Case for Good Enough Global Governance. *Foreign Affairs* 93(1):58–73.

Pauwels, E. 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.

Pekkanen, S. M.; Solís, M.; and Katada, S. N. 2007. Trading Gains for Control: International Trade Forums and Japanese Economic Diplomacy. *International Studies Quarterly* 51(4):945–970.

Perrault, R.; Shoham, Y.; Brynjolfsson, E.; Clark, J.; Etchemendy, J.; Grosz, B.; Lyons, T.; Manyika, J.; Mishra, S.; and Niebles, J. C. 2019. The AI Index 2019 Annual Report. Technical report, AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA.

Picker, C. B. 2001. A View from 40,000 Feet: International Law and the Invisible Hand of Technology. *Cardozo Law Review* 23:151–219.

Rolnick, D.; Donti, P. L.; Kaack, L. H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A. S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; Luccioni, A.; Maharaj, T.; Sherwin, E. D.; Mukkavilli, S. K.; Kording, K. P.; Gomes, C.; Ng, A. Y.; Hassabis, D.; Platt, J. C.; Creutzig, F.; Chayes, J.; and Bengio, Y. 2019. Tackling Climate Change with Machine Learning. *arXiv:1906.05433 [cs, stat]*. arXiv: 1906.05433.

Scherer, M. U. 2016. Regulating Artificial Intelligence

Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology* (2).

Stone, P.; Brooks, R.; Brynjolfsson, E.; Calo, R.; Etzioni, O.; Hager, G.; Hirschberg, J.; Kalyanakrishnan, S.; Kamar, E.; Kraus, S.; Leyton-Brown, K.; Parkes, D.; Press, W.; Saxenian, A.; Shah, J.; Tambe, M.; and Teller, A. 2016. Artificial Intelligence and Life in 2030. Technical report, Stanford University, Stanford, CA.

Vinuesa, R.; Azizpour, H.; Leite, I.; Balaam, M.; Dignum, V.; Domisch, S.; Felländer, A.; Langhans, S.; Tegmark, M.; and Nerini, F. F. 2019. The role of artificial intelligence in achieving the Sustainable Development Goals. *arXiv:1905.00501 [cs]*. arXiv: 1905.00501.

Wallach, W., and Marchant, G. E. 2018. An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics. 7.

Weaver, J. F. 2014. Autonomous Weapons and International Law: We Need These Three International Treaties to Govern "Killer Robots". *Slate Magazine*.

Zelli, F., and van Asselt, H. 2013. Introduction: The Institutional Fragmentation of Global Environmental Governance: Causes, Consequences, and Responses. *Global Environmental Politics* 13(3):1–13.

Zelli, F. 2011. The fragmentation of the global climate governance architecture. *Wiley Interdisciplinary Reviews: Climate Change* 2(2):255–270.