Aligning AI Regulation to Sociotechnical Change

Matthijs M. Maas[1]
University of Cambridge

In: Justin Bullock, Baobao Zhang, Yu-Che Chen, Johannes Himmelreich, Matthew Young, Antonin Korinek & Valerie Hudson (eds.). *Oxford Handbook on AI Governance* (Oxford University Press, 2022 forthcoming).

Section 3: Developing an AI Governance Regulatory Ecosystem
Pre-print draft, v.0.96, last updated: July 16th, 2021

Abstract

How do we regulate a changing technology, with changing uses, in a changing world? This chapter argues that while existing (inter)national AI governance approaches are important, they are often siloed. Technology-centric approaches focus on individual AI applications; law-centric approaches emphasize AI's effects on pre-existing legal fields or doctrines. This chapter argues that to foster a more systematic, functional and effective AI regulatory ecosystem, policy actors should instead complement these approaches with a regulatory perspective that emphasizes how, when, and why AI applications enable patterns of 'sociotechnical change'. Drawing on theories from the emerging field of 'TechLaw', it explores how this perspective can provide informed, more nuanced, and actionable perspectives on AI regulation. A focus on sociotechnical change can help analyse when and why AI applications actually do create a meaningful rationale for new regulation—and how they are consequently best approached as targets for regulatory intervention, considering not just the technology, but also six distinct 'problem logics' that appear around AI issues across domains. The chapter concludes by briefly reviewing concrete institutional and regulatory actions that can draw on this approach in order to improve the regulatory triage, tailoring, timing & responsiveness, and design of AI policy.

Keywords: AI, regulation, sociotechnical change, techlaw, regulatory rationale, regulatory target, problem logics

Introduction

How do we regulate a changing technology, with changing uses, in a changing world? As artificial intelligence ('AI') is anticipated to drive extensive change, the question of how we can and should reconfigure our regulatory ecosystems for AI change matters today. In just the past decade, advances in both AI research and in the broader data infrastructure

[1] Postdoctoral Research Associate, Centre for the Study of Existential Risk, University of Cambridge | College Research Associate, King's College, University of Cambridge | Email: mmm71@cam.ac.uk. | ORCID ID: 0000-0002-6170-9393

have begun to spur extensive take-up of this technology across society (D. Zhang et al. 2021). As a 'general-purpose technology' (Trajtenberg 2018), AI's impact on the world may be both unusually broad and deep. It may even prove as 'transformative' as the industrial revolution (Gruetzemacher and Whittlestone 2020). This may provide grounds for anticipation—and also for caution. Most of all, it is grounds for reflection on the choices that societies want to make and instil in the trajectory of this technology.

While still at an early stage of development, uses of AI technology are already creating diverse policy challenges. Internationally, AI is the subject to intense contestation. Further AI progress, along with the technology's global dissemination, are set to further raise the stakes. It is clearly urgent to reflect on the purposes and suitability of the regulatory ecosystem for AI governance.

The urgent question is; when we craft AI regulation, how should we do so? Many AI governance approaches at both the national and international level, remain hampered by siloed policy responses to individual AI applications (that is, they are technology-centric), or for AI applications' effects on individual legal fields or doctrines (that is, they are law-centric).

In contrast, this chapter argues that to craft adequate AI policies, a better regulatory perspective takes a step back, and first asks a better question: when we craft AI regulation, what are we seeking to regulate? And what are the ingredients for a more systematic regulatory template for crafting AI governance?

This chapter argues that to foster an effective AI regulatory ecosystem, policy institutions and actors must be equipped to craft AI policies in alignment with systematic assessments of how, when, and why AI applications enable broader forms of sociotechnical change (Maas 2020). It argues that this approach complements existing technology-centric and law-centric examinations of AI policies; and that, supported by adequate institutional processes, it can provide an informed and actionable perspectives on when and why AI applications actually create a *rationale* for regulation—and how they are consequently best approached as *targets* for regulatory intervention. This enables more tailored policy formulation for AI issues, facilitates oversight and review of these policies, and helps address structural accountability, alignment, and (lack of) information problems in the emerging AI governance regulatory ecosystem.

The chapter is structured as follows. It first (1) sketches the general value of a 'change-centric' approach to AI governance. The chapter then (2) proposes and articulates a framework focused on 'sociotechnical change', and explores how this model allows an improved consideration of (3) when an AI application creates a regulatory *rationale*, and (4) how it is subsequently best approached as a regulatory *target*, considering 6 distinct 'problem logics' that appear in AI issues across domains. Finally (5), the chapter reflects on some of the limits of this approach, before discussing concrete institutional and regulatory actions that can draw on this approach in order to improve the regulatory *triage, tailoring, timing & responsiveness,* and *regulatory design* of AI policy.

Towards change-centric approaches in an AI regulatory ecosystem

In response to AI's emerging challenges, scholars and policymakers have appealed to a wide spectrum of regulatory tools to govern AI. It should be no surprise that recent

years have seen increasing public demands for AI regulation (B. Zhang and Dafoe 2020), and diverse new national regulatory initiatives (Cussins 2020; Law Library of Congress 2019).

Much work to date has focused on the regulation of AI within particular domestic regulatory contexts. For instance, this has explored the relative institutional competencies of legislatures, regulatory agencies, or courts at regulating AI (Guihot, Matthew, and Suzor 2017; Scherer 2016). Others have emphasised the governance roles of various actors in the AI landscape (Leung 2019), exploring for instance how tech companies' ethics advisory committees (Newman 2020, 12–29), AI company employee activists and 'epistemic communities' (Belfield 2020; Maas 2019a), AI research community instruments (such as scientific conference research impact assessment mechanisms) (Prunkl et al. 2021), or private regulatory markets architectures (Clark and Hadfield 2019), could all help shape AI regulation.

There is also a growing recognition of the importance of global coordination or cooperation for AI governance (Feijóo et al. 2020; Kemp et al. 2019; Turner 2018). At the global level, much focus to date has been on the burgeoning constellation of AI ethics principles that has sprung up in the last half-decade (Fjeld et al. 2019; Jobin, Ienca, and Vayena 2019; Schiff et al. 2020; Stahl et al. 2021; Stix 2021). However, this is being increasingly complemented by regulatory proposals. These have ranged from relying on existing norms, treaty regimes, or institutions in public international law (Kunz and Ó hÉigeartaigh 2021; Burri 2017; Smith 2020). Others have proposed entirely new international organizations in order to coordinate national regulatory approaches (Erdelyi and Goldsmith 2018; Kemp et al. 2019; Turner 2018, chap. 6), or have compared such centralized institutions for AI to more decentralized or fragmented alternatives (Cihon, Maas, and Kemp 2020b, 2020a). Others have focused more on the role of soft law instruments (Gutierrez and Marchant 2021; Gutierrez, Marchant, and Tournas 2020), international standard-setting bodies (Cihon 2019; Lorenz 2020), or certification schemes (Cihon et al. 2021). Others have proposed the adaptation of existing informal governance institutions such as the G20 (Jelinek, Wallach, and Kerimi 2020).

This is clearly a diverse constellation of efforts. Yet there are underlying classes and patterns in these approaches to AI regulations.

For instance, one group of 'technology-centric' approaches focuses on 'AI' as an overarching class that should be considered in whole (Turner 2018). While this is more reflective of the cross-domain application and impact of many AI techniques, however, this approach is not without problems. For one, it is undercut by intractable debates over how to define 'AI' (Russell and Norvig 2016)—and by the fact that AI is not a single thing (Schuett 2019; Stone et al. 2016).

This shortfall is more addressed in the second 'technology-centric' approach, which is 'application-centric'—or, as some call it, 'use-case centric' (Schuett 2019, 5). This approach seeks to unpack the umbrella term 'AI', and split out the specific AI applications that regulation should focus on. In the past decade, many policy responses to AI have been sparked by one or another use case of AI technology—involving concrete issues that have been thrown up, or visceral problems that are anticipated—such as autonomous cars, drones, facial recognition, or social robots (Turner 2018, 218–19). As Petit puts it, this technology-centric approach to AI involves charting "legal issues from the bottom-up standpoint of each class of technological application" (Petit 2017).

Application-centric approaches remain the default response to AI governance. However, like the AI-centric approach, this orientation also has shortfalls. For one, it emphasizes visceral edge cases, and is therefore easily lured into regulating edge-case challenges or misuses of the technology (e.g. the use of DeepFakes for political propaganda) at a cost of addressing far more common but less visceral use cases (e.g. the use of DeepFakes for gendered harassment) (Liu et al. 2020). Moreover, the resulting AI policies and laws are frequently formulated in a piecemeal and ad-hoc fashion, which means that this perspective can promote siloed regulatory responses (Turner 2018, 218–19). A focus on individual applications also may inadvertently foregrounds technology-specific regulations even where these are not the most effective (Bennett Moses 2013). It moreover induces a 'problem-solving' orientation (Liu and Maas 2021), aimed at narrowly addressing local problems caused (or envisioned) by the specific use case that prompted the regulatory process (Crootof and Ard 2021).

A distinct set of regulatory responses to AI are instead law-centric. They invoke what Nicolas Petit has called a 'legalistic' approach, which 'consists in starting from the legal system, and proceed by drawing lists of legal fields or issues affected by AIs and robots." (Petit 2017, 2). This approach segments regulatory responses by departing from AI's impacts on- and within specific conventional legal codes or subjects (e.g. privacy law, contract law, the law of armed conflict), or by exploring the ways in which these create questions about the scope, intersection, assumptions, or adequacies of existing law (Crootof and Ard 2021).

To be clear, technology-, application-, and law-centric approaches to AI regulation have important insights, and must play a role in any AI governance ecosystem. Nonetheless, they have their drawbacks. Importantly, AI governance proposals could be grounded in a better understanding of how AI applications translate into cross-domain changes, and how future capabilities or developments might further shift this problem portfolio (Maas 2019b). An alternative is therefore to shift (or complement) these approaches with a framework that is not (solely) anchored in 'new technology' (whether on the umbrella term of 'AI', or on isolated AI applications), nor on isolated legal domains, but which rather examines types of change. What impacts are we concerned about?

AI regulatory ecosystem require a protocol for considering *where, and why AI change warrants regulatory intervention, and how and when this regulatory intervention should take place*. It should be able to adequately identify when AI applications create regulatory rationales, as well as the best levers to approach AI as a regulatory target.

Can we reformulate an impact-focused approach for AI regulation, that provides superior levers for regulation? To achieve this, this chapter instead draws on existing theories from the emerging paradigms of law, regulation and technology, and 'TechLaw'—"the study of how law and technology foster, restrict, and otherwise shape each other's evolution." (Crootof and Ard 2021, n. 1; Ard and Crootof 2020). In particular, it proposes to approach AI governance through the lens of 'sociotechnical change' (Bennett Moses 2007a, 2017). As such, this chapter will now turn to how this approach can bridge the gap between technology/application-centric and law-centric approaches, by guiding reflection on when and why new AI applications require new regulation—and how the resulting regulatory interventions are best tailored.

Reframing regulation: AI and sociotechnical change

It should be no surprise that changes in technology have given rise to extensive scholarship on the relation between law and these new technologies. In some cases, such work has focused in on identifying the assumed 'exceptional' nature or features of a given technology (Calo 2015). However, as noted, other scholars have influentially argued that it is less the 'newness' of a technology that brings about regulatory problems, but rather the ways it enables particular changes in societal practices, behaviour, or relations (Balkin 2015; Bennett Moses 2007b; Friedman 2001).

That is, in what ways do changes in a given technologies translate to new ways of carrying out old conduct, or create entirely new forms of conduct, entities, or new ways of being or connecting to others? When does this create problems for societies or their legal systems? Rather than focus on 'regulating technology' (either in general, or in specific applications), this scholarship accordingly puts a much greater emphasis on 'adjusting law and regulation for sociotechnical change' (Bennett Moses 2017, 574). In particular, Brownsword, Scotford, and Yeung have highlighted three dimensions of technological 'disruption': "legal disruption, regulatory disruption, and the challenge of constructing regulatory environments that are fit for purpose in light of technological disruption" (Brownsword, Scotford, and Yeung 2017, 7).

As noted, various scholars of law, regulation and technology have emphasized the importance of 'sociotechnical change'. In her 'theory of law and technological change' (Bennett Moses 2007b), Lyria Bennett Moses has argued that questions of 'law and technology' are rarely if ever directly about technological progress itself (whether incremental or far-reaching). Instead, she argues that lawyers and legal scholars who examine the regulation of technology are focused on questions of "how the law ought to relate to activities, entities, and relationships made possible by a new technology" (Bennett Moses 2007b, 591).

Indeed, Lyria Bennett-Moses has argued that from a regulatory perspective, "[t]echnology is rarely the only 'thing' that is regulated and the presence of technology or even new technology alone does not justify a call for new regulation"(Bennett Moses 2017, 575). In doing so, she calls for a shift in approach from 'regulating technology' to 'adjusting law and regulation for sociotechnical change.' (Bennett Moses 2017, 574) This shifts the focus on patterns of 'socio-technical change' (Bennett Moses 2007b, 591–92, 2007a)—instances where changes in certain technologies actually expand human capabilities in ways that give rise to new activities or forms of conduct, or new ways of being or of connecting to others (Bennett Moses 2007b, 591–92).

As such, the question of governing new technologies is articulated not with reference to a list of (sufficiently 'new') technologies (Bennett Moses 2017, 576), but is relatively 'technology-neutral'. It is this functional understanding of 'socio-technological change' that informs more fruitful analysis of when and why we require regulation for new technological or scientific progress. It can also underlie a more systematic examination of which developments in AI technology are relevant for a regulatory system to focus on.

AI as regulatory rationale

What types of sociotechnical changes (e.g. new possible behaviours or states of being) actually give rise to regulatory rationales? When a technology might create an opportunity for certain problematic behaviour, does that opportunity need to be acted upon, or can the mere possibility of that behaviour constitute a regulatory rationale? Can sociotechnical changes be anticipated? This entails a more granular understanding of the dynamics of sociotechnical change—and when or how such changes can constitute a rationale for regulation.

Varieties of sociotechnical change

When and why do AI capabilities rise to a problem that warrants legal or regulatory solutions? It is important to recognize that not all new scientific breakthroughs, new technological capabilities, or even new use cases will necessarily produce the sort of 'sociotechnical change' that requires regulatory responses.

In practical terms, this relates to the observations that new social (and therefore governance) opportunities or challenges are not created by the mere fact of a technology being conceived, or even prototyped, but rather by them being translated into new 'affordances' for some actors—relationships "between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used" (Norman 2013, 11). AI affordances can be new types of behaviour, entities, or relationships that were not previously possible (or easy), and which are now available to various actors (Liu et al. 2020).

How does technological change translate into sociotechnical change? When would this be disruptive to law? There are various types of sociotechnical changes that new AI applications can create or enable (Maas 2019c, 33; see also Crootof and Ard 2021).

(1) *Allowing older types of behaviours to be carried out with new items or entities*, including artefacts which are potentially not captured under existing (technology-specific) regulatory codes, or which blur the boundaries between existing domains or regimes, potentially causing problematic gaps, overlaps, or contradictions in how these behaviours are covered by regulation;
(2) *Absolute* or *categorical capability changes*, where AI progress expands the action space and 'unlocks' new capabilities or behaviour which were previously simply out of reach for anyone, and which could be of regulatory concern for one of various reasons;
(3) Relative capability changes, where AI increases the prominence of a previously rare behaviour, for instance because progress *lowers thresholds* or use preconditions for a certain capability (e.g. advanced video editing; online disinformation campaigns; cryptographic tools), which was previously reserved to a narrow set of actors; or because progress allows the *scaling up* of certain existing behaviours (e.g. phishing emails).
(4) *Positional changes* amongst actors, where AI applications that drive shifts in which particular state actors are dominant, while leaving the general 'rules' of that international system more or less unaltered.
(5) *Changing structural dynamics* in a (international) society, for instance, by;

i. Shifting prevalent influence between *types* of actors (e.g. away from states and towards non-state actors or private companies);
ii. Shifting the *means by* which certain actors seek to exercise 'influence' (e.g. from 'hard' military force to computational propaganda, or from multilateralism to 'lawfare', as a result of new communications technologies increasing the scope, velocity and effectiveness of such 'lawfare' efforts) (Dunlap 2008, 146–48);
iii. Altering the *norms or identities* of actors, and thereby changing the *terms* by which they conceive of their goals and orient their behaviour.

As mentioned, there may be certain AI innovations or breakthroughs that do not create very large sociotechnical changes of these forms, even if from a pure scientific or engineering standpoint they involve considerable alterations to the state of the art. Conversely, technological change or improvements also need not be qualitatively novel, dramatic, sudden, or cutting-edge for them to drive intense and meaningful change in balances of power, or in societal structures (Cummings et al. 2018, iv). The question is therefore not only how large these sociotechnical changes are, but how, or whether, they touch on the general rationales for new regulatory interventions.

Mapping regulatory rationales

There are various accounts for when and why regulatory intervention is warranted by the introduction of new technologies. For instance, Deryck Beyleveld and Roger Brownsword argue that emerging technologies generally give rise to two kinds of concerns: "one is that the application of a particular technology might present risks to human health and safety, or to the environment […] and the other is that the technology might be applied in ways that are harmful to moral interests" (Beyleveld and Brownsword 2012, 35).

However, while these may be the most prominent rationales, the full scope of reasons for regulation may extend further. In a non-technology context, Tony Prosser has argued that regulation, in general, has four grounds: "(1) regulation for economic efficiency and market choice, (2) regulation to protect rights, (3) regulation for social solidarity, and (4) regulation as deliberation" (Prosser 2010, 18).

How do these regulatory rationales relate to technological change? As Bennett Moses (2017, 578) notes, all four of these rationales can certainly become engaged by new technologies. That is, new technologies (or new applications) can:

(1) create sites for *new market failures*, warranting regulatory interventions such as technical standards or certification, to ensure economic efficiency and market choice, and remedy information inadequacies for consumers;
(2) generate many *new risks or harms*—either to human health or the environment, or to moral interests—which create a need for regulation to protect the rights of these parties (e.g. restrictions of new weapons; the ban on human cloning).

(3) create concern about *social solidarity*, as seen in concerns over the 'digital divide' at both a national and international level, creating a need for regulation to ensure adequate inclusion.
(4) create sites or pressures for the exertion of *proper democratic deliberation* over the design or development pathways of technologies. (Bennett Moses 2017, 579–83)

To be sure, as a technology-centric approach would note, these cases all involve new technologies which require regulation. However, Bennett Moses argues that in each of these cases, it is not the involvement of 'new technology' per se that provides a special rationale for regulation, above and beyond the resulting social changes (e.g. potential market failures; risks to rights; threats to solidarity; or democratic deficits) that are at stake (Bennett Moses 2017, 583). We are not worried about technology; we are worried about its *effects*.

As such, the primary regulatory concern is over the emergence of the 'sociotechnical' effects that occur. This conceptual shift can help address one limit that regulatory or governance strategies encounter if they focus too much or too narrowly on technology. As she argues:

> "… treating technology as the object of regulation can lead to undesirable technology specificity in the formulation of rules or regulatory regimes. If regulators ask how to regulate a specific technology, the result will be a regulatory regime targeting that particular technology. This can be inefficient because of the focus on a subset of a broader problem and the tendency towards obsolescence"

(Bennett Moses 2017, 584).

As such, taking a sociotechnical (rather than a technology-centric) approach, this lens helps keep into explicit focus specific *rationales* for governance in each use case of AI: on what grounds and when regulation is needed and justified?

These four accounts of rationales are valuable as a starting point for AI regulation. However, we can refine this account. For one, it is analytically valuable to draw a more granular distinction between (physical) harms to human health or the environment, and (moral) harms to moral interests (Beyleveld and Brownsword 2012).

Moreover, these categories all concern rationales for governance to step in, in response to sociotechnical changes that are affecting society (i.e. the regulatees) directly. However, there may also be cases where AI-enabled sociotechnical change creates an indirect regulatory rationale, because it presents some risk directly to the existing legal order charged with mitigating the prior risks. In such cases of 'legal disruption' (Liu et al. 2020; Maas 2019c), sociotechnical change can produce a threat to the regulatory ecosystem itself. This can be because these tools allow regulatees to more effectively challenge or bypass existing laws, resulting in potential 'legal destruction' (Maas 2019c). Alternatively, it can result because certain AI tools can drive 'legal displacement' (Maas 2019c), by offering substitutes or complements to existing legal instruments, in shaping or managing the behaviour of citizens (Brownsword 2019).

Drawing together the above accounts, one might then speak of a regulatory rationale for an AI system or application, whenever it drives sociotechnical changes (new

ways of carrying out old behaviour, or new behaviours, relations or entities) which result in one or more of the following situations:

(1) new possible market failures;
(2) new risks to human health or safety, or the environment;
(3) new risks to moral interests, rights, or values;
(4) new threats to social solidarity;
(5) new threats to democratic process;
(6) new threats directly to the coherence, efficacy or integrity of the existing regulatory ecosystem charged with mitigating the prior risks (1-5).

All this is not to say that these rationales apply in the same way in all specific contexts. Indeed, they will be weighted differently across distinct legal systems and jurisdictions—and between domestic and international law. Nonetheless, they provide a rubric for understanding when or why we (should) want to regulate a new AI application—and a reminder that it is the sociotechnical changes, not the appearance of new technology in itself, that we are concerned about.

AI as regulatory target

Along with providing a greater grounding for understanding whether, when and why to regulate new AI applications, a consideration of sociotechnical change can also shed light on the regulatory 'texture' of the underlying AI capabilities—that is, its constitution as a 'regulatory target' (Buiten 2019, 46–48).

That is, once regulators have confirmed a regulatory rationale (i.e. they have asked 'do we need regulation? For what sociotechnical change? What regulatory rationale?'), they then face the question of how to craft regulatory actions. In considering AI applications as a *target* for regulation, a sociotechnical change-centric perspective must on the one hand take stock of the material aspects of a technology (as an artefact).

Material features certainly matter from the perspective of understanding key parameters for regulation, such as:

(1) *Its trajectory and distribution*: i.e. the state of leading AI capabilities (across its different sub-fields), possible and plausible rates and directions of progress given material constraints on the design space and process (Verbruggen 2020); preconditions for acquisitions and use; and factors driving or inhibiting the technology's proliferation to various (types of) actors globally (Horowitz 2018);
(2) *Its material 'risk profile'*: how different AI paradigms or techniques can at times be associated with rather distinct types of ethics or safety issues (Hernandez-Orallo et al. 2020);
(3) *The political viability of regulation, given the 'regulation-tolerance/resistance' profile*: how certain features of the technology (e.g. the viscerality of weapons)

affect stakeholder perceptions of the imminence of various applications of the technology, and of the need for urgent regulation (Crootof 2019; Watts 2015)

(4) *Potential sites or vectors for regulatory leverage*: for instance, the degree to which proliferation of certain systems could be meaningfully halted through export control policies (Brundage et al. 2020; Fischer et al. 2021).

In these ways, such material features certainly matter, especially when considering AI regulation at the global level. For instance, scholars have argued that the modern global digital economy, far from consisting solely of ethereal digital products ungraspable by law, is instead populated by distinct 'regulatory objects', which vary in their degree of apparent 'materiality' (from high-capital submarine cables and satellite launch facilities, to ethereal cloud services), and their degree of centralization (from diverse suppliers of various 'smart' appliances, to dominant social networks or computationally intensive search engine algorithms) (Beaumier et al. 2020). Critically, some of these may not easily be subjected to global regulation, but many which can certainly be captured by various regulatory approaches.

However, for regulatory purposes, a material analysis is not sufficient. A sociotechnical change-centric perspective on AI regulation rather can and should go beyond the technology itself, and consider a broader set of 'problem logics' in play.

For instance, we can fruitfully distinguish between: 'ethical challenges', 'security threats', 'safety risks', 'structural shifts', 'common benefits', and 'governance disruption'. Distinguishing amongst these ideal-types is valuable, as these clusters can introduce distinct problem logics, and foreground distinct regulatory logics or levers (see Table 1).[1]

It is important to note that this taxonomy is not meant to be mutually exclusive, nor exhaustive. It aims to capture certain regularities which help ask productive regulatory questions. For each category, we can ask—how does the AI capability produce sociotechnical change? Why does this create a governance *rationale*? How should this be approached as governance *target*? What are the barriers, and what regulatory tools are foregrounded?

There is insufficient space to go into exhaustive detail on each of these categories within the taxonomy. However, at glance, we can pick out a number of ways in which clustering AI's sociotechnical impacts along these various problem logics, can facilitate structural regulation-relevant insights for AI regulators. These include consideration of aspects.

For one, this model enables examination of the underlying ***origins*** of the sociotechnical challenge of concern, in terms of: (1) the ***key actors*** (e.g. principals, operators, malicious users) whose newly AI-enabled or -related behaviour or decisions create the governance concerns, and (2) those *actors'* ***traits, interests, or motives*** which drive the AI-sociotechnical-problem related behaviour or decisions (e.g. actor apathy, malice, negligence, or the way the new capability sculpts choice architectures in ways shift structural incentives or strategic pressures).

Problem Logic and *questions*	**Corresponding governance rationales**	**Examples in AI** (*selected*)	**Regulatory Surface** (**O**rigin; **C**ontributing **F**actors; **B**arriers to **R**egulation)	**Regulatory Approaches** (*selected*)
Ethical challenges *What rights, values or interests does this threaten?*	• New risks to moral interests, rights or values • New threats to social solidarity • Threats to democratic process	• *Justice*: bias; explainability… • *Power*: facial recognition… • *Democracy*: AI propaganda… • *Freedom*: 'Code as Law'; 'algocracy'…	• O. Developer / user ***apathy*** (to certain implicated values) • BR. Underlying ***societal disagreement*** (culturally and over time) over how to weigh the values, interests or rights at stake	• *Product-focused:* Bans ('mend—or end'); 'machine ethics' • *Ex ante producer-focused:* oversight mechanisms; end-to-end auditing; ethics education for engineers; 'Value-Sensitive Design' • *Ex post principal*-focused: & accountability mechanisms
Security threats *How is this vulnerable to misuse or attack?*	• New risks to moral interests, rights or values • New risks to human health or safety	• *AI as tool*: DeepFakes; • *AI as attack surface*: adversarial input • *AI as shield*: fraudulent trading agents; UAV smuggling	• O. Attacker ***malice*** (various motives) • CF. Target ***apathy*** • CF. 'Offense-defense balance' of AI knowledge • BR. Target's intrinsic vulnerability (e.g. of human practices to automated social engineering attacks).	• *Perpetrator-focused*: change norms, prevent access; improve detection & forensics capabilities to ensure attribution and deterrence • *Target-focused*: reduce exposure; red-teaming; 'security mindset'
Safety risks *Can we rely on- and control this?*	• New risks to human health or safety	• *Unpredictability and opacity* • Environmental interactions • Automation bias and 'normal accidents' • 'Value misalignment'	• O. Actor ***negligence*** • CF. Behavioural features of AI systems (opacity; unpredictability; optimisation failures; specification gaming) • CF. Human overtrust and automation bias • BR. 'Many hands' problem—long and discrete supply chains	• Relinquishment (of usage in extreme-risk domains) • 'Meaningful Human Control' (various forms) • Safety engineering (e.g. reliability; corrigibility; interpretability; limiting capability or deployment; formal verification) • Liability mechanisms & tort law; open development
Structural shifts *How does this shape our decisions?*	• (all, indirectly)	• *Change calculations*: LAWS lower costs of conflict • *Increased scope for mis-calculation*: e.g. attack prediction systems	• O. Systemic ***incentives*** for actors (alters choice architectures; increases uncertainty & complexity; competitive value erosion) • BR: collective action problems	• Arms control (mutual restraint) • Confidence-Building Measures (increase trust or transparency)
Public Goods *How can we realize good opportunities with this?*	• Possible market failures	• Gains from AI interoperability • 'AI for global good' initiatives • Distributing benefits of AI	• O. Systemic ***incentives*** for various actors • BR. Overcoming loss aversion; coordination challenges re. cost-sharing, free-riding); political economy factors	• (Global) standards • 'Public interest' regulation and subsidies • 'Windfall clause' & redistributive guarantees
Governance Disruption *How does this change how we regulate?*	• New risks directly to existing regulatory order	• AI systems creating *substantive* ambiguity in law • Legal automation altering *processes* of law • Erodes political *foundations*	• O. Push towards legal efficiency • CF. Legal system ***exposure*** and dependence on conceptual orders or operational assumptions	• Provisions to render governance 'innovation-proof': technological neutrality; authoritative interpreters, sunset clauses; etc. … • Oversight for legal automation; distribution

Table 1. Taxonomy of AI problem logics

This in turn can be linked to the ***contributing factors*** which sustain or exacerbate this sociotechnical impact, such as:

(1) The range and diversity of AI-related failure modes and issue groups (Hernandez-Orallo et al. 2020, 7), including emergent interactions with other actors (human or algorithmic) in their environment (Rahwan et al. 2019), peculiar behavioural failure modes (Amodei et al. 2016; Krakovna et al. 2020; Kumar et al. 2019; Leike et al. 2017). (**safety risks**)

(2) Human overtrust and automation bias, rendering some AI systems susceptible to emergent and cascading 'normal accidents' (Maas 2018) (**safety risks**)

(3) The underlying 'offense-defense balance' of AI scientific research (Shevlane and Dafoe 2020a), and how it evolves along with more sophisticated AI capabilities (Garfinkel and Dafoe 2019) (**security threats**).

(4) The susceptibility of existing legal and regulatory systems themselves to 'disruption' by AI uses, at the level of doctrinal *substance*, law-making and enforcement *processes*, or the *political* foundations (Liu et al. 2020; Maas 2019c) (**governance disruption**).

Moreover, this model enables a study of the *barriers to regulation;* that is, the factors that drive the difficulty of formulating or implementing policy solutions, and which will themselves have to be overcome to achieve effective regulatory responses for AI, because of:

(5) the live societal or cross-cultural value pluralism (Gabriel 2020) or disagreement over the values, interests or rights affected, and how these should be weighted in the context of a specific contested AI application (**ethical challenges**);

(6) the disproportionately high costs of 'patching' vulnerabilities of human social systems (e.g. our faith in the fidelity of human voices) against AI-enabled social engineering attacks, relative to past costs of patching 'conventional' cybersecurity vulnerabilities by the dissemination of software fixes (Shevlane and Dafoe 2020b, 177). (**security threats**)

(7) the difficulty of foreseeing indirect effects of AI on the structure of different actors' choice architectures (van der Loeff et al. 2019; Zwetsloot and Dafoe 2019)—and the difficulty of resolving those situations through any one actor's unilateral action (Dafoe 2020), or to coordinate behaviour in response (**structural shifts**)

Finally, on the basis of the above, it allows a consideration of the types of *regulatory approaches* and levers that are highlighted and foregrounded for each of these challenges. It highlights the role of 'mend-it-or-end-it' debates around algorithmic accountability (Pasquale 2019), auditing frameworks (Raji et al. 2020), and underlying cross-cultural cooperation (ÓhÉigeartaigh et al. 2020) to diverse *ethical challenges*. Of perpetrator-focused and target-focused (e.g. 'security mindset' (Severance 2016)) interventions to shield against AI *security threats*. The development of support programs to guarantee *public goods* such as the use of AI in 'AI for Good' interventions (Floridi et al. 2018; ITU 2019), humanitarian uses (Roff 2018, 25; but see Sapignoli 2021), or redistributive

guarantees such as a 'Windfall clause' that sees tech companies pledge extreme future profits above a certain threshold towards redistribution (O'Keefe et al. 2020).

That is not to say that this framework provides conclusive recipes or roadmaps for regulation. Rather, it provides a beginning structuring framework for thinking through common challenges across diverse regulatory domains charged with resolving questions around seemingly separate applications of AI (Crootof and Ard 2021). Such an approach can at least avoid duplication of effort, and at best can support the formulation and spread of better, more resilient policies.

In sum, a sociotechnical-change-centric approach is not without its pitfalls or limits. Still, it can have various benefits in organizing and orienting an AI regulatory ecosystem. It prompts regulators to ask themselves: (1) when, why and how a given AI application produces particular types of sociotechnical changes; (2) When and why these changes rise to create a rationale for governance; (3) How to approach the target of regulation. As such, this can be an important regulatory complement to the insights provided by—and the interventions grounded in—technology-centric or law-centric perspectives.

Implementation: AI regulation through a sociotechnical lens

This lens of sociotechnical change does not provide single substantive answers for how to resolve each and every AI policy problem. However, it can help answer common recurring questions in AI policy around institutional choice and regulatory timing and design (Bennett Moses 2017, 585–91). In particular, regulatory actors can improve governance for AI challenges in terms of regulatory *triage, tailoring, timing and responsiveness, and design*.

Regulatory triage

In the first place, the sociotechnical-change-centric perspective on AI can help in carrying out regulatory *triage.* This is not just of value to AI regulation: indeed, it fits in with a broader initiative in recent legal scholarship towards exploring questions of 'legal prioritization' (Winter et al. 2021). However, within the AI regulation ecosystem, this lens helps focus attention on the most societally disruptive impacts of the technology, and as such helps re-focus scarce regulatory attention. This reduces the risk that regulatory attention is over-allocated on visceral applications of AI which may not ultimately prove scalable, or on 'legally interesting puzzles', at a cost of more opaque but prevalent indirect impacts. Better triage can be a valuable corrective to approaches that select, organize or prioritize AI policy issues based on high-profile but non-representative incidents, popular-cultural resonance, or 'fit' to pre-existing legal domains.

Moreover, if regulatory bodies focus less on the 'newness' of AI technology, or on the steady stream of each new AI application, but rather on which downstream sociotechnical impacts in fact create particular regulatory rationales, they can step back from a reactive firefighting mode, and help defuse or dissolve the so-called 'pacing problem' (Marchant 2011). Regulatory triage is also aided by the ways in which this

framework can *expand regulatory actors' scope of analysis* of which sociotechnical impacts are relevant for regulatory consideration. While technology-centric approaches can highlight the direct challenges of AI (in the areas of *ethics, security,* and *safety*), the sociotechnical-change-centric perspective also allows regulators to consider interventions for various indirect sociotechnical changes, including the ways AI systems can *shift incentive structures*, how to *realize beneficial opportunities* and public goods around AI technology, or how AI applications can *disrupt* the regulatory tools or systems which these regulators would rely on.

What does that entail in practice? Improving triage around AI regulation could involve (1) improving information infrastructures or 'technical observatories' (Clark [forthcoming] (in this volume)), to not only equip regulators with relevant and up-to-date technical information around AI techniques and applications, but also think through how these relate to downstream sociotechnical impacts.

This may help ensure regulators are less easily dazzled by the 'newness' of new AI applications themselves (Mandel 2017), and enable them to become more aware of how different analogies can highlight different regulatory narratives in potentially counterproductive ways (Crootof and Ard 2021). This can also involve (2) setting up a cross-ecosystem agency to focus on 'legal foresighting' (Laurie, Harmon, and Arzuaga 2012), and forecasting methodologies (Avin 2019; Ballard and Calo 2019) aimed at eliciting AI's technologies' disparate sociotechnical impacts, link these to potential and actual regulatory rationales, and study the shifting material textures and problem logics around that application. In particular, this can support more democratic and inclusive stakeholder debate over the choices affected parties would seek to make around the deployment of potentially disruptive AI breakthrough capabilities (Cremer and Whittlestone 2021).

Regulatory tailoring and scope

Secondly, and relatedly, the sociotechnical change-centric lens helps in *tailoring* regulatory solutions to effective clusters of AI techniques, applications, -users, and societal effects. Rather than consign regulators to confront self-similar AI challenges (e.g. around meaningful human control; susceptibility to adversarial attack; unaccountable opacity in algorithmic decision-making) many times across individual legal domains, (Crootof and Ard 2021, 1), this approach highlights common themes, underlying material value chains, or usage problem logics of AI, as they are expressed in these various domains.

Practically, improved regulatory tailoring can require (3) the establishment of various institutions and meta-regulatory oversight mechanisms—connoting "activities occurring in a wider regulatory space, under the auspices of a variety of institutions, including the state, the private sector and public interest groups [which] may operate in concert or independently" (Grabosky 2017, 150). In so doing, such mechanisms could foster improved cross-regime dialogue of AI policy (Cihon, Maas, and Kemp 2020b). This can support regulatory harmonization or the bundling of regulatory interventions for various AI applications where appropriate. It can also examine how and where different regimes and institutions can exploit the same regulatory levers (e.g. compute

hardware production) that intersect on the AI development value chain. (4) Establishing mechanisms and fora for dialogue amongst various actors in the AI space that may have a hand in shaping the overarching 'problem logic'—in terms of the problem's origins, contributing factors, or barriers to regulation. The aim of such discussions would ideally be to reconfigure some of these wider conditions to be more supportive or conducive to AI regulation, or—where they are not very tractable, to explore alternative levers or vectors for regulation. Finally, it can promote the exchange of best practices or lessons learned around how regulators can address some of the problem logics that generate the problem or impede regulation.

Regulatory timing

Thirdly, in terms of regulatory *timing and responsiveness*, a study of AI's sociotechnical changes highlights the inadequacies of governance strategies that are grounded either in an attempt to predict sociotechnical changes in detail, or reactive responses which prefer to 'wait out' technological change until its societal impact has become clear—which demands a threshold of clarity that is in fact rarely achieved, even decades after a technology's deployment (Horowitz 2020). Rather, it emphasises the importance of anticipatory and adaptive regulatory approaches (Maas 2019b). This helps mitigate some of the information problems facing AI regulation, by helping ensure AI regulation can remain adaptive and 'scalable' to ongoing sociotechnical change, given the profound lack of information about future pathways. This could be pursued by (5) incorporating provisions such as sunset clauses that prompt re-examination (at the domestic level) or designating authoritative interpreters (at the international level).

Regulatory design

Fourthly, in terms of *regulatory design*, the sociotechnical change lens highlights when and why governance should prefer technology-neutral rules versus technology-specific rules. By considering the specific regulatory or governance rationale in play, we may understand when or whether technological neutrality is to be preferred. Generally, Bennett-Moses (2017, 586) argues that "regulatory regimes should be technology- neutral to the extent that the regulatory rationale is similarly neutral".

In this view, the point is not, to find a regulatory strategy that already details long lists of anticipated future applications of AI. The idea is rather to develop institutional mechanisms that are up to the task of managing distinct problem logics—new ethical challenges, security threats, safety risks, structural shifts, opportunities for benefit, or governance disruptions—in a way that can be relatively transferable across- or agnostic to the specific AI techniques used to achieve those affects. Establishing clearer guidelines about formulation of AI-specific regulations, and the circumstances in which these should rely on standards or rules, and when they should be tech-specific or tech-neutral (Crootof and Ard 2021).

Conclusion

This chapter has introduced, articulated, and evaluated a 'sociotechnical change-centric' perspective on aligning AI regulation.

It first briefly sketched the general value of a 'change-centric' approach to the problems facing the AI governance ecosystem. The chapter next articulated a framework focused on Lyria Bennett Moses's account of regulation for sociotechnical change. It explored how, when, and why law and regulation for AI ought to tailor themselves to broad *sociotechnical change* rather than local *technological change*. It applied this model to AI technology, in order to show how this model allows a better connection of AI applications to five types of sociotechnical change—and how these in turn can be mapped to six types of regulatory *rationales*.

It then turned to the mirror question of how, having established a need for governance, regulators might craft policy interventions to the particular regulatory target of AI. This involved a consideration of both the material textures of AI applications, but especially demands focus on the 'problem logics' involved. It argued that socio-technical changes created by AI applications can be disambiguated into six specific types of challenges—ethical challenges, security threats, safety risks, structural shifts, public goods, and governance disruption—which come with distinct problem features (origins, contributory factors, barriers to regulation), and which may each be susceptible to (or demand) different governance responses.

Finally, the chapter concluded by reflecting on the limits and uses of this approach, before sketching some indicative institutional and regulatory actions that might draw on this framework to improve regulatory *triage, tailoring, timing & responsiveness,* and *regulatory design* of AI policy.

To be clear, an emphasis on sociotechnical change is not a new insight in scholarship on law, regulation and new technology. However, in a fragmented and incipient AI governance landscape, it remains a valuable tool. In sum, 'sociotechnical change' should be considered not a new or substitute paradigm for AI governance, but rather a complementary perspective. Such a lens is subject to its own conditions and limits, but when used cautiously, can offer regulators a more considered understanding of which of AI's challenges are possible, plausible, or already-pervasive—and how these might be best met.

Acknowledgements

For valuable comments and feedback on early drafts, I thank Jess Whittlestone, Jack Clark, and Harry Surden. I also thank Christina Korsgaard for her support throughout the writing process. Any remaining errors are all my own.

Notes

[1] An earlier version of this framework is presented and unpacked in further detail in (Maas 2020, 166–86). Note, this version referred to the 'public goods' as the 'common goods' problem logic, instead.

References

Amodei, Dario et al. 2016. "Concrete Problems in AI Safety." *arXiv:1606.06565 [cs]*. http://arxiv.org/abs/1606.06565 (May 13, 2017).

Ard, BJ, and Rebecca Crootof. 2020. "The Case for 'Technology Law.'" *Nebraska Governance & Technology Center*. https://ngtc.unl.edu/blog/case-for-technology-law (March 16, 2021).

Avin, Shahar. 2019. "Exploring Artificial Intelligence Futures." *AIHumanities*. http://aihumanities.org/en/journals/journal-of-aih-list/?board_name=Enjournal&order_by=fn_pid&order_type=desc&list_type=list&vid=15 (February 25, 2019).

Balkin, Jack M. 2015. "The Path of Robotics Law." *California Law Review Circuit* 6: 17.

Ballard, Stephanie, and Ryan Calo. 2019. "Taking Futures Seriously: Forecasting as Method in Robotics Law and Policy." In Miami, 22. https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/Calo_Taking-Futures-Seriously.pdf.

Beaumier, Guillaume et al. 2020. "Global Regulations for a Digital Economy: Between New and Old Challenges." *Global Policy* 11(4): 515–22.

Belfield, Haydn. 2020. "Activism by the AI Community: Analysing Recent Achievements and Future Prospects." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York NY USA: ACM, 15–21. http://dl.acm.org/doi/10.1145/3375627.3375814 (February 12, 2020).

Bennett Moses, Lyria. 2007a. "Recurring Dilemmas: The Law's Race to Keep Up With Technological Change." *University of New South Wales Faculty of Law Research Series* 21. http://www.austlii.edu.au/au/journals/UNSWLRS/2007/21.html (July 3, 2018).

———. 2007b. "Why Have a Theory of Law and Technological Change?" *Minnesota Journal of Law, Science & Technology* 8(2): 589-606.

———. 2013. "How to Think about Law, Regulation and Technology: Problems with 'Technology' as a Regulatory Target." *Law, Innovation and Technology* 5(1): 1–20.

———. 2017. "Regulating in the Face of Sociotechnical Change." In *The Oxford Handbook of Law, Regulation, and Technology*, eds. Roger Brownsword, Eloise Scotford, and Karen Yeung. , 573–96. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49 (May 13, 2017).

Beyleveld, Deryck, and Roger Brownsword. 2012. "Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning." *Law, Innovation & Technology* 4(1): 35–65.

Brownsword, Roger. 2019. "Law Disrupted, Law Re-Imagined, Law Re-Invented." *Technology and Regulation*: 10–30.

Brownsword, Roger, Eloise Scotford, and Karen Yeung. 2017. "Law, Regulation, and Technology: The Field, Frame, and Focal Questions." In *The Oxford Handbook of Law, Regulation and Technology*, eds. Roger Brownsword, Eloise Scotford, and Karen Yeung. Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-1 (January 3, 2019).

Brundage, Miles et al. 2020. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." *arXiv:2004.07213 [cs]*. http://arxiv.org/abs/2004.07213 (April 16, 2020).

Buiten, Miriam C. 2019. "Towards Intelligent Regulation of Artificial Intelligence." *European Journal of Risk Regulation* 10(1): 41–59.

Burri, Thomas. 2017. "International Law and Artificial Intelligence." *German Yearbook of International Law* 60: 91–108.

Calo, Ryan. 2015. "Robotics and the Lessons of Cyberlaw." *California Law Review* 103: 513–64.

Cihon, Peter. 2019. *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Oxford: Center for the Governance of AI, Future of Humanity Institute, University of Oxford. Technical Report. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf (April 18, 2019).

Cihon, Peter, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. 2021. "AI CERTIFICATION: Advancing Ethical Practice by Reducing Information Asymmetries." *IEEE Transactions on Technology and Society*: 1–1.

Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 2020a. "Fragmentation and the Future: Investigating Architectures for International AI Governance." *Global Policy* 11(5): 545–56.

———. 2020b. "Should Artificial Intelligence Governance Be Centralised?: Design Lessons from History." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York NY USA: ACM, 228–34. http://dl.acm.org/doi/10.1145/3375627.3375857 (February 12, 2020).

Clark, Jack. [forthcoming]. "Technical Observatories for Better AI Governance." In *The Oxford Handbook of AI Governance*, eds. Valerie Hudson and Justin Bullock. Oxford Univ. Press.

Clark, Jack, and Gillian K Hadfield. 2019. "Regulatory Markets for AI Safety." In , 9.

Cremer, Carla Zoe, and Jess Whittlestone. 2021. "Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI." *International Journal of Interactive Multimedia and Artificial Intelligence* 6(5): 100–109.

Crootof, Rebecca. 2019. "Jurisprudential Space Junk: Treaties and New Technologies." In *Resolving Conflicts in the Law*, eds. Chiara Giorgetti and Natalie Klein. , 106–29. https://brill.com/view/book/edcoll/9789004316539/BP000015.xml (March 15, 2019).

Crootof, Rebecca, and B. J. Ard. 2021. "Structuring Techlaw." *Harvard Journal of Law & Technology* 34. https://papers.ssrn.com/abstract=3664124 (August 28, 2020).

Cummings, Mary L. et al. 2018. *Artificial Intelligence and International Affairs: Disruption Anticipated*. Chatham House. https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Cussins, Jessica. 2020. "National and International AI Strategies." *Future of Life Institute*. https://futureoflife.org/national-international-ai-strategies/ (June 22, 2020).

Dafoe, Allan. 2020. "AI Governance: Opportunity and Theory of Impact." https://www.allandafoe.com/opportunity (September 20, 2020).

Dunlap, Charles. 2008. "Lawfare Today: A Perspective." *Yale Journal of International Affairs*: 146–54.

Erdelyi, Olivia J, 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*, , 95–101. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_13.pdf.

Feijóo, Claudio et al. 2020. "Harnessing Artificial Intelligence (AI) to Increase Wellbeing for All: The Case for a New Technology Diplomacy." *Telecommunications Policy*: 101988.

Fischer, Sophie-Charlotte et al. 2021. *AI Policy Levers: A Review of the U.S. Government's Tools to Shape AI Research, Development, and Deployment*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment-%E2%80%93-Fischer-et-al.pdf (March 26, 2021).

Fjeld, Jessica et al. 2019. *Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches*. Berkman Klein Center for Internet & Society at Harvard University. https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf.

Floridi, Luciano et al. 2018. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28(4): 689–707.

Friedman, David D. 2001. "Does Technology Require New Law?" *Public Policy* 71: 16.

Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30(3): 411–37.

Garfinkel, Ben, and Allan Dafoe. 2019. "How Does the Offense-Defense Balance Scale?" *Journal of Strategic Studies* 42(6): 736–63.

Grabosky, Peter. 2017. "Meta-Regulation." In *Regulatory Theory*, Foundations and applications, ed. Peter Drahos. ANU Press, 149–62. https://www.jstor.org/stable/j.ctt1q1crtm.17 (March 4, 2021).

Gruetzemacher, Ross, and Jess Whittlestone. 2020. "The Transformative Potential of Artificial Intelligence." *Communications of the ACM*. https://arxiv.org/abs/1912.00747.

Guihot, Michael, Anne F. Matthew, and Nicolas Suzor. 2017. "Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence." *Vanderbilt Journal of Entertainment & Technology Law*. https://papers.ssrn.com/abstract=3017004 (July 2, 2018).

Gutierrez, Carlos Ignacio, and Gary Marchant. 2021. *A Global Perspective of Soft Law Programs for the Governance of Artificial Intelligence*. Sandra Day O'Connor College of Law, Arizona State University. https://lsi.asulaw.org/softlaw/the-report/.

Gutierrez, Carlos Ignacio, Gary E. Marchant, and Lucille Tournas. 2020. "Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance." *JURIMETRICS* 61(1). https://papers.ssrn.com/abstract=3775271 (March 7, 2021).

Hernandez-Orallo, Jose et al. 2020. "AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues." In *European Conference on Artificial Intelligence*, , 8. https://ecai2020.eu/papers/1364_paper.pdf.

Horowitz, Michael C. 2018. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review*. https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/ (May 17, 2018).

———. 2020. "Do Emerging Military Technologies Matter for International Politics?" *Annual Review of Political Science* 23(1): 385–400.

ITU. 2019. "AI for Good Global Summit 2019 Insights." https://itu.foleon.com/itu/aiforgood2019/home/ (October 6, 2019).

Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 2020. "Policy Brief: The Creation of a G20 Coordinating Committee for the Governance of Artificial Intelligence." *AI and Ethics*. https://doi.org/10.1007/s43681-020-00019-y (October 30, 2020).

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence*: 1–11.

Kemp, Luke et al. 2019. "UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance." https://digitalcooperation.org/wp-content/uploads/2019/02/Luke_Kemp_Submission-to-the-UN-High-Level-Panel-on-Digital-Cooperation-2019-Kemp-et-al.pdf.

Krakovna, Victoria et al. 2020. "Specification gaming: the flip side of AI ingenuity." *Deepmind*. https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity (May 12, 2020).

Kumar, Ram Shankar Siva et al. 2019. "Failure Modes in Machine Learning Systems." *arXiv:1911.11034 [cs, stat]*. http://arxiv.org/abs/1911.11034 (January 7, 2020).

Kunz, Martina, and Seán Ó hÉigeartaigh. 2021. "Artificial Intelligence and Robotization." In *Oxford Handbook on the International Law of Global Security*, eds. Robin Geiss and Nils Melzer. Oxford University Press. https://papers.ssrn.com/abstract=3310421 (January 30, 2019).

Laurie, Graeme, Shawn H. E. Harmon, and Fabiana Arzuaga. 2012. "Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty." *Law, Innovation & Technology* 4(1): 1–33.

Law Library of Congress. 2019. *Regulation of Artificial Intelligence in Selected Jurisdictions*. The Law Library of Congress. https://www.loc.gov/law/help/artificial-intelligence/regulation-artificial-intelligence.pdf.

Leike, Jan et al. 2017. "AI Safety Gridworlds." *arXiv:1711.09883 [cs]*. http://arxiv.org/abs/1711.09883 (December 1, 2017).

Leung, Jade. 2019. "Who Will Govern Artificial Intelligence? Learning from the History of Strategic Politics in Emerging Technologies." University of Oxford. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665.

Liu, Hin-Yan et al. 2020. "Artificial Intelligence and Legal Disruption: A New Model for Analysis." *Law, Innovation and Technology* 12(2): 205–58.

Liu, Hin-Yan, and Matthijs M. Maas. 2021. "'Solving for X?' Towards a Problem-Finding Framework to Ground Long-Term Governance Strategies for Artificial Intelligence." *Futures* 126: 22.

van der Loeff, Agnes Schim, Iggy Bassi, Sachin Kapila, and Jevgenij Gamper. 2019. "AI Ethics for Systemic Issues: A Structural Approach." In Vancouver, Canada. http://arxiv.org/abs/1911.03216 (January 13, 2020).

Lorenz, Philippe. 2020. *AI Governance through Political Fora and Standards Developing Organizations*. Stiftung Neue Verantwortung. https://www.stiftung-nv.de/de/publikation/ai-governance-through-political-fora-and-standards-developing-organizations.

Maas, Matthijs M. 2018. "Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, New York, NY, USA: Association for Computing Machinery, 223–28. https://doi.org/10.1145/3278721.3278766 (September 8, 2020).

———. 2019a. "How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons." *Contemporary Security Policy* 40(3): 285–311.

———. 2019b. "Innovation-Proof Governance for Military AI? How I Learned to Stop Worrying and Love the Bot." *Journal of International Humanitarian Legal Studies* 10(1): 129–57.

———. 2019c. "International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order." *Melbourne Journal of International Law* 20(1): 29–56.

———. 2020. "Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks." University of Copenhagen. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3833395.

Mandel, Gregory N. 2017. "Legal Evolution in Response to Technological Change." *The Oxford Handbook of Law, Regulation and Technology*. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-45 (September 26, 2018).

Marchant, Gary E. 2011. "The Growing Gap Between Emerging Technologies and the Law." In *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*, The International Library of Ethics, Law and Technology, eds. Gary E. Marchant, Braden R. Allenby, and Joseph R. Herkert. Dordrecht: Springer Netherlands, 19–33. https://doi.org/10.1007/978-94-007-1356-7_2 (January 28, 2019).

Newman, Jessica Cussins. 2020. *Decision Points in AI Governance*. Berkeley, CA: Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf (September 3, 2020).

Norman, Donald A. 2013. *The Design of Everyday Things*. Revised and expanded edition. New York, New York: Basic Books.

ÓhÉigeartaigh, Seán S. et al. 2020. "Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance." *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00402-x (May 17, 2020).

O'Keefe, Cullen et al. 2020. "The Windfall Clause: Distributing the Benefits of AI for the Common Good." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York NY USA: ACM, 327–31. http://dl.acm.org/doi/10.1145/3375627.3375842 (February 12, 2020).

Pasquale, Frank. 2019. "The Second Wave of Algorithmic Accountability." *LPE Project*. https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/ (September 3, 2020).

Petit, Nicolas. 2017. *Law and Regulation of Artificial Intelligence and Robots - Conceptual Framework and Normative Implications*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper. https://papers.ssrn.com/abstract=2931339 (May 11, 2020).

Prosser, Tony. 2010. *The Regulatory Enterprise: Government, Regulation, and Legitimacy*. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199579839.001.0001/acprof-9780199579839 (June 23, 2020).

Prunkl, Carina E. A. et al. 2021. "Institutionalizing Ethics in AI through Broader Impact Requirements." *Nature Machine Intelligence* 3(2): 104–10.

Rahwan, Iyad et al. 2019. "Machine Behaviour." *Nature* 568(7753): 477.

Raji, Inioluwa Deborah et al. 2020. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, Barcelona,

Spain: Association for Computing Machinery, 33–44. https://doi.org/10.1145/3351095.3372873 (February 23, 2020).

Roff, Heather M. 2018. "Advancing Human Security Through Artificial Intelligence." In *Artificial Intelligence and International Affairs: Disruption Anticipated*, Chatham House. https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.

Russell, Stuart, and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River: Pearson.

Sapignoli, Maria. 2021. "The Mismeasure of the Human: Big Data and the 'AI Turn' in Global Governance." *Anthropology Today* 37(1): 4–8.

Scherer, Matthew U. 2016. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law & Technology* (2). http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf (March 5, 2018).

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. "What's Next for AI Ethics, Policy, and Governance? A Global Overview." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York NY USA: ACM, 153–58. http://dl.acm.org/doi/10.1145/3375627.3375804 (February 12, 2020).

Schuett, Jonas. 2019. "A Legal Definition of AI." *arXiv:1909.01095 [cs]*. http://arxiv.org/abs/1909.01095 (January 6, 2020).

Severance, C. 2016. "Bruce Schneier: The Security Mindset." *Computer* 49(2): 7–8.

Shevlane, Toby, and Allan Dafoe. 2020a. "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, New York, NY, USA: Association for Computing Machinery, 173–79. https://doi.org/10.1145/3375627.3375815 (February 21, 2020).

———. 2020b. "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?" In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, New York: ACM. http://arxiv.org/abs/2001.00463 (January 9, 2020).

Smith, Bryant Walker. 2020. "New Technologies and Old Treaties." *AJIL Unbound* 114: 152–57.

Stahl, Bernd Carsten et al. 2021. "Organisational Responses to the Ethical Issues of Artificial Intelligence." *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01148-6 (February 23, 2021).

Stix, Charlotte. 2021. "Actionable Principles for Artificial Intelligence Policy: Three Pathways." *Science and Engineering Ethics* 27(1): 15.

Stone, Peter et al. 2016. *Artificial Intelligence and Life in 2030*. Stanford, CA: Stanford University. http://ai100.stanford.edu/2016-report (February 26, 2017).

Trajtenberg, Manuel. 2018. *AI as the next GPT: A Political-Economy Perspective*. National Bureau of Economic Research. Working Paper. http://www.nber.org/papers/w24245 (October 22, 2018).

Turner, Jacob. 2018. *Robot Rules: Regulating Artificial Intelligence*. New York, NY: Springer Berlin Heidelberg.

Verbruggen, Maaike. 2020. "In Defense of Technological Determinism." In Brussels, Belgium.

Watts, Sean. 2015. "Regulation-Tolerant Weapons, Regulation-Resistant Weapons and the Law of War." *International Law Studies* 91: 83.

Winter, Christoph et al. 2021. *Legal Priorities Research: A Research Agenda*. Legal Priorities Project. https://www.legalpriorities.org/research_agenda.pdf (January 15, 2021).

Zhang, Baobao, and Allan Dafoe. 2020. "U.S. Public Opinion on the Governance of Artificial Intelligence." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York NY USA: ACM, 187–93. http://dl.acm.org/doi/10.1145/3375627.3375827 (February 12, 2020).

Zhang, Daniel et al. 2021. *Artificial Intelligence Index Report 2020*. Stanford, CA: AI Index Steering Committee, Human-Centered AI Initiative, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf (March 3, 2021).

Zwetsloot, Remco, and Allan Dafoe. 2019. "Thinking About Risks From AI: Accidents, Misuse and Structure." *Lawfare*. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure (February 12, 2019).