‘Solving for X?’ Towards a problem-finding framework that grounds long-term governance strategies for artificial intelligence

Liu, Hin-Yan; Maas, Matthijs Michiel

Published in:
Futures The journal of policy, planning and futures studies

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Futures 126 (2021) 102672

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

'Solving for X?' Towards a problem-finding framework to ground long-term governance strategies for artificial intelligence

Hin-Yan Liu *, Matthijs M. Maas

Artificial Intelligence and Legal Disruption Research Group, Faculty of Law, University of Copenhagen, Denmark

ARTICLE INFO

Keywords:
Problem-finding
Governance puzzles
Governance disruptors
Macrostrategic trajectories & destinations
Governance goldilocks zone
Artificial intelligence

ABSTRACT

Change is hardly a new feature in human affairs. Yet something has begun to change in change. In the face of a range of emerging, complex, and interconnected global challenges, society's collective governance efforts may need to be put on a different footing. Many of these challenges derive from emerging technological developments – take Artificial Intelligence (AI), the focus of much contemporary governance scholarship and efforts. AI governance strategies have predominantly oriented themselves towards clear, discrete clusters of pre-defined problems. We argue that such 'problem-*solving*' approaches may be necessary, but are also insufficient in the face of many of the 'wicked problems' created or driven by AI. Accordingly, we propose in this paper a complementary framework for grounding long-term governance strategies for complex emerging issues such as AI into a 'problem-*finding*' orientation. We first provide a rationale by sketching the range of policy problems created by AI, and providing five reasons why problem-solving governance approaches to these challenges fail or fall short. We conversely argue that that creative, 'problem-*finding*' research into these governance challenges is not only warranted scientifically, but will also be critical in the formulation of governance strategies that are effective, meaningful, and resilient over the long-term. We accordingly illustrate the relation between and the complementarity of problem-solving and problem-finding research, by articulating a framework that distinguishes between four distinct 'levels' of governance: problem-solving research generally approaches AI (governance) issues from a perspective of (Level 0) 'business-as-usual' or as (Level 1) 'governance puzzle-solving'. In contrast, problem-finding approaches emphasize (Level 2) 'governance Disruptor-Finding'; or (Level 3) 'Charting Macrostrategic Trajectories'. We apply this theoretical framework to contemporary governance debates around AI throughout our analysis to elaborate upon and to better illustrate our framework. We conclude with reflections on nuances, implications, and shortcomings of this long-term governance framework, offering a range of observations on intra-level failure modes, between-level complementarities, within-level path dependencies, and the categorical boundary conditions of governability ('Governance Goldilocks Zone'). We suggest that this framework can help underpin more holistic approaches for long-term strategy-making across diverse policy domains and contexts, and help cross the bridge between concrete policies on local solutions, and longer-term considerations of path-dependent societal trajectories to avert, or joint visions towards which global communities can or should be rallied.

* Corresponding author.
E-mail addresses: hin-yan.liu@jur.ku.dk (H.-Y. Liu), Matthijs.Maas@jur.ku.dk (M.M. Maas).

https://doi.org/10.1016/j.futures.2020.102672
Received 1 May 2020; Received in revised form 19 November 2020; Accepted 1 December 2020
Available online 7 January 2021

1. Introduction

Change is not a new feature in human affairs. Nor is the need for some forms of governance in response to ongoing, and at times challenging, developments. Yet over the last two centuries, and acutely within the last few decades, something has begun to change in change. Daniel Deudney has observed how "the emergence and global spread of a modern civilization devoted to expanding scientific knowledge and developing new technologies has radically increased the rate, magnitude, complexity, novelty, and disruptiveness of change" (Deudney, 2018, 223). Accelerating climate collapse feedback loops are driven by economic and industrial systems intimately intertwined with fossil fuel energy extraction (Di Muzio, 2015); arsenals of nuclear weapons are held at hair-trigger launch alert, under a continuous state of functional 'thermonuclear monarchy' (Scarry, 2016); highly interlinked and efficient global supply chains, transport networks, and economies prove acutely vulnerable to pandemics; and emerging and potentially catastrophic risks from new technologies (Bostrom & Cirkovic, 2008; Bostrom, 2013; see also Rayfuse, 2017) ensure that our global society is increasingly grasped in the throes of 'turbo change' (Deudney, 2018, 223; Rosa, 2013). Accordingly, an array of new technologies are reshaping the stakes and rapidity of the societal, strategic, and even geophysical and ecosystem changes which we must manage, both today and over the long-term. The renowned biologist E.O. Wilson once playfully quipped that "The real problem of humanity is the following: we have paleolithic emotions; medieval institutions; and god-like technology" (Wilson, 2009). If that is so, can the diverse elements and processes constituting global governance evolve along with the changing stakes and nature of the challenge?

A key question for any broader project examining how 'strategy' ought to be generally reconfigured within society's complex networks of communities, organizations, or systems in order to adopt meaningful and effective long-term perspectives (van Assche, Verschraegen, Gruezmacher, & Boezeman, 2020) concerns how governance needs to be reoriented towards the long-term perspectives that flow from these urgent and unpredictable technological challenges. Moreover, the question may prove particularly salient and critical in the technological context. In the coming decades, existing governance instruments may well face a historical 'inflection point' in their relation to this series of emerging and 'converging' technologies (Pauwels, 2019). Our capacity to collectively manage our relationship to such technologies, such as artificial intelligence (AI), may prove to be both a key objective for, and a key stress (or litmus) test of, our governance regimes. How or why might we have to adapt or replace governance systems in the face of the 'technology tsunami' (Danzig, 2017)? Of course, this is not to say that we cannot influence, channel, or resist such technological shocks—indeed, contrary to frequent depictions of the regulatory 'pacing problem' (e.g. Hagemann, Huddleston, & Thierer, 2018; Marchant, 2011), it should be kept in mind that governance structures also provide the landscape that shapes and guides technological development, such that we do not need to resign ourselves to reactively responding to shocks (Crootof & Ard, 2021, 12) but can also seek to shape paths of development in advance. Faced with a 'tsunami', we need not just worry around how to manage or weather it, but also about how we might channel or even surf it.

Nonetheless, how might we go about making such governance changes? How can we ensure that we do not find ourselves "addressing twenty-first century challenges with twentieth-century laws" (Jensen, 2014, 253) and why might such extant approaches prove problematic in the first place? We face an increasingly complex historical challenge of evolving our governance systems—or, if necessary, developing new ones—to be up to the task of responsibly managing this array of potent challenges, and to do so not just for the near-, but also for the long-term.

This paper questions the premise that the extant regulatory and governance orders will remain fully intact and functional throughout the turbulent period ushered in by complex, 'transversal' emerging issues (see Morin et al., 2019). It does so taking its departure from one of these trend-drivers—artificial intelligence. We argue that AI and its applications will outfox contemporary regulatory and governance orders. Specifically, AI holds the potential to reveal, enable, or drive society towards new regulatory and governance equilibria, and is thus a prime candidate for triggering regulatory disruption. Again, this is not to say that we have to be merely reactive: we can and must influence and shape patterns of regulatory disruption; however, to do so, we have to be aware of the dynamics, pathways, and vectors of change (Liu et al., 2020).

What does this reveal? We argue that an underlying problem lies in the fact that many contemporary regulatory and governance orders (and their respective academic fields) continue to implicitly adopt or manifest problem-*solving* orientations. In other words, regulation and governance are endeavours that respond to recognised, defined, and compartmentalised problems. Overlooked in such an orientation, however, are the inherent and contingent limitations that needlessly straight-jacket the endeavour. By prematurely settling upon an overly and necessarily narrow, and somewhat contingent, range of issues to be addressed these problem-solving efforts may increase the probability of a turbulent or even catastrophic transition towards an AI-permeated world.

Accordingly, our aim in this paper is to shed light on the potential of, and prospects for a problem-*finding* orientation for formulating and anchoring long-term strategy assemblages for the governance of AI—strategies that take stock of the ways in which the processes, instruments, assumptions and even aims of existing governance orders will be shaped and disrupted by AI. Thus our overarching aim in this paper is to chart the underexplored terrain beyond the boundaries of current problem-*solving* regulatory and governance debates. By actively and systematically searching the potential problem-space opened up by AI, this paper aims to increase the confidence that efforts geared towards addressing presently-identified problems triggered by AI are relevant over the long-term, and do constitute the significant questions that need to be addressed, as well as to identify new potential problems that AI might raise for regulation and governance, which have been underexplored to date.

Accordingly, this paper proceeds as follows. We first ground a *rationale* for a problem-finding governance framework. We sketch the policy problems created by AI technology, and provide 5 reasons on the basis of which a problem-*solving* governance approach to these challenges fails or falls short. We then extend the argument by arguing that problem-solving governance approaches are ill-equipped to engage with 'wicked problems' created by AI. We conclude the rationale by grounding the epistemic and scientific validity of a complementary problem-*finding* AI governance (research) program. Secondly, we then set out our *framework* of AI governance across

four strategic 'levels' which correspond to AI governance approaches that take for granted different parameters, contraints and failure-modes of scholarship, which variably approaches AI issues from a perspective of (0) 'business-as-usual'; (1) 'puzzle-solving'; (2) 'governance disruptor-finding'; or (3) 'charting macrostrategic trajectories'. Thirdly, we provide *reflections* on nuances, implications, and shortcomings of this long-term governance framework. We discuss how to best coordinate AI governance strategies amongst and between these complementary four levels, in order to avoid the failure modes that might occur when restricting analysis or policy to any one of these levels. We assess how 'problem-finding' scholarship can be applied in problem-solving roles. We theorize how these four levels may constitute a 'Goldilocks Zone for Governance' which, in a larger perspective, highlights the assumptions and boundaries of any governance system; and we discuss how we should navigate the linkages between strategies and overarching narratives. We conclude by suggesting that these lessons are critical in structuring governance responses—to AI; and to other vectors of change—that are meaningful and resilient into the long term.

2. From problem-solving to problem-finding: a rationale

In the first place, it is important to come to understand the plausible shortfalls of problem-solving approaches. But to do so, we must understand what characterizes these approaches; why they fall short of providing long-term governance responses to AI; and why an alternative and complementary problem-*finding* paradigm may be both pragmatically effective, as well as epistemically and scientifically warranted.

2.1. Problem-solving governance is necessary but insufficient to AI

While there are a wide range of philosophical, theoretical, and technical perspectives on AI (Bhatnagar et al., 2018; Corea, 2020; Domingos, 2015; Legg & Hutter, 2007; Rapaport, 2020), in practical terms it can be understood as a generally enabling 'omni-use technology' (Clark, 2018). More precisely, it can be understood as a sociotechnical system in practice which consists of four layers (Maas, 2020, 37):

(1) a portfolio of algorithms and computational *techniques* (e.g. symbolic approaches; supervised learning, unsupervised learning; reinforcement learning; GANs…) which are embedded in physical computing platforms or distributed across networks.
(2) In turn, these can be used for automating and improving the accuracy, speed, and/or scale of (machine) decision-making in complex or large (data) environments; yielding a set of *capabilities* that can be used to support, substitute for- or improve upon human performance in specific tasks (including but not limited to 'pattern recognition', 'data classification', 'prediction', 'data generation', 'anomaly detection', 'optimization', 'independent decision-making', etc.) (Scharre & Horowitz, 2018).
(3) These tasks are individually narrow, but because these capabilities are relevant across diverse (industrial or strategic) contexts, they consequently support a spectrum of useful *applications* across diverse domains.
(4) These applications in turn produce diverse forms of new behaviour resulting in *sociotechnical changes*.

While today AI usage is still conditional on a range of resources—such as the availability of sufficient training data, computing hardware, technical talent, and organizational adaptation (Buchanan, 2020; Horowitz, 2018; Hwang, 2018)—these barriers are all falling at varying rates. As such, in terms of societal impact and governance stakes, artificial intelligence can be functionally modelled as a strategic 'General-Purpose Technology' along the lines of steampower, electricity, or computing (Leung, 2019; Trajtenberg, 2018).

Given the sheer breadth of its underlying technologies, capabilities, and potential applications,[1] it should be no surprise that AI throws up a vast range of issues across the policy-space, as reflected in the emerging literature studying distinct law and policy questions (Calo, 2017; Dafoe, 2018; Guihot, Matthew, & Suzor, 2017; Hoffmann-Riem, 2020; Scherer, 2016; Turner, 2018). Yet, the identified literature on discrete policy issues adopts a general problem-*solving* orientation.[2] Under this paradigm, policy scholarship compartmentalises AI's immense potential problem-space into specific problems—from the safety of autonomous vehicles to misuse of 'deep fakes', and from new criminal uses to 'killer robots'—that are often defined or segmented according to disciplinary perspectives

[1] Given this, it may not be meaningful in many cases to speak of 'AI' as a singular thing to be regulated. Nonetheless, in the context of this paper, while we highlight and distinguish between different problem areas and domains, we also emphasize the importance of recognizing commonalities in problem logics and themes across conventional domains (See also Crootof & Ard, 2021, 11); as such, for convenience, we will in some parts of our analysis refer to the suitcase term of 'AI' in the singular. We thank a reviewer for prompting this clarification.

[2] As always, no term is without its difficulties. There are some conceptual challenges with designating existing scholarship or governance approaches as 'problem solving'. For one, this (a) does not (directly) interrogate existing governance actors' privileged ability to determine which are and are not 'problems' to be 'solved' (e.g. certain AI-enabled technologies such as DeepFakes in some sense shift around (epistemic) power; that may make them a 'problem' for governments; but from the perspective of certain political actors they more obviously present themselves as an opportunity. However, since these latter actors do not conventionally 'do governance', their perspective is not easily presented under this framework). In the second place, and relatedly, (b) the term 'problem-solving' appears to at least implicitly pre-suppose or imply that the pre-existing state of affairs was broadly 'unproblematic' before it was rudely disrupted by a new technological 'problem' which upset that functional state of affairs—with the implication that governance might again be broadly unproblematic once that problem has been adequately solved. That would obviously be a highly normative and debatable assumption, which many (emerging) governance actors would challenge. However, it is a shortfall that primarily afflicts 'problem-solving' governance approaches—and we would emphasize that problem-finding approaches can bring in a wider range of actors, agendas, and values. We thank Henrik Palmer Olsen for prompting this reflection.

(Crootof & Ard, 2021; Petit, 2017). There are several interrelated limitations that flow from attempts to *solve* pre-packaged policy problems which, taken together, suggest that this is a necessary but insufficent strategy for long-term governance, especially in the context of societally-disruptive technologies such as AI.

The problem-*solving* orientation takes departure from issues that are discursively recognized or constructed (by policymakers as well as by scholars in a field) as a 'clear and present danger': that is, they often focus on issues that have given rise to vivid (if anecdotal) challenges such that they are discernible (and discerned) by public and policymakers; which are accordingly highlighted (and, in some cases, 'securitized' (Stritzel, 2014)) within government strategies; and which are validated, accepted, operationalized, and funded as legitimate and well-bounded research topics within existing disciplinary fields. In this way, the problem-solving orientation is in a sense a demand-driven (rather than supply-driven) approach to research program orientation and policy formulation. At its limits, it constitutes a 'local search' for the next urgent, externally-validated and -packaged problem that can be fitted to analysis by the discipline's available methodological or conceptual tools, rather than a 'global search' across the broader landscape of possible problems (let alone each problem's landscape of possible perspectives). To an extent, this orientation is understandable. Scholars only have so much time, so much attention. For all that interdisciplinarity is often praised, it can often appear a risky, difficult affair (Fish, 1989).

Yet while perhaps appealing and even understandable from the perspective of a pre-existing academic division of expertise and labour, however, five problems adhere to a problem-solving approach. In brief, these are (1) the unreflexive over-reliance on narrow *metaphors* or analogies in the face of multifarious, complex phenomena; (2) the *tautological* manner in which existing governance assemblages surface primarily those problems that are already (or most) legible to them; (3) the constrained *path-dependency*; (4) inability to clearly *distinguish* between symptomatic and root-cause problems; (5) the promotion of a *false sense of security* through a rapid delineation of the problem-space that prematurely closes-off large parts of the governance solution-space.

2.1.1. Unreflexive reliance on over-narrow metaphors

In the first place, in many areas, and especially where it comes to new technologies, problem-solving perspectives often, implicitly or explicitly, rely on drawing metaphors or analogies in order to fit a new problem or topic within the pre-existing boxes. Yet, as any given analogy or metaphor used to describe new technologies by necessity is incomplete (Balkin, 2015; Crootof, 2018), problem-*solving* perspectives that invoke them are necessarily partial and incomplete, and potentially even misleading, as they express only certain disconnected aspects of the overall challenge posed by the technology. This insufficiency of a problem-*solving* approach is reflected in the ancient parable of the 'blind men and the elephant' which warns against concluding that elephants are like pythons, simply because one has exhaustively touched one elephant's trunk (Saxe, 1872). One can see this same risk in the context of the multifarious debates round AI, which often explicitly or implicitly trade on highly different metaphors or understandings (Brockman, 2019; Parson, Re, Solow-Niederman, & Zeide, 2019). Amongst scholars writing on AI governance, many take comprehensively different views, not just on the individual questions of what AI *is*, how AI *operates*, how we *treat or relate to* AI technology, how we *use* AI in society, or what (unintended) *impacts* AI will have on our society down the road—but indeed take different positions on which of these above frames is even the relevant one for the policy question at hand. For instance, does the regulation of social robots turn on questions of whether the artefact *is* 'software' or a 'cyber-physical system' capable of physical damage (Calo, 2015, 2017)? On questions of how humans will *relate* to humanoid systems (Darling, 2012)? Or on questions of how they might provide new attack surfaces for criminal (mis)use (Brundage et al., 2018; Caldwell, Andrews, Tanay, & Griffin, 2020; Hayward & Maas, 2020; King, Aggarwal, Taddeo, & Floridi, 2019)? Each of these local perspectives provides pieces of the puzzle that can be valuable and legitimate, but often there is no reflection on how the specific narrow perspective—not just on these questions, but even on which of these questions is the appropriate one to ask—is but one of various ways to frame and understand the salient features or key impacts of the technology (See also Cave & Dihal, 2019; Cave et al., 2020).

Indeed, the 'blind man and the elephant' metaphor may itself be overly restrictive: by extending or reframing the parable, we can see that a problem-solving approach risks misapprehending the problem in many more ways. After all, the problem is not only that an overtly narrow view may result in us (1) believing that an elephant is like a snake, on the basis of studying its trunk (i.e. *mis-characterizing problem X on the basis of its local facet)*. Rather, the very focus on 'the elephant in isolation' occludes a range of other potential issues. To stay within the framing of the parable[3] ; our narrow focus on the 'elephant' in front of us may lead us to (2) fail to notice the tiger that is crouching up behind us (*missing out on urgent 'out-of-context' problems or development Y*); (3) fail to highlight how this elephant came to potentially be brought out of its habitat and social context and presented to us, and how it came to be constructed and designated as a salient object in need of closer examination and apprehension (*missing out on the socio-political roots shaping the research agenda on X*); (4) fail to apprehend how, rather than apprehending the 'essence' of an individual elephant, it is also relevant to understand how that elephant fits within a broader ecosystem—or what the health of the elephant tells us about the broader ecosystem (*missing out on the interrelation of problem X with seemingly 'unrelated' entities, problems, or developments*); (5) base our understanding of the elephant on a momentary snapshot, rather than a longitudinal study of an elephant's developments over the course of its lifecycle (*taking an overtly static rather than dynamic view*); (6) take for granted that an elephant can be appropriately and sufficiently apprehended through touch—and appropriately and sufficiently described at the level of human-scale features—rather than considering how additional (and different) insights into the elephant could be gleaned from using scientific tools that enable us to study it at

[3] There is, of course, an irony in critiquing the prominence of (legal) metaphors within problem-solving AI policy research, through the use of this parable. However, we argue that in this instance our extension of the 'blind men and the elephant' metaphor works to highlight the contingency of taking for granted a classic or narrow interpretation of a certain parable (and the diversity of alternate and even contradictory problem formulations that could be derived or accomodated even within one metaphor-paradigm).

different levels of analysis, from fundamental physics, cellular biology, or etiology (*taking for granted a narrow methodological toolkit to studying problem X at a narrow level of granularity*).

This illustrates the epistemic limits and shortfalls of problem-*solving* approaches to understanding and governing new technologies. These approaches frequently focus on: the here and now; the direct harm; the last person to touch the technology prior to an accident; or the degree to which any solution can be found which mitigates the direct fallout; or redresses the directly (legally-recognized) injury. For example, legal problems that flow from AI applications might revolve around questions of liability for entities occupying a liminal position between agent and object, which befuddle existing legal doctrine (Liu, 2012; see generally Turner, 2018). Accommodating this challenge may be necessary for the law to regain coherence, but does little to address competing plausible models of AI applications such as a networks, or a systems, approach to understanding AI (Ekelhof, 2019; Liu, 2019a). This approach or mindset suggests that legal problems are just that, 'problems' that can be identified and interrogated with legal tools, but that this process itself skews the form that the underlying challenges are deemed to take, and the types of (legal) responses that are legitimate to propose in their wake.

2.1.2. Tautological responses

Second, a problem-solving approach ensures that any problems that are surfaced by examination are often *tautological* but somewhat arbitrary in relation to the significance of their potential sociotechnical impact. AI applications might generate legal problems because those are the problems that AI generates for the existing law—more than for the underlying society. In practice, this might often highlight legally or philosophically 'unusual' problems which are clearly and obviously blurring our existing 'symbolic order'—that is, the fundamental distinctions and categories which a general society, broad governance network, or specific legal order relies upon to draw boundaries and understand the relevant reality (see Douglas, 1966)—over less 'legible' but more 'socially disruptive' developments that provide scholars with less 'satisfying' analytical or legal puzzles. One might consider for instance the way by which discourses on self-driving cars have become dominated, for better or worse, by their linkage to the high-profile 'trolley problem' thought experiment from ethics, a development many have begun to critique as having led such governance debates down a dead-end (Bauman, Peter McGraw, Bartels, & Warren, 2014; Cunneen et al., 2020; De Freitas, Anthony, & Alvarez, 2019; Himmelreich, 2018; Jaques, 2019; Wolkenstein, 2018). A key problem here is that within problem-solving approaches there is often no reliable or unambiguous appraisal process through which to gauge whether or why those problems that are generated or focused on, are likely to be the significant and persistent problems into the future, and not just solely 'legally interesting'. If problem-solving approaches get hooked on temporary, doctrinal questions which solve unrepresentative 'edge' problems (especially problems which might, in due course, end up being 'dissolved' anyway as a result of technological developments or 'out-of-domain' social or regulatory changes), however, this bodes ill for their viability or sufficiency into the long-term.

2.1.3. Limiting path-dependencies and 'brittleness'

Third, there are clear *path dependencies* inherent to problem-solving approaches. The problems that arise are frequently understood as variations on previously validated (if not necessarily 'solved') problems, and responses often deploy analogies to those that were invoked in past iterations (Crootof, 2018, 2019b; Mandel, 2017). This sets strong constraints upon the trajectory of problem-solving efforts, ensuring diminishing—or even 'negative'—returns when these efforts are brought to bear on genuinely fresh challenges or unseen cases.

Curiously, in this way a legal system is not unlike a contemporary machine learning algorithm itself: it works entirely and solely from its training data, and so can be structurally *biased* (depending on what elements or cases are under- or over-represented in its 'training set'), or *brittle* when it encounters genuinely unprecedented 'edge cases' that combine certain strange features of past cases, or bring in genuinely unprecedented ('out-of-training-distribution') features—as in the old legal maxim that 'hard cases make bad law' (Davis & Stark, 2001). Moreover, just as machine learning systems have proven susceptible to so-called 'adversarial input' (Goodfellow, Shlens, & Szegedy, 2014; Hendrycks, Zhao, Basart, Steinhardt, & Song, 2019), a legal or political system that hems closely (and predictably) to certain past categories may be adversarially 'spoofed' by certain, well-resourced actors who can construct or structure their 'input' in ways that exploit these features of the system in order to produce 'output' to their liking. Such strategies have been well-docmented in the legal context, consider US companies invoking 'free speech' as category to secure unrestricted campaign finance laws (Citizens United v Federal Election Commission, 2010). In more political arenas, consider the ways in which certain actors have 'securitized' issues such as migration, in order to re-deploy and re-direct the age-old state logic of 'security' towards certain desired political ends (Huysmans, 2000).

Beyond these general shortfalls, what is the problem with such problem-solving path dependencies? Given the breadth and depth of challenges envisaged by the widespread introduction of AI into society, it is arguable that a doctrinal solution to the liability questions raised by AI will at best produce marginal returns. This is especially since retention or reaffirmation of the contemporary legal paradigm many instead hinder attempts to a fuller exploitation of the benefits that the technology would otherwise be able to offer.

That is not to say, of course, that this 'problem-solving path-dependency' is the only such rigidity limiting governance. Indeed, as scholars in the field of Evolutionary Governance Theory have demonstrated, there are always various other 'dependencies'—from path dependencies to interdependencies, and from goal dependencies to material dependencies—which bound governance actors' ability to strategize far beyond their found governance path (van Assche et al., 2020, 7). Such path dependencies can therefore only be adequately understood together with various interdependencies and goal dependencies (the reflexive and at times self-fulfilling impact of visions for the future on current governance), as well as material dependencies such as the technologies entrenched in the system's discourse, policy, organization, or regimes of cooperation and coordination.

Thus, while many elements of governance are contingent and could in principle be reconfigured—even radically so—van Assche,

Beunen, & Duineveld, 2014, 5) also note that "[o]ne cannot jump from each branch in the evolutionary tree to each imaginable other branch". Nonetheless, in some cases the path-dependency of problem-solving approaches appears particularly constrained, and sufficiently contingent, that further exploration could be fruitful. In particular, the insight that the existing regulatory order is but one possible 'governance strategy' equilibrium, does not mean that the 'leap' to other such equilibria will be easy (indeed, if it were, the present state would not be so much of an attractor at all, and would likely have 'decayed' to one). Rather, the point is that while this present equilibrium may have proven proficient at managing sociotechnical changes thus far, it may no longer be appropriate nor sufficient in securing long-term interests—or realizing long-term perspectives—in the 'turbo-change' era of AI.

2.1.4. Inability to differentiate between root problems and surface challenges

Fourth, a different objection to problem-solving, related to the first, turns on the fact that it is 'centrality-blind'—that is, it does not easily (or at best only retrospectively) identify and prioritize governance bottlenecks or cross-cutting themes. Presently-identified problems may be differentiated into *symptomatic* and *root-cause* problems, but the problem-solving orientation does not systematically draw this distinction. Thus, much problem-solving work becomes squandered in ameliorating the symptoms rather than the root-causes of a problem: questions of AI authorship arise in the context of intellectual property law (Kaminski, 2017); questions of responsibility are asked in the the context of autonomous weapons systems (AWS) (Jain, 2016); and questions of liability are raised in the context of accidents caused by autonomous vehicles (Boeglin, 2015; Schellekens, 2015). While legitimate legal problems, these are all symptoms of the underlying uncertainty over the liminal status of AI applications, which is the root-cause legal problem. The point here is that there are neither appraisal processes, nor confidence more generally, as to whether the policy problems posed are the central or core problems, and not merely their manifestation in certain areas. This ensures that there may also be a degree of disconnection amongst problem-solving governance approaches, such that there is potentially much double work or redundancy by virtue of an uncertainty as to the "level" of the problems that are tackled and what parallel efforts are underway.

2.1.5. Promoting a false sense of security

Fifth and finally, there is the *false sense of security* inherent within the problem-solving approach. Most fundamentally, it implicitly supports a world-view whereby we face only 'problems', when in fact many of the challenges we face today might be better understood and approached, at least upon first encounter, as 'mysteries' (Chomsky, 1976). Through emphasising that new challenges already "appear to be within the reach of approaches and concepts that are moderately well understood" (Chomsky, 1976, 281), problem-solving approaches suggests that while we may not yet have the solutions, we have an idea as to what those might look like and how we might get there. In other words, we may not have the solution at hand, but we surely have already delineated the space within which we are sure it can (and therefore ensure it must) be found. This can be juxtaposed against 'mysteries', which "remain as obscure to us today as when they were originally formulated" (Chomsky, 1976, 281). In this context, problem-solving suggests greater understanding and mastery of challenges than might otherwise be justified. This becomes an issue when epistemic certainty is not so warranted. More superficially are considerations around the appropriateness, adequacy, and sufficiency of problem-solving approaches. What we mean here is that, even where policy problems are solved, there is not necessarily a connection between this solution and the relevance of the solution to the continuing and future stability of the regulatory and governance orders. This relates back to the absence of processes for assessing the relevance of solving particular problems to the underlying challenges at the source of those problems.

A different way of articulating the objections to an exclusively problem-solving approach can be captured concisely in the context of "wicked problems" which may serve as a bridge to our proposed emphasis upon problem-*finding*.

2.2. Problem-solving governance is an inadequate response to "wicked" AI problems

Introduced in an influential 1973 article, Rittel and Webber introduced the concept of 'wicked problems'—which they contrasted to 'tame', soluble problems in some scientific fields such as mathematics, chemistry or engineering—as problems that are difficult or impossible to solve because our requirements for a solution are often incomplete, changing, or in tension (such as in situations where there are no undisputable public goods); because policy problems cannot be definitively described; or because the effort to solve one aspect of a wicked problem may reveal or create other problems. By representing "wicked problems" as an array of complex, open-ended, and intractable problems,[4] Brian Head (2008) developed an alternative rubric under which many governance challenges, including those posed by AI, can be productively examined.[5]

To be sure, it should be cautioned that merely labelling a problem as "wicked" may not necessarily assist in solving it—in the same way that labelling a system as 'complex', or a phenomenon as 'emergent' by itself does not assist in actually understanding any of its dynamics or behaviour, and rather should be understood better as a signpost highlighting these as topics in urgent need of further investigation. However, the benefits of the 'wicked problems' framework are less about providing a clear solution rubric,[6] and aims

[4] In this section, the terminological reversion to 'problems', as opposed to 'mysteries' (Chomsky, 1976), is unavoidable given the use of 'problem'-terminology in the wicked problems literature itself. The question of how to integrate these into a concept of 'wicked mysteries' is left to future work.

[5] Although the 'wicked problems' framework has also come under critique in recent years, on both conceptual and practical grounds. See for instance (Noordegraaf, Douglas, Geuijen, & Van Der Steen, 2019; Peters & Tarpey, 2019; Turnbull & Hoppe, 2019).

[6] Although see the distinction between 'coping', 'taming', or 'solving' in Daviter (2017).

more to focus attention 'on the understandings that have shaped problem-identification and thus the frames for generating problem-solutions' (2008, 106) in the first place. The interrogation of upstream factors impinging upon problem-identification may serve as a response to the missing processes for identifying, ascertaining and evaluating the centrality or significance of downstream problems impugned in the section above. Indeed, in conceptualising wicked problems as the convergence of uncertainty, complexity and value-divergence, Head (2008, 106) suggests that failures to adequately respond to wicked problems may be due to the fact that:

- The "problems" are poorly identified and scoped;
- The problems themselves may be constantly changing.
- Solutions may be addressing the symptoms instead of the underlying causes.
- People may disagree so strongly that many solution-options are unworkable.
- The knowledge base required for effective implementation may be weak, fragmented or contested.
- Some solutions may depend on achieving major shifts in attitudes and behaviours […] but there are insufficient incentives or points of leverage to ensure that such shifts are actualised (Head, 2008, 106).

A common denominator to these failure modes inheres within poor mappings of the potential problem-space. Not only might the problem formulation itself be mistaken or misaligned, but the 'true' nature of the problem might not be one which is directly connected to the external real-world challenges that are most readily perceived as problems. Instead, as this list suggests, the root of the challenge might instead lie in social dynamic, economic, and political considerations. Yet, since subsequent problem-solving endeavours are sculpted in large part by the outcome of problem-identification and -definition processes, overly narrow or static problem initial formulations may lead to inadequate or brittle responses that fall short of addressing such problems. In slogan form: solving individual pieces of the puzzle provides little or no protection against against the larger threat posed by wicked problems. Such local responses may at times be orthogonal to addressing the underlying problem—and at worst may be damaging, either in their direct results, or because they forestall deeper responses.

Scholars have in recent years increasingly highlighted how legacy institutions and policy toolkits derived from the management of 'complicated' problems may prove ill-suited for the management of 'complex' problems (Kreienkamp & Pegram, 2020; Morin et al., 2019). Applied to the challenges posed by AI governance, the wicked problem framework suggests that problem-*solving* approaches by themselves may not be sufficient (Rittel & Webber, 1973, 160–67; see also Gruetzemacher, 2018). This is because AI itself not only poses complex, wicked problems, but because the application of this technology can also be disruptive to governance responses themselves, as it can alter the key parameters of a governance path; the interests of different actors in the network; the legal principles and processes; the regulators' and regulatees' values; or the regulatory modalities at play in the policymaking portfolio. To address the core characteristics of such changes, persistent (continuous) and pervasive (wide-ranging) problem-*finding* approaches should be articulated. As such, we will shortly seek to elaborate upon and structure one such problem-*finding* approach across four strategic levels below, as a means to give granularity to the multifaceted wicked problems generated by the infusion of AI into society. Before doing so, however, we will first reflect on the epistemic and scientific foundations of a problem-*finding* approach.

2.3. The epistemic validity of a problem-finding approach to long-term AI governance

The problem-finding/problem-solving distinction has parallels in the division between normal science as 'puzzle-solving' and revolutionary science as 'paradigm-shifting' (Kuhn, 1970); in the earlier-alluded-to separation of ignorance into 'problems' and 'mysteries' (Chomsky, 1976), and in the scientific creativity of research communities variably taking the form of incremental and secure 'cold searches' or far-reaching yet uncertain 'hot searches' (Currie, 2019). Indeed, the distinct psychological learning strategies displayed throughout human childhood have been likened to foundational AI research paradigms, which also balance algorithms between paired phases of '*exploration*'—wandering far; gathering information, and phases of '*exploitation*'—acting on the information gathered to the greatest effect (Gopnik et al., 2017, 7893; in: Currie, 2019, 5). One might also compare the idea, in algorithmic evolutionary optimization, of 'simulated annealing' (Derman, 2018) where, by introducing volatility and 'shaking up' its search, an algorithm can find its way to better solutions outside the incremental search-space.

The fact that a self-similar dyad can be found in philosophy, logic, evolution, computing, and even human psychology, is indicative of not only its scientific validity, but also its critical role in governance strategies. Analogously, while problem-*solving* work ('puzzle-solving; 'addressing problems'; 'cold searches'; 'exploitation') is clearly the bread and butter of many scientific research communities, open-ended problem-*finding* research approaches ('paradigm-shifting'; 'pursuing mysteries'; 'hot searches'; 'exploration') are and ought to be a natural and critical complement. This is for four reasons.

First, such problem-finding work may itself *produce foundational academic insights* (for example about the interrelation of law and governance priorities across different scales or levels). Indeed, meta-science research has suggested that, even as exploratory research papers face a scientific 'bias against novelty', they are more likely to become amongst the top 1% highly cited papers, and to inspire follow-on highly cited research across many disciplines in the long run (Wang, Veugelers, & Stephan, 2017).

In the second place, open-ended, problem-finding research agendas can also work in a *reflexive equilibrium* with incremental policy-solving research agendas, by providing broader-scope insight into how individual policy areas operate. In this way, such research agendas can address some of the 'institutional mismatch' to addressing a set of 'transversal' issues (including, but not limited to, AI) which drive externalities beyond or across familiar problem categories and domains (Morin et al., 2019, 2–3).

Thirdly, and importantly, problem-finding approaches can at times *reveal new 'crucial considerations'* (Bostrom, 2014a): new arguments or propositions which, if true, reveal the need for not just small course corrections, but major overhauls in direction or

priorities within our governance clusters and actions. For instance, research agendas into avenues to 'automatically detect' media forged by 'DeepFakes' are dependent on a crucial assumption that such detection measures can always be developed relatively quickly, and that this strategy is viable more or less indefinitely. If, however, the 'offence-defence balance' of AI research in this area is such that (the dissemination of) progress in these techniques aids attackers more than defenders (Shevlane & Dafoe, 2020); or if, at the limit, progress in DeepFakes will mean that within a few short years detection will simply be functionally impossible (Engler, 2019), this narrow research agenda would be overturned. A problem-finding orientation can therefore frame a lodestar to aim towards, rather than merely identifying obstacles present and apparent in our current trajectory.

Finally, problem-finding approaches enable us to forestall or short-circuit Collingridge's 'dilemma of control', which holds that "[w]hen change is easy, the need for it cannot be foreseen; when change is apparent, change has become expensive, difficult, and time consuming" (Collingridge, 1980, 11). In that sense, problem-finding and problem-solving approaches fall respectively into each side of this dilemma. In this context, the power problem inherent in attempting change when it is 'expensive, difficult, and time consuming' reveals the futility of foregrounding problem-solving approaches in relation to long-term perspectives. Problem-solving approaches, somewhat obviously, become operational precisely where the desire or need for change confronts the inertia of a complex and interconnected world that frustrate such efforts. Long-term perspectives, however, reside in the information problem aspect of the dilemma: problem-finding approaches thus attempt to explore the potential problem-space while 'change is easy' and course corrections and changes are still possible. Yet, the dilemma nature of the challenge suggests that efforts of both of its sides are necessary, and the long-term perspective merely suggests that a recalibration of these efforts needs to be effected.

3. A proposed problem-*finding* framework across four strategic levels

We propose a theoretical framework or typology for mapping the long-term strategic landscape which we apply to the seemingly distinct cluster of policy domains surrounding AI.[7] In our understanding, we subdivide governance responses to AI's problems into four strategic 'levels'.[8] Levels 0 & 1 concern default, 'problem-*solving*' responses to governance[9] : (0) 'business-as-usual' governance demands the simple application of existing governance structures to the new problems; whereas (1) 'governance puzzle-solving' admits at least a certain need to reconceptualise, stretch or reconfigure existing legal doctrinal categories or governance concepts, in order to address those problems. Conversely, Levels 2 & 3 concern systemic challenges that currently lie beyond the boundaries of contemporary legal frameworks. They involve scholarship aimed at (2) finding potential 'disruptors' of core governance assumptions; or at (3) charting macrostrategic trajectories & destinations. Together, this creates a comparative framework within which to compare the problem orientations, contribution orientations, and limits of these distinct analytical and governance perspectives (see Table 1). Accordingly, such research lines call upon problem-*finding* approaches that convert, contextualize, prioritize, and structure the unknown problem-space into clusters or research agendas of problems that could then be addressed at Levels 0 & 1.

3.1. *Level 0 - 'business-as-usual' governance*

The 'default' strategy involves an adherence to 'normal' governance processes, norms, structures and concepts, however constructed. While it does not deny the emergence of certain problems or challenges, it can be slow to recognize them—and even when it does, it will deny their fundamental 'newness'. This governance level is therefore rigorously focused on solving problems within (or by extension or application of) the existing governance system. In this sense, Level 0 governance is *narrowly* problem-*solving:* it recognizes the local 'problem', but requires that any solutions are chosen from the existing set of solutions made available. It is reactive, and seeks to re-establish or re-balance the old 'governance equilibria'.

[7] To be clear, while we use AI as a focal technology, the general argument is applicable to the way long-term governance would engage with any new technology (cf. Liu et al., 2020). See also the critique, of legal scholars' problem-solving responses to technology, in Tranter (2011). We thank Roger Brownsword for prompting this clarification.

[8] We should make a few important caveats regarding our usage of the term 'levels' in this typology. In the first place, by referring to 'levels', we do not mean to imply a ranking of importance: another way to read or refer to them would be '*types*' of governance strategies—although we prefer to retain the imperfect term 'levels' for the simple reason that it emphasises more (in a way that 'types' or 'categories' does not do) how the different varieties of governance strategies are not mutually exclusive and isolated, but interlinked components of a larger governance system. In the second place, and pragmatically, this terminology should not be confused with the existing scholarship on 'multi-level governance' (See Bache & Flinders, 2004; Zürn, 2012), which explores how policymaking power and authority are distributed both throughout the hierarchies of government, as well as horizontally towards other non-state actors. In our usage, by contrast, 'levels' correspond not to the 'level' (that is, the actor) at which a given strategy is rooted or situated, but rather to the 'level' at which a strategy seeks to exert leverage (that is, its scope, or the degree to which it variably takes for granted, or interrogates the problem formulation, parameters, and larger 'outside context'. In that sense, these 'levels' are ordinal ('Level 3' takes a broader scope than 'Level 2', which takes a broader scope than 'Level 1'), but do not necessarily connote ranked or hierarchical levels of influence (e.g. 'Level 3' strategy are not by definition or always upstream from 'lower' levels, even if they might offer greater possibilities for creation and consideration of long-term perspectives); rather these levels co-constitute one another.

[9] Again, to avoid terminological confusion, it should be kept in mind that these Levels 0 and 1 are 'problem-*solving*'. We nevertheless include them as a key baseline component in our integrated framework which (encompassing the directly problem-*finding* strategies at Levels 2 and 3) we collectively refer to as the 'problem-*finding*' framework for AI governance. This is because we consider an integrated governance framework capable of carrying out 'problem-finding' governance as one that is capable of *encompassing* and *complementing* existing 'problem-solving' approaches, not of wholly replacing them.

For instance, AI policy work at this level seeks to accommodate the issues raised by AI within the extant legal frameworks, arguing that legal doctrinal changes to accommodate the technologies are not necessary—in the terminology of cyberlaw, that distinct, domain-specific legal innovations would be as superfluous as articulating a comprehensive 'law of the horse' (Easterbrook, 1996). This position holds that all appropriate 'metaphors' to be encoded explicitly in law or implicitly in broader governance discourse can be found in past technologies (Mandel, 2017). Because these are easy to understand, much of the problem-*solving* work done today in distinct sub-fields on AI policy is situated at Level 0. For example, it is manifested in attempts to simply extend the existing legal framework and principles of international humanitarian law (IHL), to cover any and all issues that are raised by AWS (Anderson, Reisner, & Waxman, 2014; Schmitt, 2013).

While it might be easy or tempting to dismiss such work out of hand, such a response would likely be neither fair nor warranted. Indeed, to be clear, a 'business-as-usual' response is not an obviously incorrect orientation in many cases (though it may be an insufficient one at a systemic level, across all cases). It relies on a certain degree of conceptual laxity—the willingness to implicitly (and at times perhaps unreflexively) stretch and re-interpret existing governance categories, just so that we can ensure their easy application to new cases. Such an approach avoids continuous, and costly, transaction costs of figuring out whether or not every new case or technology is new. Indeed, to an extent, the 'business-as-usual' heuristic is critical to governance: if we lacked one, we might easily find ourselves paralyzed: we would risk 'overfitting' (in the machine-learning sense) on our past governance experiences—leading one to argue, for instance, that a car crash involving a neon-painted car could not possibly be covered under existing laws, since no similarly-coloured car was ever involved in a car crash in the given jurisdiction, and this case was therefore clearly without precedent. Lyria Bennett Moses likewise argues that "[m]ost of the time, a law predating technological change will apply in the new circumstances without any confusion. For example, traffic rules continue to apply to cars with electric windows, and no sane person would seek to challenge these laws as inapplicable or write an article calling for the law to be clarified." (Bennett Moses, 2007, 596).

Moreover, an approach of implicit 'reinterpretation' and extension of existing governance categories may have some limited (temporary) merit *even* in the face of actual underlying real technological changes. All else being equal, holding up a 'pretence of continuity' might avoid certain externalities in the form of continuous regulatory uncertainty, or (in the case of high-stakes technologies) continuous contestation or attempted renegotiation by stakeholders. It can therefore be a necessary governance strategy in areas where reaching original political agreement on the existing governance arrangement proved thorny or difficult—such as for instance in international arms control agreements on certain types of new weapons (Crootof, 2019b; Maas, 2019c; Picker, 2001).

Nonetheless, while these Level 0 responses may be somewhat understandable, and even necessary in some cases, it is likely that both their drawbacks and inadequacy are underestimated. In the first case, such moves can be brittle: when pushed too far, attempts to capture obviously (or to the public, apparently) revolutionary technologies under age-old governance approaches may fail the 'laugh test', with the result that 'the law runs out' (Crootof, 2019b, 17), threatening the credibility of governance efforts within that specific domain, as well as, at length, the more general legitimacy of the governance system that produced or allowed it (as critics can add another anecdote to a list of 'absurdities' produced by the extant system). In other cases, and especially in global governance contexts, certain actors may simply not accept an extension of 'business-as-usual' to new technologies, leading to new explicit or implicit contestation over specific norms, entire strategies, or certain actors' legitimacy or authority to steward those policies (see also Zürn, 2018).

More fundamentally, however, because Level 0 approaches do not seek to (deeply) track or understand the nature of the ongoing changes (beyond the bare minimum necessary to address local symptomatic problems), they cannot be expected to address these in the long run. Instead, such governance structures may implicitly build up a 'technology debt', becoming hollowed out by changing practices, and ultimately being rendered into obsolete 'jurisprudential space junk' (Crootof, 2019a) as a result of changing practices. Ultimately, this approach also cannot reckon with possibly new, unanticipated features or area cross-overs, and therefore cannot provide the foundation for a problem-finding approach. In sum, Level 0 response might be necessary for any governance system, yet they are not (and potentially cannot be) sufficient. At their best (or at most), they might serve as 'holding actions' that prevent near-term problems from 'unwrenching' or distracting society so badly or so frequently that it cannot even reckon with long-term trends. But by itself it is certainly not able to adopt a creative problem-finding approach capable of providing a foundation for long-term governance, because in many cases they 'deny' that there is any deeper problem (beyond the surface disruption to be stilled). In doing so, it risks building up a 'problem debt'.

3.2. *Level 1 – governance puzzle-solving*

Closely related to Level 0 governance, but somewhat distinct, are perspectives that approach specific new problems as 'governance puzzles' to be solved. Work at this level still narrowly or primarily emphasizes the importance of fixing the direct problem at hand—that is, it takes for granted a narrow rationale—but it opens up for the possibility that doing so may require innovation and (even far-reaching) change in the regulatory tools or governance processes. As such, in contrast to Level 0, Level 1 governance is at least *broadly* problem-*solving:* it admits, at least in principle, that solving the direct problem in front of us may require innovations and changes in the set of available solutions, however, it does still narrowly seek to maintain or restore the existing status *ex ante*.

In an AI context, this could be found in debates about discovering ways to re-conceptualize privacy (in the face of new AI challenges) in order to functionally restore the equivalent of past privacy protections to citizens. It may also extend to proposed innovations in governance approaches—such as the posited use of technologies such as AI systems as 'privacy protector' (Els, 2017; Gasser, 2016) in order to secure established goals or rights. As noted, Level 1 analysis is largely similar to Level 0, but it admits to a certain degree of novelty in the challenge, and accordingly to a flexibility in governance response, while anchored upon an allegiance or commitment to the existing status: the end-product is the same, even if the route is different.

Table 1
Taxonomy of problem-solving (0-1) and problem-finding (2-3) AI governance strategies.

Level	Governance Perspective	Problem orientation	Aim & purpose	Limits
0	**Business-as-usual**	• *Solving* local societal 'problems' created by new AI systems	• Denies 'newness'; doctrinal changes not necessary • Seeks accommodation within current legal framework, through (1) extension of legal framework; (2) re-interpretation of new problem as old problem	*Does not try to anticipate long- term* • Reliance on metaphor • Brittle; interpretations fail 'laugh test' • Fail to understand underlying nature of problem, or track cross-sector manifestations
1	**Puzzle-Solving**	• *Solving* local societal 'problems' created by new AI systems • *Solving* doctrinal inconsistencies in the law revealed	• *Broadly* problem-solving: still focuses on fixing the direct problem at hand, but allows that this may require doctrinal legal changes	*Does not try to anticipate long- term* • (as Level 0) • Still mostly aims to maintain or restore existing status *ex ante*
2	**Disruptor-Finding**	• *Finding* AI developments that may 'disrupt' assumptions of or conditions for the governance system, such as (1) changes to the problem portfolio; (2) to regulators' goals; (3) to governance tools, (4) to societal values	• Does not take for granted scope and nature of governance problems at hands • Aims to understand 'out of bound' strategic barriers or deflectors	• May *miss the forest for the trees* (relative to Level 3)
3	**Charting Macrostrategic Trajectories**	• *Finding* 'crucial considerations' for how AI could shift overall trajectory of society towards (a) *terminal trajectories;* (b) *vulnerable worlds*; (c) *exposed worlds*; (d) *bad worlds*; (e) *good worlds*	• Focus first on identifying overarching trends and dependencies which might be resistant to 'piecemeal' governance responses at Levels 0−2. • Identify leverage points for intervention to shift trajectory	• May underappreciate Level 2 governance disruptors, and therefore become a set of *grand futures, derailed by detail*

3.3. Level 2 – governance disruptor-finding

In contrast, work at Level 2 begins to realize a problem-*finding* orientation. Its focus is less on the narrow problems as they are given, presented, or highlighted within pre-existing disciplinary boundaries, and instead seeks to explore underlying patterns, linkages or under-illuminated problem clusters. Recognizing that 'law' is only one 'regulatory modality' (Lessig, 1998, 1999), Level 2 scholarship shifts focus from changes beyond narrowly understood laws or government regulation, towards to the broader *regulatory* system—and how developments in a particular technology (such as AI) may end up disrupting or deflecting governance efforts (either those aimed at the technology, or even in general). Accordingly, Level 2 analyses can emphasize at least four distinct angles.

First, such scholarship can explore how ongoing technological progress—or altered social practices that seize upon pre-existing but formerly marginal affordances in new, salient ways—can *shift or expand a technology's 'problem portfolio*' in ways that render initial, now path-dependent, regulatory efforts inadequate or even counterproductive. Such work can expand the subjective problem portfolio to correspond better to the evolving opportunity/risk profile as manifested by the *affordances* (Gläveanu, 2012; Norman, 2013) of a new technology. To be sure, this is not to say that all such structural change comes from changes in technology, or that all technological change brings about structural shifts such as this. Nonetheless, at times such shifts can be key and disruptive to extant strategic trajectories. For example, the risk profile of AWS can shift with technical progress: issues that appeared critical at an early stage (the potential to violate IHL principles), may soon become matched or eclipsed by far-reaching challenges in other domains: for instance, military AI systems may generate risks of operational safety, 'flash wars', or strategic instability (Danzig, 2018; Geist & Lohn, 2018; Maas, 2019b; Scharre, 2016b; Sharikov, 2018). Moreover, much greater than the direct risks from misuse or accident, might be indirect risks from 'structure'—deriving from the way new AI systems or capabilities shape the landscape of incentives around actors, in potentially hazardous ways (van der Loeff, Bassi, Kapila, & Gamper, 2019; Zwetsloot & Dafoe, 2019). Furthermore, AI development in areas such as one-shot learning (Ram, 2019), 'simulation transfer' (OpenAI et al., 2018), synthetic data, and others can lower the data or expertise needs for using AI, and thereby can lower the proliferation threshold of certain AI systems to new, non-treaty parties. Indeed, in general, greater data efficiency can have considerable and far-reaching governance implications (Tucker, Anderljung, & Dafoe, 2020). In these cases, neither 'solving the AWS problem' at Level 0 by extending IHL principles, nor tailoring legal principles to AWS at Level 1 does much good. Thus, policy responses at Levels 0–1 may no longer track—and may even occlude—the changing character of threats posed by AI systems. Instead, Level 2 articulates more adaptive or 'preventative security governance' approaches (Garcia, 2016, 2018) or 'innovation-proof' governance (Maas, 2019c; see also Crootof, 2019b). Thus in our AWS example, Level 0–1 approaches suggest winning a legal battle, but in reality the regulatory war is lost.

Second, AI can *change policymakers' goals* in formulating law. The regulatory mind-set may shift from 'legal coherentism' towards 'regulatory instrumentalism' or even 'technocracy' (Brownsword, 2018), potentially spelling the 'death of [regulation through] rules and standards' (Casey & Niblett, 2017). Moreover, AI-sparked shifts in the international balance of power could destabilise the global legal order, just as previous technologies have upset international law (Picker, 2001). AI could do so by differentially empowering illiberal states (Danzig, 2017; Harari, 2018; Wright, 2018), or by eroding the buy-in of powerful states, into the multilateral international legal order (Danzig, 2017; Deeks, 2020; Maas, 2019a). In the context of our AWS example, finding new political equilibria that can support global restrictions on AWS might become more urgent than reconceptualising 'autonomy'.

Third, reconfigurations between the different regulatory modalities may shift *the very operational foundations* (the 'wedge' or 'contact point') of governance strategies wholesale, by (further) *displacing* the primacy of concrete 'law' in regulating behaviour or in aiming to pursue strategies. For instance, it has been explored how AI systems enable regulation through 'microdirectives' (Sheppard, 2018) and 'technological management' (Brownsword, 2015, 2016, 2019b), potentially feeding into systems of 'algocracy' (Danaher, 2016). AI also facilitates behavioural manipulation through 'hypernudging' (Yeung, 2017, 2018), or invisible influences embedded in AI-mediated adaptive choice architectures (Susser, 2019). Likewise on global level, the use of AI systems might well help alter the processes by which international law or broader global governance is produced or enforced (Deeks, 2020; Maas, 2019a). While such developments may themselves give rise to concern, and therefore may become the object of (long-term) governance strategies themselves, such shifts should simultaneously be examined insofar as they alter the operating parameters—including what Deudney (2018) describes as 'material-contextual' factors, and what van Assche et al. (2020) describe as the 'material' dependencies of the human-made environment.

As such, even if AWS challenges were to be 'solved' within Levels 0–1, a general retrenchment in the relative regulatory strength of 'normative' (international) law relative to unilateral technological tools (Brownsword, 2018) may neutralise these gains. For instance, where AI mediates or undercuts meaningful human decision-making, global AWS regulations anchored in maintaining 'meaningful human control' over narrowly defined 'autonomous' weapons systems, will overlook AI's behavioural influences, 'automation bias', and propensity to 'normal accidents' (Borrie, 2016; Carvin, 2017; Maas, 2018; Scharre, 2016a). In such cases, human agents remain nominally engaged in lethal decision-making, but in fact are consigned to the 'moral crumple zone' (Elish, 2016).

Fourth, AI can change core *values* (Danaher, 2018) and even fundamental *rights* (Liu, 2019b; Liu & Zawieska, 2020), both altering the yardstick against which we measure impact of AI and potentially defusing our means (or indeed motive) of resistance. In our AWS example, if military AI applications enable a pivot towards conflict prediction and pre-emption (De Spiegeleire, Maas, & Sweijs, 2017), the shift from high-casualty drone strikes towards 'invisible wars' that subtly enable the prediction and interdiction of 'enemies' (Deeks, 2018) might lessen the reputational penalties from waging high-tech wars. Yet, the corresponding shift to a paradigm that heightens scrutiny or accountability of such 'hidden violence' (Kahn, 2002) may not develop in time, or to the same intensity. Finally, the fact that technological change can 'reveal rights' (Parker & Danks, 2019) can demonstrate shortcomings in the existing human rights umbrella. This limits human rights' ability to 'push-back' effectively against AI: in our AWS example, attempting to fit the wrongs precipitated by AWS as claims made through the existing human rights framework reveals severe shortcomings that structurally understate the injury (Liu, 2018, 2019c).

In sum, whereas Level 0–1 work often takes for granted the scope and nature of the problems at hand, a Level 2 analysis investigates how the technology, its uses, its indirect effects, and the direct problems, may aggregrate into overarching trends that can *change the terms* of governance. It can as such reckon with 'out of bound' strategic barriers or deflectors (changes to the problem portfolio; to regulators' goals; to governance tools, or to societal values) that would completely surprise and change the terms of analyses at Levels 0–1—potentially rendering any solutions arrived at as superfluous, contrary to the new regulatory goals, or out of step with the nature and values of the society they are meant to serve. Simultaneously, analysis at this level can reckon with these 'medium-term' disruptors that might frustrate/undercut governance efforts that only take account of Level 3 destinations.

3.4. Level 3 – charting macrostrategic trajectories & destinations

Finally, the third level reframes the AI policy-debates at Levels 0–2 through the lens of foundational axiological and 'macrostrategic' (Bostrom, 2016) questions about what kind of worlds we do and do not want to reach with AI. Work at this level examines (I) *analytically*, how the injection of AI systems (or indeed other technologies) might inadvertently alter the trajectory of human society in the mid- to long-term (Brundage, 2018; Baum et al., 2019); and it examines (II) *strategically,* how we might identify or constitute inflection points and opportunities for intervention for re-directing that trajectory towards preferable worlds.

Some work to date has focused on mapping (technological) vectors that could alter the macrostrategic trajectory. In most cases, such work has focused on (a) *terminal trajectories,* the most extreme scenarios whereby technogenic 'turbo-change' gradually or suddenly leads society towards global catastrophes or even existential risks (Moynihan, 2020). Such work has explored certain scenarios of low—or unknown—probability but high impact, such as the catastrophic or even existential threat to humanity that future AI systems might pose if not well-aligned with human values (Bostrom, 2014b; Everitt, Lea, & Hutter, 2018; Russell, 2019; Yudkowsky, 2008a). In such cases, the world might get many 'near-term' policy issues at Levels 0–2 right (for example, such as achieving a Level 0 ban on AWS; altering notions of liability to account for unpredictable AI systems at Level 1; or putting in principles around the shift towards regulation by 'technological management' at Level 2)—yet in the long-term still see this come to naught.

But the space of possible macrostrategic destinations surely extends beyond this. Indeed, recent Level 3 thinking therefore has explored *(b) vulnerable worlds* which concern ways in which worlds that are still 'safe' today may nonetheless be set on a trajectory within which continued technological progress inexorably ensures that, as Bostrom (2019, 3) predicts, "a set of capabilities will at some point be attained that make the devastation of civilization extremely likely, unless civilization sufficiently exits the semianarchic

default condition." Related to this, other scholarship has extended this further, by exploring the ways in which diverse, complex, path-dependent socio-technological trends, which in isolation do not rise to an 'existential risk', might converge and interact in ways that gradually, but steadily and irreversibly, increase the systemic *vulnerability* of our societies (Kuhlemann, 2018; Liu, Lauta, & Maas, 2018). This could include the accumulating effects of certain technologies—as with industry-fuelled climate change, a trajectory where a "great many actors face incentives to take some slightly damaging action such that the combined effect of those actions is civilizational devastation" (Bostrom, 2019, 7). It could also involve intersecting or compounding effects of different technological disasters, such as the climactic 'termination shock' that would follow if future global geo-engineering programs were to be suddenly interrupted as a result of regional (nuclear) war (Baum, Maher, & Haqq-Misra, 2013).

A further variation on this theme examines trajectories towards *(c) exposed worlds*. Whereas vulnerable worlds can involve disastrous scenarios as a result of many small accumulating or interacting failures, errors, or hazards, an exposed world—or what some have called a 'fragile world' (Manheim, 2018)—is one that makes certain choices that have rendered it susceptible to previously-modest shocks or hazards, including hazards that society previously might have weathered much better. Under some scenarios, technologies such as AI might lead us to more exposed worlds, insofar as larger parts of our economy and knowledge base become tied up in certain to-us opaque infrastructures dependent on global connectivity and constant and reliable electricity supplies.

Alternatively, AI could inadvertently lead societies into *(d) bad worlds*: 'progress traps' (Wright, 2005) or 'inadequate equilibria' (Yudkowsky, 2017) where society reaches a future destination where it is not destroyed, nor necessarily vulnerable or exposed, but where it is *bad* and (nearly) *inescapable* in the sense that subsequent recovery out of this state towards better trajectories is no longer possible. Even if these effects are not absolutely 'intrinsic' to AI technologies, it may still be that many of the uses to which AI is put within our society, nonetheless could: imperil human dignity (Brownsword, 2017); erode the 'social suite' and relational capacities of humans, in ways that drive creeping negative 'social spillovers', such as a diminished ability to cooperate (Christakis, 2019); or increasing algorithmic 'inscrutability' may lead us to a 'new dark age' of uncertainty, surveillance, or the death of sociality or empathy (Bridle, 2018). Such long-term trajectories towards 'flawed realization' (Bostrom, 2013, 19) might emerge if AI systems would have long-term and lasting impacts on our world and society which would (on net) be considered ethically adverse or at least sub-optimal across a broad range of ethical views.

Finally, while these are sometimes under-examined relative to these other scenarios, it is also important for Level 3 scholarship to orient itself towards *(e) Good worlds.* These are axiological descriptions of certain societal destinations—or trajectories—that do not merely avoid these above enumerated risks, but articulate clear principles or 'desiderata' (Bostrom, Dafoe, & Flynn, 2019) for what good society we would like to pursue or societal trajectory we would like to be set on. Of course, in one sense, perspectives that articulate how we want to proceed ('journey' or 'trajectory') or where we want to end up in ('destination') in the long term, are both alternative and complementary strategies for avoiding 'bad worlds' or other pitfalls in development. These should invoke different types of assessment. The (troubled) narratives of utopia versus dystopia seem to consider macrostrategic destinations in their exploration of contours and configurations of established worlds, whereas ethical quandaries seem to relate to macrostrategic trajectories insofar as their consideration or their outcomes affect pivotal points in pursuit of a path ahead. At the very least, setting out the parameters of 'good worlds' provides a benchmark to assess whether or not we are either on the path to a bad world, or are actually in some sense in a bad world already. This would be useful since we presently seem to lack objective points of reference to indicate what type of world we are in. It should, however, be noted that such shared positive perspectives for long-term societal macrostrategy remain relatively underdeveloped—especially in terms of linking them to shared narratives and visions.

The insight of Level 3 problem-finding scholarship is that, even if we managed to solve the discrete AI policy-problems—such as the global regulation of military AI—at lower levels, this reactive, fragmented firefighting approach may simply not suffice in aggregate to shift societies out of overall-captive trajectories towards the 'attractor states' of vulnerable, fragile or bad worlds. Rather, AI might drive long-term sociotechnical effects or shifts in constitutional societal values, shifts which might be passively unobserved or even unobservable (cf. Cirkovic, Bostrom, & Sandberg, 2010) by society, or which might even be actively veiled by certain interested parties, until it is too late to marshal resistance or change. The point is that AI can lock in such dependencies in ways that even the broad analyses at Level 2 may not fully appreciate, missing the forest for the trees. At the same time, Level 3 work highlights positive leverage points which problem-oriented work at Level 0–2 might pass by: opportunities where AI itself opens up loci for positive interventions that can shift this trajectory—promoting beneficial applications of 'AI for global good' (Cave & Ó hÉigeartaigh, 2018) which empower human autonomy, or help consolidate a 'postwork utopia' (Danaher, 2017, 2019a). In this way, we emphasise the importance of articulating governance lodestars; to pivot towards a prospective, teleological approach that sketches a spectrum of societal trajectories that not merely avoids catastrophic outcomes (Bostrom, 2013, 2019), but which articulates what we seek to gain—and where we want to go. If developed, such work would highlight a potential new chapter for law and regulation, ensuring governance can speak to both the perils and promise of AI.

4. Reflections on problem-finding AI governance strategies

The above gives an analysis of four governance orientations, and their implications for constituting a long-term governance approach to the problems created by AI. We therefore will now provide four meditations on some of the nuances of our model; as such we will explore (a) the interrelation and mutual complementarity of governance across the four levels; (b) the potential reapplication of problem-finding governance research (Level 2 and 3) in a problem-solving mode; (c) how all four strategies are themselves merely situated within a larger 'Governance Goldilocks Zone' that circumscribes and articulates the very limits of what can be meaningfully governed; and (d) the interrelationship between governance strategies at these four levels, and broader (self-fulfilling) narratives or visions.

4.1. Coordinating strategies across levels: from partial failures to co-evolution

We want to emphasize that the point of the four-level framework is not to discard 'low-level' scholarship or governance, and fully replace it with 'high-level' scholarship. Indeed, in one sense, we should be careful to how we understand the distinction between these four levels; these are primarily mapped to the different ways in which governance strategies are oriented, and which (problem-; legal-; and societal) parameters they takes for granted.

However, while this distinction between the four levels—and indeed, the terminology of 'levels'—seems to imply an ascending hierarchy or scale, that linkage should not be drawn too far. That is, it may be tempting to map the axis of problem-solving to problem-finding governance to many other scales, such as a distinction between 'near term' vs. 'long term' governance, 'specific- 'vs. 'general' governance, or 'small scale' vs. 'large scale' issues. Yet while these other scales may show some correlation to our fourfold distinction, they should not be drawn too widely: it is after all possible (if not necessarily widespread) to adopt a Level 0 or Level 1 governance approach to problems that will likely only emerge in the 'long-term future' (as seen in the debates over legal personhood for robots). Conversely, it is also possible to adopt a Level 3 perspective which does not seek to take in the full spectrum of societal and technological activity, but which narrowly focuses in on one specific topic (e.g. 'over the next decade, will AI progress be more constrained by computing hardware or by data?') which is held to be a crucial consideration or pivot point for determining the longer term trajectory. As such, while the four Levels correspond somewhat to other scales, they should not be taken as equivalent.[10]

More practically, the point of our framework here is not to argue for the categorical superiority of problem-finding approaches to old problem-solving 'dogma'. Indeed, we must beware not to replace the particular failures of a governance system that is almost solely problem-*solving* with the particular shortfalls of another governance system that is almost entirely problem-*finding*. Rather, the point is to use problem-finding scholarship to complement and build beyond the existing problem-solving work, while understanding how all these four strategic orientations are necessary and *complementary*—and how work must range and coordinate across them.

In particular, we want to emphasise that each governance system has to find its own form of complementarity, which balances problem-solving with problem-finding strategies. This is a balance that can differ per governance domain depending on local conditions and interactions. To illustrate this, we now examine three types of common failure modes in AI governance which takes stock of only one or another approach, and which thereby risk failing to 'governing for the long-term': (a) reactive work that is restricted to levels 0–1, and therefore *does not even try* to anticipate the long-term; (b) work that examines Level 2 governance disruptors, but which ignores Level 3 macrostrategic destinations, and which therefore *misses the forest for the trees*; (c) work that considers Level 3 trajectories, but ignores Level 2 disruptors, and therefore risks becoming a set of *grand futures, derailed by detail*.

In the first place, scholarship that only focuses on Level 0 is reactive and focused on solving direct problems in the present. In that sense, it is *not even trying* to pursue governance solutions that are scalable into the long-term. As discussed above, such approaches might have some limited uses, but as discussed above, they risk being blindsided both temporally and sectorally. In relation to Level 3 macrostrategic trajectories, this approach seems inadequate. It is reactive, a holding action at best, and does not reflect on longer-term trajectories or dependencies, meaning that we would have to be very 'lucky' to find ourselves in 'good worlds' through Level 0 work alone.

In the second place, work that works up to Level 2, but does not consider the Level 3 macrostrategic destination, risks *not seeing the forest for the trees*. In that way, it may manage to identify and eventually address many cross-sectoral problems, but over time could still leave us stuck in a potential progress trap. That is not to say that such work is 'senseless', but rather that it too needs to be grounded in a picture of how constitutional shifts at distinct levels—and their associated governance strategies—cohere and converge into a long-term governance trajectories.

Thirdly, and conversely, any work that focuses only on Level 3, but which tries to solely reason 'down' to the other levels, runs the risk of ending up as a set of *grand futures, derailed by detail*. Specifically, by ignoring Level 2 disruptors, such strategies may make foundational assumptions—about the available or effective instruments of law; about the values of policymakers and societies—which are on track to get sidelined by Level 2 eddies and deflecting vectors even in the medium term. For example, some proposals to govern and avert potential future catastrophic risks arising from advanced artificial intelligence systems have envisioned governance strategies grounded in a comprehensive inter-state global treaty regime (cf. Wilson, 2013) or housed within the United Nations (Castel & Castel, 2016; Nindler, 2019; However see Cihon, Maas, & Kemp, 2020). While it is certainly valuable to explore all avenues—certainly historically proven ones—the risk is that such a move constitutes an attempted leap, from a Level 3 macrostrategic goal (e.g. a catastrophic trajectory to avoid), immediately down to possible Level 0 or Level 1 solutions ('global treaties') that are patterned on the solution package perceived to be available or the norm today. In doing so, however, the risk is that such proposals fail to engage with key Level 2 insights. For instance, (a) some have argued traditional international law instruments based on state consent are very poorly equipped to engage extreme but unknown catastrophic risks (van Aaken, 2016; though see Vöneky, 2018); (b) in recent decades, the parameters of global governance have already been shifting, with many arguing that it has trended away from formal international law-making, towards diverse 'regime complexes' made up of heterogeneous actors and informal governance arrangements (Alter & Raustiala, 2018; Morin et al., 2019; Pauwelyn, Wessel, & Wouters., 2014); (c) such proposals may not sufficiently engage with insights of how and where the deployment of AI itself may affect or erode the political scaffolding or legitimacy of 'hard' international law itself (Deeks, 2020; Maas, 2019a). Any of these might constitute a 'crucial consideration' against the project of trying to secure the Level 3 macrostrategic objective of avoiding disastrous long-term trajectories by attempted Level 0 or Level 1 tools (i.e.

[10] We thank Henrik Palmer Olsen for spurring this line of thought.

treaties) alone.

Ultimately, the aim is not to elevate some of these strategies over others, but instead to help provide an inter-community framework or 'translation zone', that can help scholars situate themselves within scholarship at different levels, and to translate their work across to other communities as well as policymakers and the broader network of governance actors. Such an approach is critical also to fully levarage both the divergent 'scanning' function of problem-finding scholarship at Levels 2–3, as well as the convergent 'rallying' function of inviting groups and actors behind preferable Level 3 trajectories, in ways that are cognizant of possible Level 2 'governance disruptors', as well as the intricacies of concrete Level 0–1 policy implementation.

4.2. Problem-finding strategies applied in problem-solving mode

Having set out the interrelation and complementarity of problem-solving and problem-finding approaches across the four levels, we should now make a special observation with regards to the ways in which approaches at these four levels can reflect- or adapt to one another. After all, one reading of our argument above would be that while we have reservations about a problem-solving governance approach pursued in isolation, they can work alongside problem-finding approaches, because these two frames generate a sufficient complementarity. In another perspective, however, one might argue that even the problem-finding paradigms (Levels 2–3) can involve (or may require) 'problem-solving' features—but that these are different to the isolated (Level 0 or Level 1) versions of the problem-solving approach.[11] This second interpretation has much to recommend it.

As noted above, for instance, Level 3 (problem-finding macrostrategic) scholarship on AI can be understood to include an *analytical* mode, which explores chiefly how the injection of AI systems (or indeed other technologies) may inadvertently alter the trajectory of human society in the mid- to long-term. This type of analysis can be purely problem-finding. It is important, however, to note that much work in this paradigm also involves an *interventional* mode: such work does not merely seek to 'find' previously unseen but crucial dangers; but it also aims to explore how we might identify or create governance inflection points and opportunities for intervention, to re-direct the global trajectory towards preferable (i.e. 'good') worlds. We should therefore draw a distinction between Level 3 work that aims to *find* new crucial considerations, and Level 3 work that aims to *solve* these challenges through interventions or inflection points.

Paradoxically, this latter approach therefore appears to revive aspects of problem-solving analysis in Level 3 work. Nonetheless, the type of problem-solving here is frequently distinct from Level 0 or Level 1 problem-solving, because it need not be as closely wedded to pre-existing solution sets. That is, while it admits that any new problems (such as 'crucial considerations') excavated by problem-finding approaches (Levels 2–3) could be amenable to problem-solving approaches, it does not automatically presume that existing responses are necessarily adequate or superior to alternate governance responses. Nonetheless, the interrelation of the problem-solving and problem-finding paradigms in (AI) governance, and the way they can shape, alter and inform one another, remains a key nuance to this framework, which remains to be further explored.

4.3. A goldilocks zone for governance? Beyond the four-level framework

Moreover, while we distinguish four strategic levels in our governance framework, it is worth stepping back and taking stock of the broader long term implications that this framework raises, particularly in terms of how governance (of any form) categorically relates to differential, technology-driven changes.[12] The animating question here is whether there is something to be found beyond these four levels? That is, these four levels of problem-solving and problem-finding governance strategies may together describe (or cover; or constitute) the governance landscape in a complete and comprehensive manner. Yet the variety amongst them—and the lower and upper 'limits' of Level-0 and Level-3 governance strategies respectively—imply that these four levels of governance are in a sense themselves confined within a 'Goldilocks Zone' for governance. The 'Goldilocks Zone' is the informal astronomical term for the 'circumstellar habitable zone'—that narrow region around a star where the surface temperature on some planetary bodies is "just right" for water to be present in the liquid phase, implying they could support the emergence of life (NASA, 2020). To extend this analogy, we might then come to think of a 'Governance Goldilocks Zone' as that narrow band of core parameters, assumptions, or boundary conditions for problem 'governability' within which the very project of governance (of either a problem-solving or problem-finding variety) is capable of residing or thriving in the first place.

Within this broader context, we can recognize how one can only 'do governance' (of any type or sort) under assured conditions of *autonomy* (freedom to determine behaviour and to act) and *influence* (behaviours and actions are causally connected to outcomes), and these constitute the underlying presumptions of our four level framework.[13] The long-term governance perspective, however, enables us to step outside of these presumptions in order to understand the implications of a Governance Goldilocks Zone.

That is, to draw another analogy (and at the risk of mixing metaphors), we might allude to the fundamental states of matter to illustrate the character of the problem landscape beyond this narrow Governance Goldilocks Zone. One might accordingly think of the Levels as corresponding to varying states of matter at distinct temperatures. In this heuristic, one could see Level -1 governance as

[11] We thank Roger Brownsword for spurring this line of thought.

[12] We thank Victoria Sobocki for discussions prompting this section.

[13] This has similarities to Roger Brownsword's account of the core 'existence conditions' for human life (such as the maintenance of the core infrastructure conditions), and the generic conditions for agency, which he holds to be the 'first regulatory responsibility', and a 'regulatory red line' (Brownsword, 2019a, 90–95).

corresponding to a Bose-Einstein Condensate (BEC) (a state of matter that occurs at extremely low temperatures near absolute zero, where molecular motion nearly stops and atoms begin to clump together); Level 0 would be a solid state; Level 1 a liquid state; Levels 2–3 the gaseous state; and Level 4 the plasma state. These suggest varying levels of particle movement and intuitive grasp of behaviour, but for the purposes of this section, we focus on the Bose-Einstein condensate and the plasma that bookend the fundamental states of matter. This because the Bose-Einstein condensate state provides an example of situations where change is difficult, and the plasma state presents an example as to the difficulties inherent for intuitive predictability of behaviour that lie beyond our daily experience of matter.

4.3.1. Level '-1': governance BEC: stasis in reality or in governance responses

Stretching backwards into the space 'beneath' Level 0 would imply *stasis*, both actual and perceived, with both forms of stasis converging to foreclose the impetus or opportunity for altering the existing governance landscape. Thus, at Level 'negative-1' (-1),[14] we have stepped outside of a governance framework that can be responsive and adaptive to (technologically induced) change, either because there is no actual change (*stasis in reality*), or no perceived or recognised change (*stasis in governance responses*) in the socio-otechnical landscape. Of course, no society has ever been in a perfect form of internal stasis—let alone a form of stasis resilient to any outside shocks. Nonetheless, some soft forms of these conditions might have historically applied to (possibly pre-industrial revolution) societies, which did undertake investigations into the governance implications of (socio)technical change, perhaps because no such change was easily perceptible or in memory. Under these conditions, not even Level 0 'business-as-usual' governance is engaged, because no new (sociotechnical) situations (appear to) present themselves for examination.

This static nature of Level -1 conditions results in governance self-entrenchment, because the very possibility of subsequent change (in governance) is systemically resisted: there are no new 'problems' that are presented for Level 0–1 approaches to 'solve'; and there is no background change that can provide an easy seed, impetus, or justification for Level 2–3 approaches to go out and 'find' potential challenges. In the course of discarding a dynamic view of governance, where its aims, objectives, values, methods, and actors are in constant and mutual feedback, Level -1 amounts in a sense to a condition of 'absolute zero', where there is no movement (at least from within the system) and therefore no prospect for interaction, recombination, or phase transitions.[15] There are no moments of 'legal disruption' or uncertainty (Liu et al., 2020), and as such Level -1 undercuts the assumption of governance autonomy by removing the perception or reality that there is change, and it can thus be perilous from a govenance perspective because alternatives are no longer imagined, perceived or pursued—let alone actualised. It lies outside the governance Goldilocks zone, because it is 'uninhabitable' for governance initiatives—providing no 'sustenance' or activation energy.

4.3.2. Level '4': 'plasma' governance beyond sight, comprehension, or control

At the other end of the scale, we can consider the space extending beyond Level 3 as analogous to plasma, a fundamental state of matter that is often misunderstood and difficult to intuit from a perspective that is more familiar with the more quotidian states of matter. In the context of long term governance, Level 4 appears so deeply unfathomable as to defy conceptualisation from our present standpoint. It suggests the minimum requirement to reconsider our contemporary axiological configuration (Danaher, 2020), and questions the desirability of our present notions, techniques and objectives for governance (Bostrom et al., 2019).

Indeed, at Level 4, it may be that our attempts at 'doing governance' may range from suboptimal through to being downright harmful over the long-term, because opportunities and prospects are opened up in this space that is currently beyond our ability to comprehend or fathom. From such a long-term governance perspective, Level 4 suggests that 'doing governance' might not only be futile, but may also be both counter-productive.

The 'ingovernability' of Level 4 problems—that is, the features which drive a certain problem outside the governance Goldilocks Zone—may derive from three complementary sources, each of which is individually sufficient to erode or even foreclose our potential to 'do governance': (1) we *cannot see* governance problems; (2) we *cannot comprehend* the governance problems; (3) we fundamentally *cannot control* the governance problems (with today's tools). Respectively, these factors range from primarily undermining autonomy towards undercutting influence although of course these factors comprise of a mix of these two dimensions.

With regards to (1) governance problems that *we cannot see*, there are a wide range of blinkers and veils that prevent us from seeing clearly all potential threats. Generally speaking, accurate judgments of catastrophic shocks are beset by various cognitive and epistemic biases (Yudkowsky, 2008b), making their governance subject to a 'tragedy of the uncommons' (Wiener, 2016). In extreme cases, the absence of observable or recognized precedent around truly existential disasters can create an 'anthropic shadow' (Cirkovic et al., 2010),[16] which can exacerbate the effects of our 'availability bias', further restricting our ability to see what might really be at stake. As Cirkovic, Sandberg, and Bostrom put it: 'the observation selection effect implicit in conditioning on our present existence prevents us from sharply discerning magnitudes of extreme risks close (in both temporal and evolutionary terms) to us' (2010, 1500). Even where we might perceive certain threats in the abstract, the intangible and distributed nature of certain global processes (such as

[14] It should be emphasized that by calling these categories 'Levels', we are not suggesting that they lie within the same continuous 'problem-finding' framework.

[15] Of course, to extend the analogy (or recognize some of its limits); there are of course materials which achieve unusual properties (e.g. super-conductivity) at extremely low temperatures. Do governance approaches achieve such states as they crystallize below Level 0? We leave that question for future work.

[16] 'Anthropic bias can be understood as a form of sampling bias, in which the sample of observed events is not representative of the universe of all events, but only representative of a set of events compatible with the existence of suitably positioned observers' (Cirkovic et al., 2010, 1495–96).

climate change) can make them 'hyperobjects' (Morton, 2013), complicating our apprehension of these problems (or the appropriate levers we might use to affect them). While these constitute the prominent examples, it becomes obvious that the prospect for long-term governance is severely eroded under conditions of invisibility. Because these governance problems lie beyond the epistemic 'event horizon', they do not prompt problem-solving investigations; and even problem-finding investigations may only hit on them 'by chance'—as theoretical possibilities discovered in the course of exploring some (adjacent) other problem. While Level 3 problem-finding work may sometimes help at spotting such challenges, for lack of a 'signal', many of them may remain governance 'dark matter' that is structurally outside of the Governance Goldilocks Zone.

Our *lack of comprehension* under (2) merges into such conditions of opacity, mirroring the distinction drawn between sensation and perception on the one end, and our inability to control outcomes under (3). For instance, John Danaher has anticipated that algorithms may contribute to a coming era of 'Techno-Superstition', which combines opacity (a growing lack of public understanding of how the world—or an AI system—actually works) with the unwarranted illusion of control over AI systems (Danaher, 2019b). Within this context, we may appreciate that there is a problem (posed by AI), yet our inability to properly *comprehend* it may place us within the equivalent of a 'governance black box' which, Danaher proposes, we find ourselves in, in relation to AI systems. Thus, the arguments here are the same as those Danaher presents in relation to techno-superstition: lack of understanding, illusion or control, erosion of achievement, loss of autonomy, and undermining of human agency. Applied to Level 4, these arguments would suggest that we are not able to 'do governance' in a meaningful manner under such conditions.

Finally, (3) with regards to (AI; or general) problems that are either intrinsically or currently deeply *beyond our control,* problems are again outside the Governability Zone. There are at least two ways we might lack 'control' over a governance problem: in the first case, one might imagine a case where a certain problem is in principle governable, but all power and control have become centralised to a single actor in a perfect autocracy, thereby locking out participation by all others. In this case, if the 'singleton' actor is uninterested in addressing a problem, governance would have no purchase either. In the second, more common case, the challenge might either be one fundamentally beyond our present technological means[17] ; it might require a level of perfect collaboration and cooperation that is beyond our (present) political means; the prospect for control has either been lost, or is recognised as not existing.[18] The common denominator is the loss of control which shuts out the very possibility to 'do governance', either through the loss of participatory input in the first form or the disconnection between action and outcome in the second form.

Taken together, these suggest that the governance strategies with which we are familiar, and which we have mapped to Levels 0–3, may in fact reside in a Goldilocks Governance Zone, which circumscribes the limits within which it is possible to meaningfully contemplate or enact any strategic long-term governance.[19] As the underlying presumptions of autonomy and influence fall away at either end, be it the stasis of Level -1, or the unfathomable invisibility, incomprehensibility or impotence that characterise Level 4, the Goldilocks Governance Zone comes into view and becomes demarcated. Yet, without attempting to define the four strategic levels for long-term governance, identifying such a sweetspot would have been a difficult endeavour.

4.4. Strategies and narratives

While we make a strong claim in support of the complementary (and under-appreciated) value of problem-finding scholarship to such long-term governance scholarship, the framework we have proposed here is by no means a definitive one, and there will be many directions within which it can be developed and enriched further. To highlight but one point of potential interest, along with governance strategies, it will be key to examine how overarching narratives and encompassing long-term visions which frame strategies, can percolate across the four levels. There is already extensive scholarship on the role, dynamics, and risks of imagined futures (Beckert, 2016) or of 'sociotechnical imaginaries' (Jasanoff & Kim, 2015) with which medium-term technological projects such as 'smart cities' are often imbued (Sadowski & Bendor, 2019). Likewise, there are a host of emerging Level 3 descriptive or aspirational narratives revolving around our societal macro-strategic societal state, trajectory, or destination—whether 'turbochange' (Deudney, 2018), the 'great challenges' framework (Torres, 2018), a 'vulnerable world' (Bostrom, 2019), or 'existential security' (Sears, 2020)—which provide kernels or seeds of various (possibly contradictory or competing) long-term perspectives and projects.

This raises three questions: in the first place, how are such *long-term narratives embedded and manifested* in Level 0 to Level 2 partial strategies? In the second place, how do these *narratives relate to, and exert political effects*, alongside pre-existing 'utopian' or 'dystopian' visions, ideals or narratives of the future (Berenskoetter, 2011)? Finally, in the third place, are there ways in which such narratives (or

[17] For example, while we may harden the resilience of our societal infrastructures (e.g. electrical grid) to solar flares, we do not have the technical capabilities to affect the solar mechanics in order to reduce their prominence.

[18] For example, by taking complex adaptive systems seriously and recognising the ramifications that actions and inputs are largely untethered from producing the intended outcomes over the span of long-term governance perspectives. Alternatively, the viewpoint advanced by Nassim Nicholas Taleb over the course of *Incerto* could also be applied to suffuse randomness and uncertainty into processes upon which we project causation and the illusion of control.

[19] That is not to say that these conditions are rigid or fixed. To exploit the habitability analogy of the Goldilock Zone further, it is worth noting (a) that not all areas on earth fall within the habitable zone, as there are extremes in environmental conditions (e.g. volcanoes) which restrict (at least human) habitation; and (b) that we are technologically capable of affording human habitation beyond the natural parameters of these zones (including in outer space) through life support systems. Drawing on this analogy in terms of governability, we might suspect (a) that governance (and problem 'governability') is not uniformly distributed within Levels 0−3 of our governance framework, and that there might indeed be pockets of 'ungovernability' within these zones; and (b) even the external boundaries (e.g. the 'ceiling' of Level 3 governance) are not fixed, and using institutional or technological means, it could be possible to extend the zones of governability beyond the past parameters.

their policies or analysis at different levels) can have '*self-realizing*' effects? The role of 'self-fulfilling prophecies' has a long pedigree in scholarship, particularly where it concerns the 'autogenetic effects' of social predictions or utopian and dystopian narratives (Maas, 2012). Self-fulfilling and self-negating prophecies have received study in economics (Felin & Foss, 2009), as well as in international relations (Houghton, 2009), where scholars have charted the ways in which influential theories—such as the the 'democratic peace' and 'commercial peace' theses, or the 'clash of civilizations' (Bottici & Challand, 2006; see also memorably Tipson, 1997)—might end up self-fulfilling. Beyond the social realm, however, there might also be ways in which technological 'predictions', such as Moore's Law (Mollick, 2006) or current predictions of technological unemployment, become self-fulfilling (Khurana, 2019). This will be one key dynamic to be considered in both our problem-finding framework, as well as long-term governance more broadly.

To conclude, we hope our framework can contribute to the co-evolution of governance actors (and scholars) operating within different 'Levels', as well as the coordination and unification of 'partial strategies' at these levels, into more deliberate, pluriform, and cross-domain approaches that leverage the best from both problem-solving and problem-finding.

5. Concluding thoughts

Contemporary society continues to experience significant changes to the scale, pace, and type of change in both its social, natural, and technological environment. This suggests the need to define, develop, and deploy complementary governance strategies both to meet these challenges, and to build towards more positive futures, which can take long-term perspectives into account while leveraging concrete and operationalized policies today.

In this paper, we have identified some of the shortcomings of contemporary approaches to both the study and practice of governance, which often—explicitly or implicitly—take a 'problem-solving' approach. Specifically, we argue that problem-solving approaches to complex problems (such as AI) are limited, because of (1) the unreflexive over-reliance on narrow *metaphor* in the face of complex phenomena; (2) the *tautological* manner in which existing governance lenses identify and prioritize primarily those problems that are already legible to them; (3) the constrained *path-dependency*; (4) the inability to clearly *distinguish* between symptomatic and root-cause problems; (5) the promotion of a *false sense of security* through a rapid delineation of the problem-space that prematurely closes-off large parts of the governance solution-space. We as such argued that while such 'problem-*solving*' approaches may be necessary, but are also insufficient in the face of many of the 'wicked problems' which we face, including in the realm of AI.

By contrast, we have made the case that exploratory, problem-*finding* research is critical to long-term policy and governance strategy endeavours. This is true for epistemic and scientific reasons relating to good and valid science and scholarship—because problem-*finding* work can (1) itself *produce foundational academic insights* at the interstitices of different problem domains, or at different scales of analysis; (2) it can work in a productive *reflexive equilibrium* with problem-solving scholars and strategy-makers; because (3) it can at times *reveal new strategic 'crucial considerations'* which might completely alter the balance of local-context strategy considerations; and because (4) it can enable us to *forestall the 'power' problem inherent* to the Collingridge Dilemma of technology governance.

Beyond these epistemic virtues, the proactive and explorative orientation of the problem-finding approach is crucial to any and all conceptions of the long-term perspectives, whether these are conceived as 'early warning radar', or as 'inspirational lodestar'. Accordingly, we have proposed and elucidated a rough typology for four distinct 'levels' at which one can study or formulate governance strategies: at Levels (0) '*Business As Usual*' governance and (1) '*Governance Puzzle-Solving*', work takes a predominantly problem-solving approach to issues. Conversely, problem-finding approaches focus on (2) '*Governance Disruptor-Finding*' and (3) '*Charting Macrostrategic Trajectories*'.

While within this paper, we have primarily discussed specific examples involving AI governance strategies, problem-finding approaches would likely be equally crucial for many other cross-sectorial challenges—if not more crucial. Problem-solving approaches might gain traction with regard to the clear, discrete clusters of pre-defined problems that are thrown up in relation to AI governance, by nature of the technogenic origins of the challenges. Yet, intersectional challenges diminish the crispness of their concomitant problem-sets. The result is that there may be uncertainty or ambiguity concerning how problem-solving approaches could effectively go about confronting such emergent, complex, and interconnected global challenges, without arbitrarily confining the metaphors and models in order to manufacture the necessary problem-clusters to be tackled. Thus, while problem-solving approaches may already be strained when confronted with polymorphous policy domains such as those relating to AI governance, the efficacy of problem-solving approaches may fade further where long-term governance takes multiple interacting developments into view. Long-term perspectives trade in future worlds, themselves intricate complexes of emerging challenges and possibilities of which factors such as AI and its governance comprise a minuscule subset. Thus, attempts to grapple with long-term perspectives must necessarily involve problem-finding approaches at their very core because these futures remain mysteries. In slogan form: we need to collect, catalogue, and categorise the challenges of long-term governance before we can set about answering concrete questions that are raised.

We opened this paper with an emphasis upon the notion of *change*: an apparently paradoxical observation that change is the only constant, while simultaneously itself being subjected to change. Stepping back from AI, and long-term governance – what does it mean for change to be changing?

If we are right that the very possibility of doing governance requires change (that is, a guarantee that we are not at Level -1), then change undergirds the prospect, potential and promise of governance itself. Yet, it is clear that change is not one-dimensional, instead varying according to the *rate*, *scale*, *direction*, and *duration*, of the change in question: these can be further combined to constitute gradual, disruptive change, and paradigmatic change, for example. The question of change can also be addressed through an agent perspective to focus on who is driving or inhibiting the change in question; conversely a patient approach would focus upon the redistribution of benefits and burdens in the process of aftermath of that change. Furthermore, the relativity inherent within change

requires considerations of benchmarks, those concerning the perceived status quo and expectations for the future for example, and how developments diverge or converge to these. The prospect for change also raises the possibility for encountering attractor states, which will affect the nature of observed change as well as future change.

Leveraging the adjacent possibles opened up through the process of unpacking the four strategic levels for long-term governance, we also contextualised our framework within the broader spaces of (un)governability to suggest that our framework sketched out the Governance Goldilocks Zone within which the conditions are conducive to permitting us to still able to 'do governance' in the first place. Looking beyond the boundaries of our framework suggests directions for further research to explore and identfy the underlying presumptions that support the very prospect for our framework to come into play, and to the underexplored factors at play that enable us to project governance into the long-term.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We thank the editors and anonymous reviewers for their kind comments and recommendations. We also thank Roger Brownsword, Victoria Sobocki, and the students of the Fall 2020 'Artificial Intelligence and Legal Disruption' class at The University of Copenhagen for their engaging comments and discussion. Finally, we thank Helle Krunke and Henrik Palmer Olsen for key clarifying comments and questions during the final stage of refining this paper.

References

Alter, K. J., & Raustiala, K. (2018). The rise of international regime complexity. *Annual Review of Law and Social Science, 14*(1), 329–349. https://doi.org/10.1146/annurev-lawsocsci-101317-030830.

Anderson, K., Reisner, D., & Waxman, M. (2014). Adapting the law of armed conflict to autonomous weapon systems. *International Law Studies, US Naval War College, 90*, 386–411.

Bache, I., & Flinders, M. (2004). *Multi-level governance* (1st ed.). Oxford University Press.

Balkin, J. M. (2015). The path of robotics law. *California Law Review Circuit, 6*(June), 17.

Baum, S. D., Jr., Maher, & Haqq-Misra, J. (2013). Double catastrophe: intermittent stratospheric geoengineering induced by societal collapse. *Environmentalist, 33*(1), 168–180. https://doi.org/10.1007/s10669-012-9429-y.

Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., , … Maas, M. M., et al. (2019). Long-term trajectories of human civilization. *Foresight, 21*(1), 53–83. https://doi.org/10.1108/FS-04-2018-0037.

Bauman, C. W., Peter McGraw, A., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*(9), 536–554. https://doi.org/10.1111/spc3.12131.

Beckert, J. (2016). *Imagined futures: Fictional expectations and capitalist dynamics* (1st edition). Cambridge, Massachusetts: Harvard University Press.

Bennett Moses, Lyria (2007). Why have a theory of law and technological change? *Minnesota Journal of Law, Science & Technology, 8*(2), 589–606.

Berenskoetter, F. (2011). Reclaiming the vision thing: Constructivists as students of the future. *International Studies Quarterly, 55*(3), 647–668.

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., Feyereisl, J., et al. (2018). Mapping intelligence: Requirements and possibilities. In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence 2017* (pp. 117–135). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-96448-5_13. Studies in Applied Philosophy, Epistemology and Rational Ethics.

Boeglin, J. (2015). The costs of self-driving cars: Reconciling freedom and privacy with tort liability in autonomous vehicle regulation. *Yale Journal of Law and Technology, 17*, 171–204.

Borrie, J. (2016). *"Safety, unintentional risk and accidents in the weaponization of increasingly autonomous technologies." UNIDIR resources 5*. UNIDIR. http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf.

Bostrom, N. (2013). Existential risk prevention as a global priority. *Global Policy, 4*(1), 15–31.

Bostrom, N. (2016). *"'Macrostrategy.'" Presented at the Bank of England - "Macrostrategy," Moorgate Auditiorium, 20 Moorgate, London, April 26*. https://www.youtube.com/watch?v=f9HvMLSD0jo.

Bostrom, N. (2019). The vulnerable world hypothesis. *Gobal Policy*, (September)https://doi.org/10.1111/1758-5899.12718, 1758-5899.12718.

Bostrom, N. (2014a). Crucial considerations and wise philanthropy. In *Presented at the Good Done Right, All Souls College*. https://drive.google.com/file/d/0B4kMPIEI5Mb8LXQ1YUlOLUFwLUk/view?usp=drive_open&usp=embed_facebook.

Bostrom, N. (2014b). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N., & Cirkovic, M. M. (2008). *Global catastrophic risks* (1st edition). Oxford; New York: Oxford University Press.

Bostrom, N., Dafoe, A., & Flynn, C. (2019). Public policy and superintelligent AI: A vector field approach. In S. M. Liao (Ed.), *Ethics of artificial intelligence*. Oxford University Press. http://www.nickbostrom.com/papers/aipolicy.pdf.

Bottici, C., & Challand, B.ît (2006). Rethinking political myth: The clash of civilizations as a self-fulfilling prophecy. *European Journal of Social Theory, 9*, 315–336. https://doi.org/10.1177/1368431006065715.

Bridle, J. (2018). *New dark age: Technology and the end of the future*. London; Brooklyn, NY: Verso.

Brockman, J. (Ed.). (2019). *Possible minds: Twenty-five ways of looking at AI*. New York: Penguin Press.

Brownsword, R. (2015). In the year 2061: From law to technological management. *Law, Innovation and Technology, 7*(1), 1–51. https://doi.org/10.1080/17579961.2015.1052642.

Brownsword, R. (2016). Technological management and the rule of law. *Law, Innovation and Technology, 8*(1), 100–140. https://doi.org/10.1080/17579961.2016.1161891.

Brownsword, R. (2017). From erewhon to AlphaGo: For the sake of human dignity, should we destroy the machines? *Law, Innovation and Technology, 9*(1), 117–153. https://doi.org/10.1080/17579961.2017.1303927.

Brownsword, R. (2018). Law and technology: Two modes of disruption, three legal mind-sets, and the big picture of regulatory responsibilities. *Indian Journal of Law and Technology, 14*, 1–40.

Brownsword, R. (2019a). *Law, technology and society: Re-imagining the regulatory environment* (1st edition). Abingdon, Oxon; New York, NY: Routledge.

Brownsword, R. (2019b). Law disrupted, law re-imagined, law re-invented. *Technology and Regulation*, (May), 10–30. https://doi.org/10.26116/techreg.2019.002.

Brundage, M. (2018). Scaling up humanity: The case for conditional optimism about artificial intelligence. *Should we fear artificial intelligence?* European Parliamentary Research Service - Scientific Foresight Unit (STOA). http://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., , … Dafoe, A., et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. *ArXiv:1802.07228 [Cs]*, February http://arxiv.org/abs/1802.07228.
Buchanan, B. (2020). *The AI triad and what it means for national security strategy*. Center for Security and Emerging Technology. https://cset.georgetown.edu/research/the-ai-triad-and-what-it-means-for-national-security-strategy/.
Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science, 9*(1), 14. https://doi.org/10.1186/s40163-020-00123-8.
Calo, R. (2015). Robotics and the lessons of cyberlaw. *California Law Review, 103*, 513–564.
Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review, 51*(2), 37.
Carvin, S. (2017). *Normal autonomous accidents: What happens when killer robots fail?*. SSRN Scholarly Paper ID 3161446. Rochester, NY: Social Science Research Network https://papers.ssrn.com/abstract=3161446.
Casey, A. J., & Niblett, A. (2017). The death of rules and standards. *Indiana Law Journal, 92*(4), 1401–1447.
Castel, J. G., & Castel, M. E. (2016). The road to artificial superintelligence - has international law a role to play? *Canadian Journal of Law & Technology, 14*. https://ojs.library.dal.ca/CJLT/article/download/7211/6256.
Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence, 1*(2), 74. https://doi.org/10.1038/s42256-019-0020-9.
Cave, S., & Ó hÉigeartaigh, S.án S. (2018). An AI race for strategic advantage: Rhetoric and risks. *AAAI/ACM conference on artificial intelligence, ethics and society*. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf.
Cave, S., Dihal, K., & Dillon, S. (Eds.). (2020). *AI narratives: A history of imaginative thinking about intelligent machines*. New York: Oxford University Press.
Chomsky, N. (1976). Problems and mysteries in the study of human language. In A. Kasher (Ed.), *Language in focus: Foundations, methods and systems: Essays in memory of Yehoshua Bar-Hillel* (pp. 281–357). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-1876-0_15. Boston Studies in the Philosophy of Science.
Christakis, N. A. (2019). *How AI will rewire us*. March 4, 2019. The Atlantic https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/.
Cihon, P., Maas, M. M., & Kemp, L. (2020). Should artificial intelligence governance be centralised?: Design lessons from history. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 228–234). https://doi.org/10.1145/3375627.3375857.
Cirkovic, M. M., Bostrom, N., & Sandberg, A. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis, 30*(10), 1495–1506.
Citizens United v Federal Election Commission. (2010). *558 U.S. U.S. Supreme court*.
Clark, J. (2018). *Import AI #83: Cloning voices with a few audio samples, why malicious actors might mess with AI, and the industry academia compute gap*. February 26, 2018. Import AI (blog) https://jack-clark.net/2018/02/26/import-ai-83-cloning-voices-with-a-few-audio-samples-why-malicious-actors-might-mess-with-ai-and-the-industryacademia-compute-gap/.
Collingridge, D. (1980). *The social control of technology*. Frances Pinter.
Corea, F. (2020). *AI knowledge map: How to classify AI technologies*. Medium. June 28, 2020 https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020.
Crootof, R. (2018). Autonomous weapon systems and the limits of analogy. *Harvard National Security Journal, 9*, 51–83. https://doi.org/10.2139/ssrn.2820727.
Crootof, R. (2019a). Jurisprudential space junk: Treaties and new technologies. In C. Giorgetti, & N. Klein (Eds.), *Resolving conflicts in the law* (pp. 106–129). https://brill.com/view/book/edcoll/9789004316539/BP000015.xml.
Crootof, R. (2019b). Regulating new weapons technology. In E. T. Jensen, & R. T. P. Alcala (Eds.), *The impact of emerging technologies on the law of armed conflict* (pp. 1–25). Oxford University Press.
Crootof, R., & Ard, B. J. (2021). Structuring techlaw. *Harvard Journal of Law & Technology, 34*. https://papers.ssrn.com/abstract=3664124.
Cunneen, M., Mullins, M., Murphy, F., Shannon, D., Furxhi, I., & Ryan, C. (2020). Autonomous vehicles and avoiding the trolley (Dilemma): Vehicle perception, classification, and the challenges of framing decision ethics. *Cybernetics and Systems, 51*(1), 59–80. https://doi.org/10.1080/01969722.2019.1660541.
Currie, A. (2019). Existential risk, creativity & well-adapted science. In *Studies in the history & philosophy of science, 76* pp. 39–48). https://www.sciencedirect.com/science/article/abs/pii/S0039368117303278?via%3Dihub.
Dafoe, A. (2018). *AI governance: A research agenda*. Oxford: Center for the Governance of AI Future of Humanity Institute. https://www.fhi.ox.ac.uk/govaiagenda/.
Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology, 29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1.
Danaher, J. (2017). Building a postwork utopia: Technological unemployment, life extension and the future of human flourishing. In K. Lagrandeur, & J. Hughes (Eds.), *Surviving the machine age* (pp. 63–82). Palgrave-MacMillan.
Danaher, J. (2018). Artificial intelligence and the constitutions of the future. *Philosophical Disquisitions (blog)*. September 29, 2018 https://philosophicaldisquisitions.blogspot.com/2018/09/artificial-intelligence-and.html.
Danaher, J. (2020). *Axiological futurism: The systematic study of the future of human values*.
Danaher, J. (2019a). *Automation and utopia: Human flourishing in a world without work*. Harvard University Press.
Danaher, J. (2019b). Escaping skinner's box: Ai and the new era of techno-superstition. *Philosophical Disquisitions (blog)*. October 15, 2019 https://philosophicaldisquisitions.blogspot.com/2019/10/escaping-skinners-box-ai-and-new-era-of.html.
Danzig, R. (2017). An irresistible force meets a moveable object: The technology tsunami and the liberal world order. *Lawfare Research Paper Series, 5*(1). https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf.
Danzig, R. (2018). *Technology roulette: Managing loss of control as many militaries pursue technological superiority*. Center for a New American Security. https://www.cnas.org/publications/reports/technology-roulette.
Darling, K. (April 2012). Extending legal rights to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. Extending legal rights to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects, We Robot Conference, University of Miami, .. In *We Robot Conference* (pp. 1–18). https://doi.org/10.2139/ssrn.2044797
Davis, M., & Stark, A. (2001). Conflicts in rulemaking: Hard cases and bad law. *Conflict of interest in the professions*. Oxford University Press.
Daviter, F. (2017). Coping, taming or solving: Alternative approaches to the governance of wicked problems. *Policy Studies, 38*(6), 571–588. https://doi.org/10.1080/01442872.2017.1384543.
De Freitas, J., Anthony, S. E., & Alvarez, G. (2019). *Doubting driverless dilemmas*. Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/a36e5.
De Spiegeleire, S., Maas, M. M., & Sweijs, T. (2017). *Artificial intelligence and the future of defense: Strategic implications for small- and medium-sized force providers*. The Hague, The Netherlands: The Hague Centre for Strategic Studies. http://hcss.nl/report/artificial-intelligence-and-future-defense.
Deeks, A. (2018). Predicting enemies. *Virginia Law Review, 104*, 1529–1593.
Deeks, A. (2020). High-tech international law. *The George Washington Law Review, 88*, 575–653.
Derman, E. (2018). Simulated annealing. In J. Brockman (Ed.), *This idea is brilliant: Lost, overlooked, and underappreciated scientific concepts everyone should know*. Harper Perennial. https://www.edge.org/response-detail/27033.
Deudney, D. (2018). Turbo change: Accelerating technological disruption, planetary geopolitics, and architectonic metaphors. *International Studies Review, 20*(2), 223–231. https://doi.org/10.1093/isr/viy033.
Di Muzio, T. (2015). *Carbon capitalism: Energy, social reproduction and world order*. London; New York: Rowman & Littlefield Publishers.
Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Allen Lane.
Douglas, M. (1966). *Purity and danger: An analysis of concepts of pollution and taboo* (1st edition). London; New York: Routledge.
Easterbrook, F. H. (1996). *Cyberspace and the law of the horse, 207* p. 11). The University of Chicago Legal Forum.
Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3). https://doi.org/10.1111/1758-5899.12665.

Elish, M. C. (2016). *Moral crumple zones: Cautionary tales in human-robot interaction (We robot 2016)*. SSRN Scholarly Paper ID 2757236. Rochester, NY: Social Science Research Network https://papers.ssrn.com/abstract=2757236.
Els, A. S. (2017). Artificial intelligence as a digital privacy protector. *Harvard Journal of Law & Technology, 31*(1), 19.
Engler, A. (2019). *Fighting deepfakes when detection fails*. November 14, 2019. Brookings (blog) https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/.
Everitt, T., Lea, G., & Hutter, M. (2018). *AGI safety literature review*. ArXiv:1805.01109 [Cs], May http://arxiv.org/abs/1805.01109.
Felin, T., & Foss, N. J. (2009). Social reality, the boundaries of self-fulfilling prophecy, and economics. *Organization Science, 20*(3), 654–668. https://doi.org/10.1287/orsc.1090.0431.
Fish, S. (1989). Being interdisciplinary is so very hard to do. *Profession*, 15–22.
Garcia, D. (2016). Future arms, technologies, and international law: Preventive security governance. *European Journal of International Security, 1*(1), 94–111. https://doi.org/10.1017/eis.2015.7.
Garcia, D. (2018). Lethal artificial intelligence and change: The future of international peace and security. *International Studies Review, 20*(2), 334–341. https://doi.org/10.1093/isr/viy029.
Gasser, U. (2016). Recoding privacy law: Reflections on the future relationship among law, technology, and privacy. *Harvard Law Review Forum, Law, Privacy & Technology Commentary Series, 130*(December), 10.
Geist, E., & Lohn, A. J. (2018). *How might artificial intelligence affect the risk of nuclear war?* RAND. https://www.rand.org/pubs/perspectives/PE296.html.
Gläveanu, V. P. (2012). What can be done with an egg? Creativity, material objects, and the theory of affordances. *The Journal of Creative Behavior, 46*(3), 192–208. https://doi.org/10.1002/jocb.13.
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. *ArXiv:1412.6572 [Cs, Stat]*, December http://arxiv.org/abs/1412.6572.
Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences, 114*(30), 7892–7899. https://doi.org/10.1073/pnas.1700811114.
Gruetzemacher, R. (2018). *Rethinking AI strategy and policy as entangled super wicked problems*. In, 6. New Orleans http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_70.pdf.
Guihot, M., Matthew, A. F., & Suzor, N. (2017). Nudging robots: Innovative solutions to regulate artificial intelligence. *Vanderbilt Journal of Entertainment & Technology Law*, (July). https://papers.ssrn.com/abstract=3017004.
Hagemann, R., Huddleston, J., & Thierer, A. D. (2018). Soft law for hard problems: The governance of emerging technologies in an uncertain future. *Colorado Technology Law Journal, 17*(1), 94.
Harari, Y. N. (2018). *Why technology favors tyranny*. October 2018. The Atlantic https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/.
Hayward, K. J., & Maas, M. M. (2020). Artificial intelligence and crime: A primer for criminologists. *Crime Media Culture*, (June)https://doi.org/10.1177/1741659020917434, 1741659020917434.
Head, B. W. (2008). Wicked problems in public policy. *Public Policy, 3*(2), 101–118.
Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2019). *Natural adversarial examples*. *ArXiv:1907.07174 [Cs, Stat]*, July http://arxiv.org/abs/1907.07174.
Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice, 21*(3), 669–684. https://doi.org/10.1007/s10677-018-9896-4.
Hoffmann-Riem, W. (2020). Artificial intelligence as a challenge for law and regulation. In T. Wischmeyer, & T. Rademacher (Eds.), *Regulating artificial intelligence* (pp. 1–29). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-32361-5_1.
Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*. May 15, 2018 https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/.
Houghton, D. P. (2009). The role of self-fulfilling and self-negating prophecies in international relations. *International Studies Review, 11*(3), 552–584. https://doi.org/10.1111/j.1468-2486.2009.00873.x.
Huysmans, J. (2000). The European union and the securitization of migration. *JCMS Journal of Common Market Studies, 38*(5), 751–777. https://doi.org/10.1111/1468-5965.00263.
Hwang, T. (2018). *Computational power and the social impact of artificial intelligence*. March https://arxiv.org/abs/1803.08971.
Jain, N. (2016). Autonomous weapons systems: New frameworks for individual responsibility. In N. Bhuta, S. Beck, R. Geiß, H.-Y. Liu, & K.ß Claus (Eds.), *Autonomous weapons systems - law, ethics policy* (pp. 303–324). Cambridge University Press.
Jaques, A. E. (2019). *Why the moral machine is a monster* (p. 10).
Jasanoff, S., & Kim, S.-H. (Eds.). (2015). *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power* (1st edition). Chicago; London: University of Chicago Press.
Jensen, E. T. (2014). The future of the law of armed conflict: Ostriches, butterflies, and nanobots. *Michigan Journal of International Law, 35*(2), 253–317.
Kahn, P. W. (2002). The paradox of riskless warfare. *Philosophy & Public Policy Quarterly, 22*(3), 2–8.
Kaminski, M. E. (2017). Authorship, disrupted: AI authors in copyright and first amendment law symposium - future-proofing law: From RDNA to robots (Part 2). *U.C. Davis Law Review, 51*, 589–616.
Khurana, R. (2019). *The threat of automation is a self-fulfilling prophecy*. July 5, 2019. Palladium Magazine (blog) https://palladiummag.com/2019/07/05/the-threat-of-automation-is-a-self-fulfilling-prophecy/.
King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2019). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics*, (February)https://doi.org/10.1007/s11948-018-00081-0.
Kreienkamp, J., & Pegram, T. (2020). Governing complexity: Design principles for the governance of complex global catastrophic risks. *International Studies Review, viaa074*(October). https://doi.org/10.1093/isr/viaa074.
Kuhlemann, K. (2018). Complexity, creeping normalcy and conceit: Sexy and unsexy catastrophic risks. *Foresight*, (November)https://doi.org/10.1108/FS-05-2018-0047.
Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Legg, S., & Hutter, M. (2007). *A collection of definitions of intelligence*. *ArXiv:0706.3639 [Cs]*, June http://arxiv.org/abs/0706.3639.
Lessig, L. (1998). The new Chicago school. *The Journal of Legal Studies, 27*(S2), 661–691. https://doi.org/10.1086/468039.
Lessig, L. (1999). The law of the horse: What cyberlaw might teach. *Harvard Law Review, 113*(2), 501. https://doi.org/10.2307/1342331.
Leung, J. (2019). *Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies*. Oxford: University of Oxford. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665.
Liu, H.-Y. (2012). Categorization and legality of autonomous and remote weapons systems. *International Review of the Red Cross, 94*(886), 627–652. https://doi.org/10.1017/S181638311300012X.
Liu, H.-Y. (2018). The power structure of artificial intelligence. *Law, Innovation and Technology, 10*(2), 197–229. https://doi.org/10.1080/17579961.2018.1527480.
Liu, H.-Y. (2019a). From the autonomy framework towards networks and systems approaches for 'autonomous' weapons systems. *Journal of International Humanitarian Legal Studies, 10*(1), 89–110.
Liu, H.-Y. (2019b). The digital disruption of human rights foundations. *Human rights, digital society and the law: A research companion*. London: Routledge.
Liu, H.-Y. (2019c). From the autonomy framework towards networks and systems approaches for 'autonomous' weapons systems. *Journal of International Humanitarian Legal Studies, 10*(1), 89–110. https://doi.org/10.1163/18781527-01001010.
Liu, H.-Y., Lauta, K. C., & Maas, M. M. (2018). Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures, 102*, 6–19. https://doi.org/10.1016/j.futures.2018.04.009.

Liu, H.-Y., Maas, M., Danaher, J., Scarcella, L., Lexer, M., & Van Rompaey, L. (2020). Artificial intelligence and legal disruption: A new model for analysis. *Law, Innovation and Technology, 0*(0), 1–54. https://doi.org/10.1080/17579961.2020.1815402.
Liu, H.-Y., & Zawieska, K. (2020). From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. *Ethics and Information Technology, 22*, 321–333. https://doi.org/10.1007/s10676-017-9443-3.
Maas, M. M. (2012). *The forging of our futures: The temporal construction of political logics, the performative self-fulfillment of prophetic visions, and the need for a post-positivist 'transformational idealism' in IR*. Utrecht, The Netherlands: University College Utrecht. https://www.academia.edu/8007823/The_Forging_of_Our_Futures_The_Temporal_Construction_of_Political_Logics_the_Performative_Self-Fulfillment_of_Prophetic_Visions_and_the_Need_for_a_Post-Positivist_Transformational_Idealism_in_IR.
Maas, M. 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* (pp. 223–228). https://doi.org/10.1145/3278721.3278766.
Maas, M. M. (2020). *Artificial intelligence governance under change: Foundations, facets, frameworks*. Copenhagen, Denmark: University of Copenhagen.
Maas, M. M. (2019a). 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.
Maas, M. M. (2019b). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy, 40*(3), 285–311. https://doi.org/10.1080/13523260.2019.1576464.
Maas, M. M. (2019c). 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. https://doi.org/10.1163/18781527-01001006.
Mandel, G. N. (2017). Legal evolution in response to technological change. *The Oxford handbook of law, regulation and technology*. https://doi.org/10.1093/oxfordhb/9780199680832.013.45. July.
Manheim, D. (2018). *Systemic fragility as a vulnerable world*. https://philpapers.org/archive/MANSFA-3.pdf.
Marchant, G. E. (2011). The growing gap between Emerging technologies and the law. In G. E. Marchant, B. R. Allenby, & J. R. Herkert (Eds.), *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem* (pp. 19–33). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-1356-7_2. The International Library of Ethics, Law and Technology.
Mollick, E. (2006). Establishing Moore's law. *IEEE Annals of the History of Computing, 28*(3), 62–75. https://doi.org/10.1109/MAHC.2006.45.
Morin, J. F., Dobson, H., Peacock, C., Prys-Hansen, M., Anne, A., Bélanger, L., , … Dietsch, P., et al. (2019). How Informality can address emerging issues: Making the most of the G7. *Global Policy, 10*(2), 267–273. https://doi.org/10.1111/1758-5899.12668.
Morton, T. (2013). Hyperobjects: Philosophy and ecology after the end of the world. *Hyperobjects: Philosophy and ecology after the end of the world* (pp. 1–230).
Moynihan, T. (2020). Existential risk and human extinction: An intellectual history. *Futures, 116*(February), 102495. https://doi.org/10.1016/j.futures.2019.102495.
NASA. (2020). *"Goldilocks zones and stars." NASA science mission directorate*. January 31, 2020 https://science.nasa.gov/goldilocks-zones-and-stars.
Nindler, R. (2019). The United Nation's capability to manage existential risks with a focus on artificial intelligence. *International Community Law Review, 21*(1), 5–34. https://doi.org/10.1163/18719732-12341388.
Noordegraaf, M., Douglas, S., Geuijen, K., & Van Der Steen, M. (2019). Weaknesses of wickedness: A critical perspective on wickedness theory. *Policy and Society, 38*(2), 278–297. https://doi.org/10.1080/14494035.2019.1617970.
Norman, D. A. (2013). *The design of everyday things* (revised and expanded edition). New York, New York: Basic Books.
OpenAI, M. A., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., et al. (2018). *Learning dexterous in-hand manipulation. ArXiv:1808.00177 [Cs, Stat]*, August http://arxiv.org/abs/1808.00177.
Parker, J., & Danks, D. (2019). *How technological advances can reveal rights*. In, 7 http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_129.pdf.
Parson, E., Re, R., Solow-Niederman, A., & Zeide, A. (2019). *Artificial intelligence in strategic context: An introduction*. February 8, 2019. AI Pulse (blog) https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/.
Pauwels, E. (2019). *The new geopolitics of converging risks: The UN and prevention in the era of AI*. United Nations University - Centre for Policy Research. https://i.unu.edu/media/cpr.unu.edu/attachment/3472/PauwelsAIGeopolitics.pdf.
Pauwelyn, J., Wessel, R. A., & Wouters, J. (2014). When structures become shackles: Stagnation and dynamics in international lawmaking. *European Journal of International Law, 25*(3), 733–763. https://doi.org/10.1093/ejil/chu051.
Peters, B. G., & Tarpey, M. (2019). Are wicked problems really so wicked? Perceptions of policy problems. *Policy and Society, 38*(2), 218–236. https://doi.org/10.1080/14494035.2019.1626595.
Petit, N. (2017). *Law and regulation of artificial intelligence and robots - conceptual framework and normative implications*. SSRN Scholarly Paper ID 2931339. Rochester, NY: Social Science Research Network https://papers.ssrn.com/abstract=2931339.
Picker, C. B. (2001). A view from 40,000 feet: International law and the invisible hand of technology. *Cardozo Law Review, 23*, 151–219.
Ram, N. (2019). *One shot learning in AI innovation*. January 25, 2019. *AI Pulse* (blog) https://aipulse.org/one-shot-learning-in-ai-innovation/.
Rapaport, W. J. (2020). What is artificial intelligence? *Journal of Artificial General Intelligence, 11*(2), 52–56. https://doi.org/10.2478/jagi-2020-0003. Special Issue "On Defining Artificial Intelligence"—Commentaries and Author's Response.
Rayfuse, R. (2017). Public international law and the regulation of emerging technologies. *The Oxford handbook of law, regulation and technology*. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-22.
Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155–169.
Rosa, H. (2013). In J. Trejo-Mathys (Ed.), *Social acceleration: A new theory of modernity*. Columbia University Press.
Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking. https://www.amazon.com/Human-Compatible-Artificial-Intelligence-Problem-ebook/dp/B07N5J5FTS.
Sadowski, J., & Bendor, R. (2019). Selling smartness: Corporate narratives and the Smart City as a sociotechnical imaginary. *Science, Technology & Human Values, 44*(3), 540–563. https://doi.org/10.1177/0162243918806061.
Saxe, J. G. (1872). The blind men and the elephant. *The poems of John Godfrey Saxe*. Boston: J. Osgood.
Scarry, E. (2016). *Thermonuclear monarchy: Choosing between democracy and doom* (1st edition). W. W. Norton & Company.
Scharre, P. (2016a). *"Autonomous weapons and operational risk." Ethical autonomy project. 20YY future of warfare initiative*. Center for a New American Security. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf.
Scharre, P. (2016b). Flash war - autonomous weapons and strategic stability. In *Presented at the understanding different types of risk*. http://www.unidir.ch/files/conferences/pdfs/-en-1-1113.pdf.
Scharre, P., & Horowitz, M. C. (2018). *Artificial intelligence: What every policymaker needs to know*. Center for a New American Security. https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know.
Schellekens, M. (2015). Self-driving cars and the chilling effect of liability law. *Computer Law & Security Report, 31*(4), 506–517. https://doi.org/10.1016/j.clsr.2015.05.012.
Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology, 2*(Spring). http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf.
Schmitt, M.el (2013). Autonomous weapons systems and international humanitarian law: A reply to the critics. *Harvard National Security Journal Features*.
Sears, N. A. (2020). Existential security: Towards a security framework for the survival of humanity. *Global Policy, 11*(2), 255–266. https://doi.org/10.1111/1758-5899.12800.
Sharikov, P. (2018). Artificial intelligence, cyberattack, and nuclear weapons—A dangerous combination. *The Bulletin of the Atomic Scientists, 74*(6), 368–373. https://doi.org/10.1080/00963402.2018.1533185.
Sheppard, B. (2018). Warming up to inscrutability: How technology could challenge our concept of law. *The University of Toronto Law Journal, 68*(Supplement 1), 36–62. https://doi.org/10.3138/utlj.2017-0053.

Shevlane, T., & Dafoe, A. (2020). 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* (pp. 173–179). https://doi.org/10.1145/3375627.3375815.
Stritzel, H. (2014). Securitization theory and the Copenhagen school. In H. Stritzel (Ed.), *Security in translation: Securitization theory and the localization of threat* (pp. 11–37). New Security Challenges Series. London: Palgrave Macmillan. https://doi.org/10.1057/9781137307576_2. UK.
Susser, D. (2019). Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. *Proceedings of the AAAI / ACM conference on artificial intelligence, ethics and society 2019*, 7.
Tipson, F. S. (1997). *Culture clash-ification: A verse to Huntington's curse*, 1997. Foreign Affairs https://www.foreignaffairs.com/articles/1997-03-01/culture-clash-ification-verse-huntingtons-curse.
Torres, P. (2018). Facing disaster: The great challenges framework. *Foresight, 21*(1), 4–34. https://doi.org/10.1108/FS-04-2018-0040.
Trajtenberg, M. (2018). AI as the next GPT: A political-economy perspective. *Working paper 24245. National bureau of economic research*. https://doi.org/10.3386/w24245.
Tranter, K. (2011). The law and technology enterprise: Uncovering the template to legal scholarship on technology. *Law, Innovation and Technology, 3*(1), 31–83. https://doi.org/10.5235/175799611796399830.
Tucker, A. D., Anderljung, M., & Dafoe, A. (2020). Social and governance implications of improved data efficiency. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 378–384). https://doi.org/10.1145/3375627.3375863.
Turnbull, N., & Hoppe, R. (2019). Problematizing 'wickedness': A critique of the wicked problems concept, from philosophy to practice. *Policy and Society, 38*(2), 315–337. https://doi.org/10.1080/14494035.2018.1488796.
Turner, J. (2018). *Robot rules: Regulating artificial intelligence*. New York, NY: Springer. Berlin Heidelberg.
van Aaken, A. (2016). Is international law conducive to preventing looming disasters? *Global Policy, 7*(S1), 81–96. https://doi.org/10.1111/1758-5899.12303.
van Assche, K., Beunen, R., & Duineveld, M. (2014). *Evolutionary governance theory. SpringerBriefs in economics*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00984-1.
van Assche, K., Verschraegen, G., Gruezmacher, M., & Boezeman, D. (2020). Strategy for collectives and common goods: Coordinating strategy, long term perspectives and policy domains in governance. *Futures*.
van der Loeff, A. S., Bassi, I., Kapila, S., & Gamper, J. (2019). *AI ethics for systemic issues: A structural approach*. Canada: Vancouver. http://arxiv.org/abs/1911.03216.
Vöneky, S. (2018). Human rights and legitimate governance of existential and global catastrophic risks. In S. Vöneky, & G. Neuman (Eds.), *Human rights, democracy, and legitimacy in a world of disorder* (pp. 139–162). Cambridge University Press. https://papers.ssrn.com/abstract=3363552.
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436.
Wiener, J. B. (2016). The tragedy of the uncommons: On the politics of apocalypse. *Global Policy, 7*(S1), 67–80. https://doi.org/10.1111/1758-5899.12319.
Wilson, E. O. (2009). *An intellectual entente*. September 10, 2009. Harvard Magazine https://harvardmagazine.com/breaking-news/james-watson-edward-o-wilson-intellectual-entente.
Wilson, G. (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. *Va. Envtl. LJ, 31*, 307.
Wolkenstein, A. (2018). What has the trolley dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology, 20*(3), 163–173. https://doi.org/10.1007/s10676-018-9456-6.
Wright, N. (2018). *How artificial intelligence will reshape the global order: The coming competition between digital authoritarianism and liberal democracy*. July 10, 2018. Foreign Affairs https://www.foreignaffairs.com/articles/world/2018-07-10/how-artificial-intelligence-will-reshape-global-order.
Wright, R. (2005). *A short history of progress*. New York: CARROLL & GRAF.
Yeung, K. (2017). 'Hypernudge': Big data as a mode of regulation by design. *Information, Communication and Society, 20*(1), 118–136. https://doi.org/10.1080/1369118X.2016.1186713.
Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance, 12*(4), 505–523. https://doi.org/10.1111/rego.12158.
Yudkowsky, E. (2017). *Inadequate equilibria: Where and how civilizations get stuck*. Machine Intelligence Research Institute.
Yudkowsky, E. (2008a). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom, & M. M. Cirkovic (Eds.), *Global catastrophic risks* (pp. 308–345). New York: Oxford University Press.
Yudkowsky, E. (2008b). Cognitive biases potentially affecting judgment of global risks. *Global Catastrophic Risks, 1*(86), 13.
Zürn, M. (2012). Global governance as multi-level governance. *The Oxford handbook of governance*. https://doi.org/10.1093/oxfordhb/9780199560530.013.0051. March.
Zürn, M. (2018). Contested global governance. *Global Policy, 9*(1), 138–145. https://doi.org/10.1111/1758-5899.12521.
Zwetsloot, R., & Dafoe, A. (2019). *Thinking about risks from AI: Accidents, misuse and structure*. February 11, 2019. Lawfare https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure.