Journal of European Public Policy

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjpp20

Governing AI – attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence

Tim Büthe, Christian Djeffal, Christoph Lütge, Sabine Maasen & Nora von Ingersleben-Seip

To cite this article: Tim Büthe, Christian Djeffal, Christoph Lütge, Sabine Maasen & Nora von Ingersleben-Seip (2022) Governing AI – attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence, Journal of European Public Policy, 29:11, 1721-1752, DOI: 10.1080/13501763.2022.2126515

To link to this article: https://doi.org/10.1080/13501763.2022.2126515

JOURNAL OF EUROPEAN PUBLIC POLICY
2022, VOL. 29, NO. 11, 1721–1752
https://doi.org/10.1080/13501763.2022.2126515

Routledge
Taylor & Francis Group

INTRODUCTION

Governing AI – attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence

Tim Büthe [a,b,c,d], Christian Djeffal [b,e], Christoph Lütge [b,f], Sabine Maasen[g] and Nora von Ingersleben-Seip [a,b]

[a]Hochschule für Politik München/Munich School of Politics and Public Policy at the Technical University of Munich, Munich, Germany; [b]TUM School of Social Sciences and Technology, Department of Governance, Technical University of Munich, Munich, Germany; [c]TUM School of Management, Technical University of Munich, Munich, Germany; [d]Sanford School of Public Policy, Duke University, Durham, NC, USA; [e]TUM School of Social Sciences and Technology, Department of Science and Technology Studies, Technical University of Munich, Munich, Germany; [f]Institute for Ethics in Artificial Intelligence, Technical University of Munich, Munich, Germany; [g]Faculty of Business, Economics, and Social Sciences, University of Hamburg, Hamburg, Germany

ABSTRACT

Artificial Intelligence raises new, distinct governance challenges, as well as familiar governance challenges in novel ways. The governance of AI, moreover, is not an issue of distant futures, it is well underway – and it has characteristics akin to 'herding cats' with a mind of their own. This essay introduces the contributions to the special issue, situating them in broader political and social science literatures. It then provides a sketch of an interdisciplinary research agenda. It highlights the limits of 'explainable AI', makes the case for considering AI ethics and AI governance simultaneously, identifies as an underappreciated risk 'system effects' that arise from the introduction of AI applications, and calls for policymakers to consider both the opportunities and the risks of AI. Focusing on the (ab)uses of AI, rather than the complex, rapidly changing and hard-to-predict technology as such, might provide a superior approach to governing AI.

KEYWORDS Artificial intelligence; governance; regulation; ethics; disruptive technology; algorithms; responsibility; technology assessment

Introduction

This special issue brings together a diverse set of papers to examine distinctive challenges and opportunities in the governance of AI from a variety of perspectives, reflecting our commitment to a multi-disciplinary, multi-level and multi-dimensional approach to the governance of complex socio-

CONTACT Tim Büthe buthe@hfp.tum.de

technical processes, which requires addressing both particularities and general issues. Accordingly, the special issue contains, in addition to this introductory essay, five articles that raise matters of general concern by delving deep into specific topics. Notwithstanding their diverse perspectives and substantive foci, the papers focus on political and public policy aspects of the topic, so as to advance the social-scientific understanding of AI, as well as enable public policies guided by the spirit of 'handle with care, not fear' (Renda, 2019, pp. 21ff).

Artificial Intelligence (AI) is the focus of much controversial discussion, including among scholars and policymakers in Europe and beyond, often without a clear understanding of what makes it distinctive. It is variously described as '*the* technology *of the future*,' an already widely used 'general purpose technology,' a particular 'key technology,' or a 'set' of quite diverse technologies. Such different conceptions and definitions do not just prompt academic debates. How we define AI matters, *inter alia* for the rights of stakeholders: when laws and regulations assign specific rights and obligations to AI users (and others affected by the use of AI), those rights hinge upon what is considered an AI application or system.[1]

Ambiguity about the definitional boundaries of AI also contributes to sharply divergent reactions: Some categorically reject AI as exceedingly disruptive (Babcock *et al.*, 2019; Bostrom, 2014; Armstrong & Pamlin, 2015; Armstrong *et al.*, 2016) or fear AI as threatening important enlightenment traditions such as transparency and accountability in the exercise of power (Bryson & Theodorou, 2019; Floridi & Cowls, 2022). Others admire it, often uncritically (e.g., Shadbolt *et al.*, 2016) – and indeed AI is enabling phenomenal advances in, e.g., predictions based on pattern recognition, which make it possible to use the wealth of data that characterize the digital age to achieve otherwise unobtainable gains in economic efficiency (TCS, 2017), environmental sustainability (ITU, 2021; Khan, 2022), or medical treatments (Hamet & Tremblay, 2017; He *et al.*, 2019). Yet others consider the adoption of AI simply inevitable, similar to previous technologies that gave a competitive advantage to early adopters in the global economy and in world politics (Milner & Solstad, 2021), though they might still warn of the risks of dangerous AI-fuelled arms races (NSCAI, 2021; Walker, 2021). In sum, how we define AI also matters for normative assessment. So how might we define AI?

AI: attempt at a definition

While AI has attracted much attention, there are still substantive struggles concerning even the definition of AI. Disagreements about the definition start in computer science (Russell & Norvig, 2020, pp. 19–23), where the term was initially coined in a grant application in 1955. McCarthy *et al.* (1955) in fact did not really define the term but merely stressed that the

'study [of artificial intelligence] is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.'[2]

Having left the definition of AI largely open – and approaching it as a (set of) general purpose technologies – may have been a strength in that it allowed a range of different technologies to be subsumed under the AI concept (see, e.g., Crafts, 2021; Djeffal, 2019, pp. 256–260; Gasser *et al*., 2018, p. 3). Conversations about 'artificial intelligence' thus show considerable continuity even as data structures changed, computing capacities exploded, and algorithmic approaches evolved, e.g., from linear to Bayesian to machine learning, which is at the center of contemporary discussions of AI. Leaving the definition of AI open also facilitates continuing to subsume future technological breakthroughs under the umbrella of AI.

At the same time, vagueness due to the lack of a definition comes at a cost. Many policy instruments employ only partial or preliminary definitions – or refrain from providing a definition at all – which creates uncertainty and invites *ex post* conflicts, including costly litigation, about the applicability of any particular measure. It also creates challenges in the legislative and regulatory process as, for instance, different committees of the European Parliament currently hold different – and in some cases widely different – views on the definition of AI (2022a, 2022b). The lack of a clear definition, moreover, invites 'concept stretching' in lieu of fruitful concept development (Sartori, 1970; Collier & Mahon, 1993; Munck, 2004). Such stretching impedes analytical precision and knowledge accumulation in scholarship and encourages loose talk in lieu of productive deliberation in policy discussions.

We cannot resolve, or authoritatively settle, the definitional controversies here. But to encourage conceptual clarity, we set out an explicit definition early on:

> For purposes of this special issue, we understand AI as systems that combine means of gathering inputs from the external environment with technologies for processing these inputs and relating them algorithmically to previously stored data, allowing such systems to carry out a wide variety of tasks with some degree of autonomy, i.e., without simultaneous, ongoing human intervention. Their capacity for learning allows AI systems to solve problems and thus support, emulate or even improve upon human decision-making – though (at least at this point of technological development) not holistically but only with regard to well-specified, albeit possibly quite complex tasks.[3]

Contributors were free to diverge or take issue with this definition, but were asked to do so explicitly.

From concept to governance

The above definition is meant to foster multi- and interdisciplinary research and debate, based on recognizing both the risks and opportunities of AI.

Many of the fundamental ideas underlying AI have in principle been known since the 1950s. AI, however, tends to be 'data-hungry' and computation-intensive, which for several decades limited the potential for implementing those ideas. Only quite recently have exponential increases in sensor capabilities, data storage capacity, and computing power started to make a wide range of new – and to some extent previously unimaginable – applications of AI possible. This development has been accompanied by the proliferation of discourses about AI and a shift of technological development to specific issue areas like mobility, health, industry 4.0, smart cities, etc. These rapid developments make AI and its governance a very timely issue.

Many new AI applications hold tremendous promise for improving quality of life, reducing severe risks, and enabling leaps in factor productivity and hence economic growth – while simultaneously reducing environmental harm. Massively improved AI-driven speech recognition and ever less invasive, AI-connected sensory devices, for instance, empower mobility-impaired individuals and can reduce loneliness and isolation (DiNuovo, 2018; Kiron & Unruh, 2019). AI-powered decision-support systems reduce the risks of potentially fatal medical treatment errors (Guo & Li, 2018; Topol, 2019). AI-powered blind spot warning systems for trucks save the lives of pedestrians and bicyclists (Stephan, 2018). In addition and often simultaneously, AI promises reductions in environmental harm. A wide range of gains in factor productivity are made possible by AI-driven increases in the efficient usage of resources (Furman & Seamans, 2018). And new AI-driven solar energy bundling, a very recent breakthrough development, enables even the production of steel and cement, which requires exceptionally large amounts of energy, without emitting any greenhouse gases (Nield, 2019). Policymakers – and citizens – generally do not want to impede these benefits of AI through the way AI is governed, and they indeed might even want to encourage the beneficial development of AI (Busse & Baeva, 2022).

At the same time, AI comes with considerable risks, which are often perceived as threats – in many instances for good reasons. These threats include potentially radical, rapid changes in the workplace and in labor markets, since AI allows not just further advances in traditional automation (replacing manual labor) but enables 'intelligent tools' (Zysman & Kenney, 2017) to perform a broad range of the tasks traditionally done by highly trained people, such as examining x-rays or extracting pertinent information (and increasingly better than people). Other key concerns are unanticipated malfunctions, as well as insufficient IT security, resulting in the hacking of AI systems and other cyber security risks. AI also creates or exacerbates risks for politics and public policy: Fabricated information, including falsified or wholly invented documents, images, and data ('deep fakes'), created with powerful AI applications that make it ever harder to detect them as fakes, has already been used to manipulate the political behavior of various target groups in the

run-up to elections (and referenda such as Brexit), thus threatening the viability of anything like a deliberative process in democratic societies. And the development of sophisticated delivery systems, capable of autonomously transporting and detonating nuclear, biological, or chemical weapons, greatly increases the threat potential of those who can obtain such weapons and might accelerate the military escalation of a conflict in a foreign policy crisis.

AI thus is neither inherently good nor inherently malign or evil. Rather, AI raises new, distinct governance challenges, as well as challenges that are familiar from other cases of technology governance (e.g., Christou & Simpson, 2011) and 'disruptive technological change' (Hasselbalch, 2018), albeit in arguably novel ways. The key question for public policy, therefore, is not whether to categorically accept or reject AI but rather how to *govern* it.

Governments, individual companies and private sector bodies, civil society organizations, multi-stakeholder initiatives, and various academics have put forth a large number of proposals for governing AI. They range from norms and standards, guidelines, and policies, to laws and regulations. Many of these initiatives and public policies have been developed at the local or national level. The OECD has identified 1459 AI-related policy measures, from funding to legally binding regulations, among its member states alone (OECD.AI, 2021).[4] Increasingly, proposals for AI governance have also been developed at the transnational and international levels. The European Union, in particular, has become a major arena for AI governance. The European Commission joined 15 states to develop a strategy on AI (2018), which prompted the development of numerous strategies in member states. The President of the European Commission, Ursula von der Leyen, has designated AI as one of the five top priorities of her tenure (2019). A Proposal for an Artificial Intelligence Act is underway (European Commission, 2021), as well as a discussion of new rules for civil liability (European Parliament, 2020), etc. The European Parliament has stressed the importance of AI on numerous occasions, recently also in its resolution on AI (2022c). A European multi-stakeholder dialogue has brought different voices together (Pagallo *et al.*, 2019).[5]

What needs to be governed?

Productively and effectively governing AI requires understanding what makes AI distinctive and what therefore the most important issues are to warrant governance. Three elements or characteristics of AI stand out, discussed in greater detail in the paper by Nitzberg and Zysman (2022). First, AI relies upon very large quantities of data, making access to data a very important issue for AI governance, including the rights of individuals and

institutional or commercial 'owners' of massive databases to restrict others' access to data, as well as the limits of such rights.

AI generally delivers better outcomes the more data it can draw upon. Sometimes, it may consequently be crucial to ensure access to a wealth of data, including (if data security concerns can be addressed) personal (identified) data. Access to detailed medical histories and even genetic information, for instance, holds great promise for phenomenal progress in developing treatments for rare diseases and for personalized medicine. Yet, sometimes, we might not want AI to draw upon certain information, even if it is readily available. When AI, for example, is used to estimate credit default risks (and thus access to finance) or cost-neutral health insurance premiums, we might not want to allow it to use even basic demographic information such as sex, gender, age, or the neighborhood in which the person resides, because it would perpetuate inequities or introduce new forms of bias (see, e.g., Obermeyer *et al*., 2019).

Second, AI is distinct from traditional forms of data management and analysis in that it entails (semi-)autonomous learning, which can be the source of both risks and opportunities. How learning takes place therefore is a crucial, distinctive issue for AI governance. AI is generally very good at finding locally optimal solutions, even in very complex, multi-dimensional issue spaces, but it cannot on its own determine what to optimize. As Nitzberg and Zysman put it, decisions about 'when, where, and how' AI optimizes 'reflect' societal goals and norms (2022, p. 1763), and indeed, such decisions often are a function of trade-offs between multiple values. AI might of course be used to discern what those social norms are. And if we subscribe to a revealed preferences approach and a consequentialist ethic, AI might in fact be better at detecting human preferences and apparent priorities than most people are themselves. But notice that delegating the choice of what is to be optimized (or relying on AI decision support for making the choice) implies having consciously made an even more fundamental normative decision that we should not want machines to make for us. What our preferences are (or at least which considerations should go into a decision) is thus central to AI governance.

Third, common elements of AI use raise some issues that can be addressed at a general level, but most of the action in AI is in its myriad combinations with very different socio-technical systems such as exoskeletonous robotic devices, chatbots, or content moderation on social media and prediction engines. Such combinations can create new, distinctive functionalities in many different areas, such as healthcare, public security, games, and consumer goods and industrial production.

One key issue is therefore how general or specific AI governance can or should be. Ultimately both the general and the specific challenges raised by the use of AI will need to be addressed, and the new challenges, which AI raises for governance, can differ quite a bit across diverse AI applications.

In an ideal world, AI governance and the development of applied AI technology would therefore proceed together and iteratively. An iterative approach also would be desirable because the specific 'possibilities, applications, and risks' do not just follow deductively from characteristics of the technology, but 'only emerge in the early years' of the development of the technology (Nitzberg & Zysman, 2022, p. 1754). As Nitzberg and Zysman point out, however, given the high likelihood of path-dependent developments, such an iterative approach 'is not feasible as we have a narrow window for implementing technology governance' (2022, p. 1759).

The articles in this special issue

The issue starts with *Mark Nitzberg and John Zysman*'s article on 'The Diverse Challenges of Governing AI.' They begin by putting AI into its historical and technological context, which culminates in a discussion of the capabilities of AI and the limits that differentiate it from human intelligence – above all regarding 'fundamental aspects of human cognition and interpretation': the ability to contextualize information and human beings' ability to 'construct their own narratives and worldview' (p. 1758). Although it is therefore premature to worry about governing 'strong AI' or 'artificial general intelligence' (systems with cognitive abilities that fully match or exceed human intelligence), myriad aspects and consequences of AI use in highly specific sectors and domains might still warrant governance.

Beyond offering insights into what context-specific AI governance might entail (as discussed above), Nitzberg and Zysman emphasize that gaining the full benefits of AI use requires the deployment of the entire 'technology stack' or suite of 'intelligent tools.' This leads them to a discussion of the central role played by platforms, especially the dominant digital platform firms (DPFs), in that technology stack. DPFs are crucial, they argue, because DPFs' dominant position in the market (and often control over other elements of the technology stack) renders their practices *de facto* standards. In other words, DPFs are able to make the rules for AI without having to consult or bargain with other stakeholders. Nitzberg and Zysman, therefore, advocate regulating, above all, the platform firms in order to regain a modicum of democratic control and achieve meaningful AI governance.

Public policy can in a number of ways encourage, accelerate, nudge, constrain or even prevent technological change. To what extent it does any of these things presumably is – at least in part and at least in democracies – a function of the assessments of the specific technology by society or, more precisely, by the people who constitute the democratic polity.

One way to achieve such democratic governance of AI might be to involve the citizens in technology development and governance directly (e.g.,

Costanza-Chock, 2020). Research on participatory and deliberative democracy suggests that it is possible to meaningfully involve citizens without prior issue-area expertise in the governance of even quite complex issues (Bächtiger *et al*., 2019; Dryzek *et al*., 2019; Elstub & Escobar, 2019; Fishkin, 2018; Fung & Warren, 2011), and some are therefore optimistic that modest public financing might suffice to achieve and sustain public participation in technology governance (Büthe & Mattli, 2011, ch.9). More recent work, however, suggests that sustaining meaningful public participation in the governance of complex, fast-changing issues (which AI technologies certainly are) is much more challenging, even when some funding support is available (Alshadafan, 2020, though cf. Ada Lovelace Institute 2022, forthcoming), especially when the governance needs to be transnational or global (Grigorescu, 2015; Pauwelyn *et al*., 2022). The challenge also becomes apparent in studies such as the recent study for the German Center for Trustworthy AI by Busse and Baeva (2022), who find that most citizens feel that they insufficiently understand how AI systems work.

Another way to achieve democratic governance of technologies such as AI is to discern the preferences of citizens, then have democratic governments act on those preferences.[6] Accordingly, *Sönke Ehret* (2022) examines public preferences regarding the regulation of AI, going beyond the mostly country-specific existing literature by analyzing public preferences across a very diverse set of five countries: Chile, China, Germany, India, and the United Kingdom. As he points out, much (elite-driven) public debate – and much of the existing research – has focused on opposition to, or calls for prohibitions of, the use of AI as a function of the extent to which AI conforms to, or conflicts with, what are believed to be widely held societal norms regarding data privacy, explainability, etc. (see, e.g., Shin, 2021). Ehret builds on this prior work but focuses on how these possible drivers of preferences compare (and possibly interact) with AI-induced economic risks and opportunities as drivers of AI governance preferences.

The analysis yields a number of intriguing findings, based on a conjoint experiment in which respondents are asked, in each round, to compare two algorithms with specified characteristics. Then, they (i) decide which algorithm in each pair should be prohibited and (ii) assess each algorithm on a 7-point scale.[7] Economic considerations affect these outcome variables as we might expect: Higher risk of experiencing unemployment increases the propensity to prohibit and lowers the assessment score for an AI algorithm, whereas higher levels of job creation in the economy as a whole (due to the introduction of the AI) lower the propensity to prohibit and increase the assessment scores for the AI algorithm. These effects are very consistent across the five countries and show that economic consequences of the introduction of digital technologies are politically salient – not just as mediated by the competition via political parties (König & Wenzelburger, 2019) but

directly: expected economic consequences drive preferences for technology policy.

At the same time, normative considerations also matter: AI use scenarios that entail the collection and sharing of data without the user's *ex ante* knowledge and consent (thus violating privacy norms), as well as black box algorithmic decisions (where the functioning of AI is not explained or cannot be understood, not even by outside experts) lower the approval of the AI use. These characteristics also increase the probability that participants in Ehret's experiments choose a scenario for regulatory intervention (prohibiting AI use). Across all five countries, both normative and material/economic considerations thus appear to drive public preferences, though the effect of normative considerations is greater than the effect of economic considerations – but only in Germany and the UK, consistent with theories about post-materialist values (Inglehart, 1990). When looking separately at different measures of compliance or violations of social norms, however, cross-national differences no longer neatly fit a pattern consistent with stereotypes: For instance, Indian and Chinese citizens respond to violations of privacy norms by apps involving AI almost as strongly as UK citizens, whereas such privacy violations appear to have no effect on AI policy preferences in the OECD country Chile.

The analysis also yields a number of other intriguing findings, challenging the conventional wisdom. For example, in Chile, China, Germany and India, Ehret finds no statistically significant difference between explainability to all vs. explainability to experts, only. This finding suggests a notable willingness of citizens to defer to expert authority. Only in the UK does Ehret detect a clear preference for AI use being 'easy to explain' (rather than being explainable only to experts).

The remaining contributions focus on specific modes and mechanisms of AI governance. *Christian Djeffal, Markus Siewert and Stefan Wurster* (2022) analyze governments' national AI strategies in order to reconstruct from them the role of states and states' stances on their responsibility regarding AI. They find that states mainly fall into one of two camps: market-driven states that rely mostly on self-regulation by AI developers and implementers on one side, and entrepreneurial and regulatory states on the other.

Whereas Djeffal *et al*. focus on the plethora of measures to govern AI taken by governments on behalf of states, *Graeme Auld, Ashley Casovan, Amanda Clarke, and Benjamin Faveri* (2022) point out the increasingly prominent role of non-governmental initiatives for AI governance. In light of this development, they argue that 'AI governance cannot [anymore] be understood through the lens of state-centric policy[-making] theories or governance models alone' (2022, p. 1837).

Auld *et al*. note the diversity of non-governmental stakeholders active in this issue space, as well as the broad range of AI governance issues those actors have sought to address, but they concentrate on corporate and civil society actors governing AI *ethics* (broadly conceived) through the development, implementation, promotion, and certification of standards for ethical AI.

Similar to private governance initiatives in many other issue areas, non-governmental governance of AI ethics initially sought to prevent or push back against governmental governance. For companies, this 'oppose and fend off [governmental measures]' approach, Auld *et al*. argue, offered an opportunity to minimize state intervention in business activity and to retain a high level of autonomy. At the same time, supporting and engaging with firm- or industry-level governance initiatives allowed the companies to reduce the risk that their use of AI might result in reputational damage that might arise from being seen as deviating from broadly accepted norms. For civil society organizations, this approach (implemented in distinct, yet similarly exclusive institutions) offered an opportunity to bring pressure to bear directly on commercial targets, avoiding slow, often bureaucratic processes of governmental governance, which might be biased against non-commercial stakeholders.

Notwithstanding the continued efforts by some to keep governments at bay, Auld *et al*. detect in recent years a shift toward a different, 'engage and push' approach. Rather than try to forestall state-based governance, AI stakeholders pursuing this approach seek to feed information and specific proposals into governmental rule-making processes, given a perceived need to have governments develop AI rules to address inherently transnational issues at a global level (see also Büthe, 2022) and government's superior ability to ensure stability and a level playing field vis-à-vis competitors. This shift from 'oppose and fend off' to 'engage and push' is important because it implies changes in *how* AI should be governed and what the respective role of the state and various private interests should be in the process. Each approach, moreover, implies a distinct choice of governance venue, which institutionalizes the relative power of these actors in the governance of AI and the kinds of issues to be addressed by AI governance (as well as the mechanisms through which they are addressed).

At the same time, AI governance is still very much in flux, and Auld *et al*. identify several sources of instability that make continued change in AI governance arrangements likely. Auld *et al*., therefore, ask what AI governance might look like in the years to come, focusing on the relationship between governmental and private actors. To answer that question, Auld *et al*. draw on the broader literature on private governance (see, e.g., Auld, 2014; Bartley, 2018, 2021; Büthe & Mattli, 2011; Grabs *et al*., 2021; Green, 2014; Vogel, 2005) to identify three ideal-typical 'pathways' AI governance might take. The starting point in each case is distinct corporate and civil society motivations. Auld *et al*. submit that each approach implies a distinct set of

tactics and strategies; different kinds of interactions, and hence different governance venues (the first two are 'oppose and fend off' and 'engage and push'; the third approach is the 'lead and inspire' pathway, for which they see little evidence in the realm of AI governance yet, but which might yet become a possibility in light of developments in private governance in other fields). The authors thus provide a framework for thinking about what kind of AI governance might emerge from public-private interactions.

Seung-Hun Hong, Jonghan Lee, Sanghoon Jang, and Ha Hwang (2022) focus on the regulation of a particular application of AI, examining how South Korea and the UK have regulated the use of AI to enable autonomous vehicles (AVs). This specificity allows Hong *et al*. to conduct a thorough comparative analysis of numerous aspects of the regulatory environment for AI-driven AVs, consistent with our call for focusing on the specific applications of AI to fully understand the distinctive challenges that AI governance needs to address, and the pros and cons of different approaches to governing AI in its specific applications.

The analysis of AV-specific AI governance, at the same time, also yields broader insights. Disruptive innovative technologies such as AI-powered autonomous vehicles are rapidly evolving. Their fast-changing specifics – from details of the product design to the risks these innovations pose to society or to specific individuals – make regulating such technologies both necessary and difficult. These difficulties are exacerbated, Hong *et al*. show, by focusing regulatory efforts on specifics such as product or algorithm design, on rules as opposed to principles, etc. More flexible regulation may be a way to overcome these challenges and possibly even encourage (or at least not impede) further innovation.

Yet, we lack a way to measure 'flexibility' across the regulatory space. That space has at least three dimensions, which Hong *et al*. call rule structure, enforcement structure, and regulatory feedback, each of which has several components. To address the lack of broadly usable measures and thus allow systematic comparisons of flexibility in the regulation or governance of AI across countries and across the three dimensions. Hong *et al.* develop an index or indicator of regulatory flexibility which promises to be useful far beyond the issue of AV-specific AI use and its governance.

The co-production of AI technology and AI governance

Reading across the articles and combining their observations with our own, we highlight two aspects of AI governance that attest to the many ways in which AI technologies and AI governance co-produce each other: frames and the level of abstraction used in thinking about the key issues AI governance needs to address.

Frames of governance: creative mirrors

Technology governance depends on assumptions, normative preferences and stances taken towards the technology in question and towards society. Laws, regulations and other measures to govern AI thus do not so much reflect inherent characteristics or objective truths about the technology, but reflect political actors' *perceptions* given those actors' predisposition – and in the process of framing what needs to be addressed, they construct the technology. Measures taken to govern AI are thus akin to what Meineke, drawing on a metaphor from Goethe, called '*schaffende Spiegel*' ('creative mirrors'; Meinecke, 1948, p. 7; see also Hofmann, 2018, p. 14).

The importance of framing extends to scholarly analyses of governing AI. *Ehret* (2022), for example, frames societal changes through AI in a certain way before comparing citizens' preferences. A variety of frames concerning what AI actually means shows in the articles. While *Nitzberg and Zysman* take an open view from a liberal democratic perspective without limiting themselves to one frame, *Auld et al.* look into the development of ethical frames and reconstruct several general stances in that regard.

Issues: general or specific AI governance?

Discussions of the key issues in AI governance, including discussions of the ethical principles that might provide a moral compass for the governance of AI, often take place at very high levels of abstraction. This is appropriate insofar as fairness, justice, privacy, etc. are important issues and principles of broad, general applicability. Even the risk of economic or political harm from the concentrated power of platforms, emphasized by Nitzberg and Zysman (2022; see also von Ingersleben-Seip and Georgieva, forthcoming) is an issue that arises across a broad range of AI uses. A focus on such general aspects, however, has consequences for the components of AI, which technology scholars and policymakers are likely to 'see' and address in AI governance.[8] Approaching the issue of AI governance at a very highlevel of abstraction pushes us toward a focus on the most general characteristics or features of AI (and vice versa), whereas approaching AI governance on a more specific, concrete level is likely to result in a focus on the differentiated governance issues concerning *specific* AI *applications*.

Herding cats?

Science and Technology Studies have highlighted the crucial importance of governance for societal shaping and making of technology. Accordingly, the democratic ideal of AI governance should give all who are affected by the introduction of AI – and possibly also those who in the absence of AI

might miss out on important opportunities for improved well-being – a chance to shape the trajectory of AI technology development.[9]

AI governance, however, is enormously challenging because AI is a rapidly changing, new technology. Such moving targets are a tremendous challenge for public, democratic governance even under otherwise conducive circumstances (Abbott & Snidal, 2009; Fenwick *et al*., 2017; Vincent *et al*., 2015). Yet, AI is not just a moving target. Governing AI is an attempt to govern a technology that is fundamentally geared toward learning and adapting, *on its own*, at speeds that increasingly exceed human capacity to do so. Governing AI therefore is – and surely will increasingly become – akin to herding cats: If autonomous learning is a defining (even if still quite limited) characteristic of AI, then governing AI is an attempt to regulate something with a (more or less prominent) mind of its own.

Therefore, in lieu of drawing firm conclusions from the research presented in this special issue, we do something only slightly less risky: offer a sketch of a research and policy agenda for AI governance *beyond* what is covered in this special issue.[10]

In Lieu of a conclusion: an AI governance research agenda

In-Explainable AI?

Machine learning was not conceived to be readily understood by average citizens. In more recent years, researchers have consciously tried to reconceptualize this aspect of AI by developing the notion of 'Explainable AI' (e.g., Gunning *et al*., 2021; Samek *et al*., 2019). The literature on AI and its governance now often advocates Explainable AI as the panacea for both public acceptance and technology development in the public interest. Explainable AI narrowly conceived might merely mean that it can in principle be explained how the AI algorithm works – maybe even in ways and in language comprehensible to a lay audience (see Ehret's distinction between 'easy to explain' and 'only experts can understand '; see also König *et al*., 2022). Indeed, recent research suggests that it is possible, *post hoc*, to reconstruct how an AI system arrived at its recommendations, decisions, or other outcomes it generated (Mertes *et al*., 2022; Molnar, 2022; Weitz *et al*., 2019), though such an *ex post* reconstruction requires great effort and comes at a considerable costs – and hence is only rarely undertaken.[11]

Often, however, 'Explainable AI' seems to carry much more demanding connotations, suggesting that we can understand how a particular AI algorithm works in such a way that the recommendations or decisions resulting from the use of that algorithm become predictable. Unfortunately, this idea is misleading. Many AI algorithms 'search' for a solution in various, nonlinear ways, such that it is *not* guaranteed that they will arrive at the same

result every time. Moreover, AI is powerful as a (self)learning system precisely because (or when) it continuously takes new information into account, which means that very small changes to the data may prompt it to give a different answer to the same question (see, e.g., de Marchi *et al*., 2004, p. 372f).

Sometimes 'Explainable AI' is understood to go even further, namely to allow a human being to follow along and review the AI-based decisionmaking in real time and then approve or reject the AI-generated recommendation not just based on human intuition or in comparison with alternative courses of action or choices suggested by alternative solution methods (such as linear, matrix algebra- or calculus-based optimization) but based on having understood and followed each step in the AI-based decisionmaking in every case. Expecting or even demanding, as part of an AI governance framework, such a real-time human review before any AI-based recommendation is put to use, would undermine key benefits of using AI. AI may be far from being able to perform at the level of human cognition and emotional intelligence. But the speed with which computers are able to manage, process, and analyze massive quantities of data now exceeds the human capacity to take these steps, and computational power seems virtually guaranteed to continue to increase. Demanding that the operation of AI 'slow down' so people can follow along in each case of its use defeats the purpose of using AI.[12]

Fortunately, the expansive (unrealistic) view of Explainable AI may not be a prerequisite for legitimate and responsible AI. As Mann (1993) pointed out, there are many aspects of modern life – including seemingly mundane aspects, such as our fresh water supply, garbage collection and disposal, and mail delivery – where non-expert citizens usually do not fully understand how they work. We simply accept that they (appear) to work, and we delegate to the 'experts' the task of figuring out the details (combined with some kind of accountability mechanism to deal with evidence to the contrary, should such evidence arise). It is conceivable that AI use will at some point be viewed very similarly, but much more research is needed to identify the conditions under which such delegation may be acceptable to citizens and under what conditions it yields good results.

Ethics & governance

As noted in the introduction, AI raises new, distinct governance challenges, as well as familiar governance challenges in arguably novel ways (see also Djeffal, 2020). It has long been the case, for instance, that states find it difficult to predict problems within society and formulate adequate regulations *ex ante* (Eisenberg, 1992). This has become ever more apparent since the speed of innovation has increased exponentially since the last century (Owen *et al*., 2013), albeit in diverse, complex ways (Breznitz, 2020).

AI has further exacerbated this dilemma for policymakers and others seeing public governance as the most appropriate way to address those problems.[13] Moreover, as a consequence of the issues sketched above – such as the rapid changes in AI technology, the massive diversification of AI applications, and even the co-constitution of AI technology and AI governance – many important aspects of AI are not (or insufficiently) addressed by law or other forms of public governance (Bureau & Dieuaide, 2018). Therefore, to ensure that AI is researched, developed and implemented in a way that serves – rather than decreases – the good of society, it is ever more important to consider AI *ethics* and AI *governance* simultaneously.

Companies, research institutions and public sector bodies have started to develop a wealth of principles and guidelines for ethical AI. European initiatives include the European Commission's (2020) White Paper on Artificial Intelligence, which highlights steps to achieve an 'ecosystem of excellence and trust,' and AI4People's core principles for AI usage (Floridi et al., 2018). They also include AI4People's proposal concerning 14 priority actions, models of AI governance, and regulatory toolboxes (Pagallo *et al*., 2019) – as well as the Ethics Guidelines for Trustworthy AI issued by the High-level Expert Group on Artificial Intelligence (2019). Identifying a set of core values and principles that should guide the development of AI, however, is not just a concern in Europe or the Global North. Global efforts include the AI principles developed by the OECD (2019), as well as the IEEE's (2019) ethical considerations in the design of autonomous and intelligent systems.[14]

Overall, there seems to be a near-global consensus on five key ethical principles, namely: transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin *et al*., 2019). All of these principles and guidelines are important for laying down the ethical foundation and for providing a moral compass on how to proceed and what to consider in the development of AI, especially when the exact impacts of particular AI applications are not predictable, and consequently, concrete laws have not been formulated yet. Despite their high level of abstraction, such ethical principles can guide AI implementation in industrial engineering (see, e.g., Unruh *et al*., 2022). At the same time, with few exceptions (such as Google, Microsoft and Daimler), major AI users and developers have so far only made vague commitments regarding AI ethics (Lütge, 2020). And startups are even more unlikely to integrate ethical AI principles unless they are obliged to do so due to data-sharing relationships with high-tech firms (Bessen *et al*., 2022) – not because startups are less committed to the common good or ethical principles than established firms, but because their focus on short-term survival makes it difficult to pursue objectives where the rewards are long-term.[15]

To close the gap between abstract AI principles and concrete guidance for practical action, it is now urgent to build on AI ethics to develop more practical, concrete, use-case – specific recommendations and regulations, especially in

those fields of AI applications that are relatively firmly established. For example, in the automotive sector, policymakers need to swiftly develop a regulatory framework that, *inter alia*, clarifies what constitutes legitimate override functions or harmonizes the definition of a 'safe condition' for AVs (Lütge *et al*., 2021). Such binding regulations are often even welcomed by the private sector (consistent with Auld *et al*.'s observation of a shift away from wanting to prevent government rule-making) because they remove uncertainty about what is permitted and what restrictions must be considered, so that instead of spending time on self-regulation, manufacturers can focus their attention on development and production.

At the same time, a 'one size fits all' approach will not be sufficient. What is needed are tailored (differentiated) regulatory efforts for the many different AI applications. In developing those regulations, a starting point could be to '[…] define criteria for distinguishing between 'hard' and 'soft' rules' (Lütge, 2020, p. 6). 'Hard rules' (i.e., laws and regulations) might pertain to critical decisions and AI applications, whereas other AI applications would require only 'soft rules' such as ethics codes. To facilitate such a distinction, Germany's Data Ethics Commission (2019) has developed a 'criticality pyramid,' which ranges from AI applications with no potential for harm to applications with an untenable potential for harm.

The overarching goal of all of these measures should be to ensure socially responsible adoption of AI but, at the same time, prevent the creation of additional barriers to socially beneficial innovation. Instead of considering them separately, we propose acknowledging that AI ethics and AI governance go hand in hand; overarching ethical principles and core values are essential to adapting existing legal frameworks accordingly.

System effects in human–machine interactions

One of the most promising uses of AI is in decision-support systems (see, e.g., Campanella *et al*., 2016; He *et al*., 2019). Recent research also suggests that most people view AI much more favorably when it is employed such that it provides no more than suggestions, i.e., when the ultimate decision (and responsibility) still rests with a person. There is a risk, however, that those decisionmakers will then overly rely upon the AI support system. This may happen for a number of reasons, including an unreflected pursuit of efficiency gains made possible by AI. Most important, however, may be the risk of what Robert Jervis (1997), drawing on behavioral research in psychology and related fields, called 'system effects'.[16] The seeming certainty with which AI systems tend to present a recommendation (appearing to take most other options 'off the table') is likely to feed the assumption that the supercomputer – which provides the recommendation at breakneck speed, taking into account unimaginably large amounts of information – has

surely considered all relevant aspects. The resulting conscious or subconscious expectation that the risk of making a mistake has declined tends, in turn, to make human agents more risk-acceptant than they otherwise would be (see also Leicht-Deobald *et al*., 2019, pp. 384–386). Figuring out how to avoid such harmful system effects will be an important contribution to effective AI governance.

AI in international politics

An AI application might yield quite different results in different contexts when the algorithm has been 'trained' on culturally specific data. Attempts to establish national 'sovereignty' over many kinds of data may therefore unintentionally result in a more fragmented digital world. And censorship or similar forms of government restrictions on cross-border information exchange (Milner, 2006; Flonk, 2021) may intentionally heighten fragmentation by prompting the same AI to yield different results in different political contexts. Specific AI applications may, moreover, have a language- or culturally specific interface. The algorithmic code at the heart of AI, however, moves readily across borders – akin to other digital goods and services (Büthe, 2022). In the absence of political interference, AI might therefore spread rapidly once the advantages of specific AI applications become apparent (Drezner, 2019; Milner & Solstad, 2021), and it may affect domestic political and economic structures along the way, as other transformative technologies have done in the past (Hintze, 1975 (1906)).

While the general dynamics might be familiar from previous cases in which disruptive technologies were introduced into international politics, the specifics of how AI is likely to change international politics have only just started to attract serious scholarly attention (Allen, 2022; Chen, 2022; Cheng & Zeng, 2022; Ding & Dafoe, 2021; Kerry *et al*., 2021; Meltzer, 2021; Meltzer & Tielemans, 2022; Siegmann & Anderljung, 2022; Von Ingersleben-Seip, 2022). The development of autonomous weapons systems, which might circumvent defensive structures and technologies, possibly shifting the offense-defense balance in militarized disputes, is a prominent possibility (e.g., Gill, 2019). AI might also change international politics indirectly, for instance via its effect on economic growth. Bughin *et al*., estimate that AI could deliver additional economic output of US$13 trillion by 2030, but these benefits will be distributed very unevenly, with most of the gains accruing to developed countries.[17] At the same time, AI-enabled capabilities may allow rising powers to challenge the traditionally dominant powers (Roberts *et al*., 2021). More research is needed to understand these and other possible channels through which AI may transform international politics.

AI is also prompting changes in the patterns of competition and cooperation. Technical experts from countries across the world, including from geopolitical rivals, such as the United States and China, are collaborating on technical standards for AI within transnational standard-setting organizations. Governments, however, are much less willing to cooperate and enter into global agreements to govern AI. The lack of inter-governmental cooperation is particularly striking with respect to the ethics of AI (von Ingersleben-Seip, 2022). Even the UNESCO agreement on the Ethics of Artificial Intelligence, adopted by UNESCO members in 2021 and touted as the 'first ever *global* agreement on AI ethics,"[18] is not truly global, as the United States is not a signatory of the agreement. Given that one of the world's major AI powers has not signed up to the UNESCO guidelines, their impact has been limited so far. On a domestic level, many countries have taken steps toward technological sovereignty, reducing their dependence on geopolitical rivals. For example, in light of Chinese-American trade frictions, China has increased its investments into the AI infrastructure layer centered on AI chips, computing power and data platforms in order to be less dependent on American technologies (Ding, 2022).

The inherently transnational nature of AI technology implies that its governance might also need to take place, at least in part, at the inter- or transnational level. While this need has been recognized in some international fora, such the OECD, public, governmental governance of AI tends to be so far focused very much at the national level, as shown by Djeffal *et al*. (2022). And the partial but substantial breakdown in inter-state cooperation in the early months of the COVID-19 pandemic suggests that a grave crisis might be needed to prompt international cooperation on a broad scale with simultaneous serious depth (Bremmer, 2022). Non-governmental initiatives have gone much further in moving AI governance to the transnational level, as noted by Auld *et al*. (2022), but it is at best an open question whether substantial inter- or transnational AI governance can be achieved without governments as parties and enforcers of any agreement. And given that countries' goals with regard to the development and use of AI vary widely, it is currently hard to imagine that there will be meaningful international agreements for the governance of AI beyond relatively small subsets of countries. This absence of genuinely global governance is fueled by intense cyber rivalries that risk splintering the digital world into two blocks (United States vs. China) or perhaps even three (United States vs. Europe vs. China). If the digital world in fact ends up splitting into two (or three) different blocks, the question is whether these blocks will be rigidly separated or interoperable (Zysman & Nitzberg, 2020, p. 21). One potential outcome is that the two (or three) blocks will have a shared technical infrastructure layer but widely divergent application and content layers.[19] In sum, research on the

promise and limits of intergovernmental cooperation on AI governance – and the conditions under which it can be achieved – is greatly needed.

Risks and opportunities

AI, as a special case of the broader phenomenon of digitalization, has wide-ranging transformative potential (Maasen & Sutter, 2020, pp. 84ff; Zysman & Newman, 2006). Among the consequences is an increase in several kinds of risks: AI increases cyber security risks, may bring radical changes in the workplace, and enables 'deep fakes', as well as new threats from autonomous weapons.[20] These risks must be taken seriously.

At the same time, many AI applications hold tremendous promise for improving our quality of life, reducing risks, and enabling qualitative leaps in factor productivity and hence economic growth, with simultaneous reductions in environmental harm. Depriving citizens of those benefits (by preventing or substantially delaying the development of AI technology) has detrimental consequences that might be equally or more severe than the detrimental consequences that citizens or policymakers might seek to forestall by inhibiting or preventing the technology. The losses caused by preventing the introduction or development of the technology are, however, often invisible because the social, economic and health benefits of a technology often do not become fully apparent until quite far into the development of any given new technology (Fenwick *et al*., 2017).

Moreover, even within a single issue area, AI itself can play out very differently. For instance, AI is the basis of systems of surveillance that resemble the worst nightmares of privacy scholars and activists. At the same time, many privacy-enhancing technologies are based on AI technologies. This duality underscores the need to consider AI both a risk and an opportunity for social values. And such duality, can be found with regard to multiple social values such as opacity and transparency, discrimination and equality, as well as environmental sustainability and pollution. Interestingly, disadvantaged groups most immediately affected by a technological innovation may focus to a much greater extent on the benefits than the average citizen or voter, as Schönmann *et al*. (2022) have shown regarding assessments of care robots: People who depend upon extensive care and support in their daily lives show much greater awareness of the potential upsides of care robots than the general population. As technical and social AI innovations continue, it becomes ever more important to govern such novel technologies in ways that prevent harm and contain risks without depriving people of benefits and foreclosing opportunities for improving wellbeing.

The challenge is exacerbated by public discourses that are, especially in the realm of European public policy, often dominated by a focus on downside risks of change and on prohibitions. Identifying ways to elevate public

discourse so as to take both the risks and the opportunities seriously would be a very important contribution to both scholarship and practice of the regulation and governance of AI.

Governing not the technology but its (ab)uses

The articles in this special issue focus mostly on governing AI as a technology. This focus is in keeping with the dominant approach to technology governance, even in light of sophisticated inquiries into whether technologies and technical artifacts have 'agency' or 'politics' (Latour, 2007; Winner, 1980). Yet, it is not the only way to address the issues raised by AI.

An alternative approach might focus on how human beings and organizations concretely *use* AI. The primary goal of AI governance, after all, is not to steer the development of the technology as such, but to avoid bad and achieve better outcomes or consequences of particular uses of AI. Health-related data, for instance, is in many countries seen as particularly sensitive. AI-based analyses of such data can be used to estimate personalized risk profiles, which might then be used to deny higher-risk individuals employment, health insurance, or other benefits. Consequently, health and medical data tends to be subject to some of the strongest privacy and data security requirements. Gathering such data, combining it with other data, and analyzing health and medical data is therefore in many countries highly restricted or even prohibited. Such regulation of the technology, however, also inhibits advances in personalized medicine, delays progress in finding treatments for rare diseases, and prevents the early detection of, and effective policy responses to, population health risks and threats, such as epidemics and pandemics. It might therefore be much better, from a public interest perspective, to loosen restrictions on the data and the tools for its analysis and focus instead on prohibiting the denial of insurance coverage on health grounds (maybe complemented by public subsidies for high-risk insured persons), health-based employment discrimination, and other abuses – and vigorously enforcing such prohibitions.

Future research might hence focus on ways of governing the use and punishing the abuse of AI, rather than governing its primarily technical aspects. While surely not without its own challenges, this approach also might offer a fruitful way to deal with the unpredictability, the rapid changes, and the technical complexity of AI while reducing the need to restrict its technological development.

Notes

1. For instance, the April 2021 proposal for an EU Regulation commonly known as "the AI Act," specifies a number of rights and obligations for individuals and groups vis-a-vis the developers and users of AI applications and systems,

conditional on the risk categorization of the AI systems – and on what constitutes an 'AI system' (defined in Arts. 4–6 and Art.3, respectively, and their annexes). German works council representatives report efforts by employer associations to narrow the definition of AI in the draft legislation in such a way that employers' transparency obligations and employees' participation rights regarding the AI are limited to the machine learning components and do not apply to the rest of the algorithms, nor to any other parts of the technology stack, in which these components are embedded. How AI is defined thus matters for human resource management, including performance monitoring and evaluation, as well as hiring, promotion, and termination (not-for-attribution in-person and email communications by Tim Büthe with German works council representatives, Munich, August/September 2022).

2. One of the authors later reported that they chose the term to "get away from studying human behavior and to consider the computer as a tool for solving certain classes of problems" (McCarthy, 1989, p. 6).
3. Key to the autonomous or "intelligent" capabilities of such systems (see, e.g., Corea, 2019) are computational methods that allow for improvements in the ability to achieve the specified objective, which may occur via human-controlled ("supervised") or even entirely automated ("unsupervised") feedback into the algorithmic processing. Such improvements are often called "machine learning," which uses neuro-biologically inspired computational architectures, such as neural networks, to achieve ever better recognition of patterns in data, including clusters, sparseness, and linear and non-linear correlations across time and space (see, e.g., Bishop, 2006; for an early discussion of possible uses (and limitations) in Political Science and International Relations, see Beck *et al*., 2000; de Marchi *et al*., 2004).
4. For the European OECD member states, these efforts are examined in detail in the essay by Djeffal *et al*. (2022).
5. For a broader discussion of transnational, mostly non-governmental efforts, see the essay by Auld *et al*. (2022).
6. This approach reflects the liberal tradition (Moravcsik, 1997). For a compelling treatment, problematizing the liberal tradition beyond what we can cover in this essay, see Katznelson (1996).
7. Generalizability may be limited by the modest sample size and the non-random opt-in sampling, though cross-country comparability appears to be retained, as the country-level sub-samples are non-representative in similar ways (e.g., due to the over-representation of younger, highly educated participants).
8. The issue raised here is not just an issue for public policy: Lütge (2019) points out that companies and other stakeholder find it difficult to meaningfully adopt and actually implement principles at such a high level of abstraction into their business practices.
9. We recognize, of course, that the use and governance of AI are by no means just undertaken by democratic states and societies (see Horowitz, 2016; Feldstein, 2019; Gravett, 2020).
10. Our ability to offer firm conclusions is also limited by the diversity of substantive foci, approaches, and methods in this special issue. To advance the social-scientific understanding of AI, we have brought together papers that examine distinctive challenges and opportunities in the governance of AI from a variety of perspectives, focused on political and public policy aspects of the topic. The articles in the special issue also use a variety of approaches and

methods; they look at different instruments and other aspects of governance such as public attitudes, national strategies, ethical standards, as well as public and private regulation. The tremendous diversity enriches the special issue. At the same time, given that the articles all differ from each other in substantive focus, as well as theoretical approach, methods, etc., we cannot offer a comparative final analysis of the pros and cons of focusing governance efforts on different governance mechanisms, nor a conclusive discussion of tradeoffs (e.g., between regulatory flexibility and standardization) or a synthesis of big policy implications.

11. Conceptualized in this way, Explainable AI may be akin to a stress test, conducted on a small sample as part of quality control – though cf. Bryson and Winfield's proposal that, when AI-driven systems, such as care robots, act in unexpected ways, users 'should be able to ask it "Why did you just do that?" and receive an intelligible reply' (2017, p. 119).
12. Tim Büthe thanks Johannes Fottner, Charlotte Unruh and Charlotte Haid for fruitful discussions on this issue. Note that this is a distinct issue from the 'slow AI' movement's concern with the speed of (and priorities in) technology development rather than technology application or operation; see Auld *et al.* (2022, p. 1833).
13. We do not suggest that private governance is inherently unproblematic or necessarily more efficient or effective, just that the problems are different (see, e.g., Balleisen, 2009; Mosley, 2009; Büthe, 2013; Schleifer *et al.*, 2019; Grabs *et al.*, 2021; Kinniburgh *et al.*, 2022, forthcoming), and the literature on regulation and governance has historically mostly focused on state regulation and public governance.
14. Although these efforts are in principle global, stakeholders from the Global North have been much more vocal and visible, at least in the discourses conducted and published in Western languages. As Auld *et al.* (2022, pp. 1835–1838) point out, such preeminence of stakeholders from certain regions of the world can lead to "localization effects," which undermine the output legitimacy of governance efforts when the distinctive concerns of other regions are not appropriately addressed (see also DeMenno and Büthe's discussion of the "political information"-based policy learning channel for exerting influence over governance agendas and outcomes (2022, esp. pp. 61–63).
15. Data sharing agreements might incentivize more ethical AI, but at a cost: they force startups to share valuable information with these firms and therefore potentially reduce market competition.
16. Jervis provides a compelling example in his introduction (1997, pp. 7f, drawing in part on C. Perrow's *Normal Accidents*): After the Exxon Valdez oil spill disaster in Alaska in 1989, the U.S. government required new oil tankers that were going to use U.S. ports to be double-hulled. Having invested in the more expensive vessels, shipping companies and/or captains used the extra protection afforded by the double hulls to go faster and take greater chances, thus mostly eliminating the hoped-for environmental benefits.
17. A notable exception to the projected pattern is China, which has been investing heavily into becoming a global leader in the AI supply chain (Bughin et al., 2018, pp.3f) – though note that Chinese private AI investments declined during the COVID pandemic in 2019 and 2020 (Ding, 2022), and many economically disadvantaged Chinese provinces may not be on track to meet the ambitious goals set by the central government for the growth of their respective AI industries (Hine & Floridi, 2022, forthcoming).

18. See https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence. Emphasis added.
19. We thank Andrea Renda for pointing out this scenario.
20. For a recent overview, see Bartneck *et al*. (2021).

Acknowledgements

We thank all the authors who have submitted papers for this special issue; the numerous reviewers who through their vigorous yet constructive criticisms have pushed us all to improve the papers; the presenters, discussants, and numerous active participants of the Workshop 'Governing New Technologies: Technological Innovations, Market Competition and Public Policy,' held at the TUM Akademiezentrum Raitenhaslach, 10–12 February 2020, jointly organized by Zlatina Georgieva, Nora von Ingersleben-Seip and Tim Büthe with TUM School of Management Professor for the Economics of Innovation, Prof. Dr. Hanna Hottenrott. Some of the ideas set out in the concluding section of this essay were first developed and/or tested out in the context of discussions of the EU H2020-Project 'TRends In Global Governance and Europe's Role' (TRIGGER), the Brookings-CEPS-led transatlantic Dialogue on AI, and/or the 'Human Preference-Aware Optimization' project, led by Tim Büthe with Prof. Dr. Johannes Fottner and funded by the Institute for Artificial Intelligence. We are also grateful to Graeme Auld and John Zysman for helpful comments on previous drafts and the editors-in-chief of *JEPP*, Berthold Rittberger and Jeremy Richardson, for their support of the project.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We are grateful for financial support from the Hochschule für Politik München at the Technical University of Munich, TUM (Chair for International Relations, Prof. Dr. Tim Büthe) and the TUM Raitenhaslach Fund.

Notes on Contributors

Tim Büthe is Professor and Chair for International Relations, Hochschule für Politik München (Munich School of Politics and Public Policy) at the Technical University of Munich (Germany) and the TUM School of Management. He also is Technology Policy Scholar at the Sanford School of Public Policy, Duke University. Most of the work on this special issue was done while he was also a Senior Fellow of the Kenan Institute for Ethics, Duke University (USA). He serves as the corresponding editor for this special issue and corresponding author for this introductory essay.

Christian Djeffal is Assistant Professor for Law, Science and Technology at the Technical University of Munich (Germany).

Christoph Lütge is Professor and Chair for Business Ethics, as well as the founding director of the Institute for Ethics in Artificial Intelligence at the Technical University of Munich (Germany).

Sabine Maasen is Professor and Chair for Science Studies and Innovation Research at the University of Hamburg (Germany). Her work on this special issue was begun while she was Professor and Chair for Sociology of Science at the Technical University of Munich and director of the Munich Center for Technology in Society (MCTS) at TUM.

Nora von Ingersleben-Seip is a Doctoral Candidate at the Chair for International Relations at the Hochschule für Politik/TUM School of Social Sciences and Technology. Most of her work on this special issue was conducted while she was a Researcher for the EU H2020 TRIGGER project.

ORCID

Tim Büthe http://orcid.org/0000-0002-4724-5000
Christian Djeffal http://orcid.org/0000-0003-2098-7239
Christoph Lütge http://orcid.org/0000-0002-3870-4789
Nora von Ingersleben-Seip http://orcid.org/0000-0003-3708-4757

References

Abbott, K. W., & Snidal, D. (2009). The governance triangle. In W. Mattli & N. Woods (Eds.), *The politics of global regulation* (pp. 44–88). Princeton University Press.

Ada Lovelace Institute. (2022, forthcoming). *Inclusive AI governance: Civil society participation in standards development.*

Allen, G. C. (2022). *One key challenge for diplomacy on AI? China's military does not want to talk*. Commentary from the Center for Strategic and International Studies.

Alshadafan, A. (2020). Energy efficiency standards: The struggle for legitimacy. *International Journal of Standardization Research*, *18*(1), 1–23. https://doi.org/10.4018/IJSR.20200101.oa1

Armstrong, S., Bostrom, B., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & Society*, *31*(2), 201–206. https://doi.org/10.1007/s00146-015-0590-y

Armstrong, S., & Pamlin, D. (2015). *12 Risks that threaten human civilization*. Global Challenges Foundation. Retrieved September 10, 2022, from https://www.pamlin.net/material/2017/10/10/without-us-progress-still-possible-article-in-china-daily-m9hnk.

Auld, G. (2014). *Constructing private governance: The rise and evolution of forest, coffee, and fisheries certification*. Yale University Press.

Auld, G., Casovan, A., Clarke, A., & Faveri, B. (2022). Governing AI through ethical standards: Learning from the experience of other private governance initiatives. *Journal of European Public Policy*, *29*(11). https://doi.org/10.1080/13501763.2022.2099449

Babcock, J., Kramar, J., & Yampolskiy, R. V. (2019). Guidelines for artificial intelligence containment. In A. E. Abbas (Ed.), *Next-generation ethics* (pp. 90–112). Cambridge University Press.

Bächtiger, A., Dryzek, J. S., & Mansbridge, J. (2019). *Oxford handbook of deliberative democracy*. Oxford University Press.

Balleisen, E. J. (2009). Private cops on the fraud beat: American business self-regulation and its discontents, 1895-1932. *Business History Review*, *83*(1), 113–160. https://doi.org/10.1017/S0007680500000222

Bartley, T. (2018). *Rules without rights: Land, labor, and private authority in the global economy*. Oxford University Press.

Bartley, T. (2021). Power and the practice of transnational private regulation. *New Political Economy*, *27*(2), 188–202. https://doi.org/10.1080/13563467.2021.1881471

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *An introduction to ethics in robotics and AI*. Springer.

Beck, N., King, G., & Zeng, L. (2000). Improving quantitative studies of international conflict: A conjecture. *American Political Science Review*, *94*(1), 21–36. https://doi.org/10.2307/2586378

Bessen, J., Impink, S. M., & Seamans, R. (2022). The cost of ethical AI development for AI startups. *Proceedings of 2022 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (Oxford, UK), 92–106. https://doi.org/10.1145/3514094.3534195

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

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

Bremmer, I. (2022). *The power of crisis: How three threats – and our response – will change the world*. Simon and Schuster.

Breznitz, D. (2020). *Innovation in real places: Strategies for prosperity in an unforgiving world*. Oxford University Press.

Bryson, J., & Theodorou, A. (2019). How society can maintain human-centric artificial intelligence. In M. Toivonen & E. Saari (Eds.), *Human-centered digitalization and services: Translational systems sciences* (Vol. 19; pp. 305–323). Springer.

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer 50*(5), 116–119. https://doi.org/10.1109/MC.2017.154

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy*. McKinsey Global Institute Discussion Paper.

Bureau, M. C., & Dieuaide, P. (2018). Institutional change and transformations in labour and employment standards: An analysis of 'grey zones'. *Transfer: European Review of Labour and Research*, *24*(3), 261–277. https://doi.org/10.1177/1024258918775573

Busse, F., & Baeva, G. (2022). *Was sind die richtigen Zutaten für vertrauenswürdige Künstliche Intelligenz? Ergebnisse der ZVKI-Online-Befragung: Wissen, Nachvollziehbarkeit und bewertbare Erfahrungen – Zutaten für vertrauenswürdige Künstliche Intelligenz (KI)*.

Büthe, T. (2013). *Distributional consequences of transnational private regulation: institutional complementarity as a structural source of power in global product and financial markets*. Duke University Rethinking Regulation Working Paper no.6. https://doi.org/10.2139/ssrn.2238100

Büthe, T. (2022). De-Globalisierung in der Standardisierung und Governance digitaler Technologien? In S. A. Schirm, A. Busch, S. Lütz, S. Walter, & H. Zimmermann (Eds.), *De-Globalisierung: Forschungsstand und Perspektiven* (pp. 279–301). Nomos Verlag. https://doi.org/10.5771/9783748932901-279

Büthe, T., & Mattli, W. (2011). *The new global rulers: The privatization of regulation in the world economy*. Princeton University Press.

Campanella, P., Lovato, E., Maronoe, C., Fallacara, L., Mancuso, A., Ricciardi, W., & Specchia, M. L. (2016). The impact of electronic health records on healthcare quality: A systematic review and meta-analysis. *European Journal of Public Health*, *26*(1), 60–64.

Chen, J. (2022). Cyber and influence operations. In W. C. Hannas & H. Cheng (Eds.), *Chinese power and artificial intelligence: Perspectives and challenges* (pp. 189–215). Routledge.

Cheng, J., & Zeng, J. (2022). Shaping AI's future? China in global AI governance. *Journal of Contemporary China*, *31*(137), 1–17. https://doi.org/10.1080/10670564.2022.2107391

Christou, G., & Simpson, S. (2011). The European union, multilateralism, and the global governance of the internet. *Journal of European Public Policy*, *18*(2), 241–257. https://doi.org/10.1080/13501763.2011.544505

Collier, D., & Mahon, J. E. (1993). Conceptual 'stretching' revisited: Adapting categories in comparative analysis. *American Political Science Review*, *87*(4), 845–855. https://doi.org/10.2307/2938818

Corea, F. (2019). *An introduction to data: Everything you need to know about ai, big data and data science*. Springer.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press.

Crafts, N. (2021). Artificial intelligence as a general-purpose technology: An historical perspective. *Oxford Review of Economic Policy*, *37*(3), 521–536. https://doi.org/10.1093/oxrep/grab012

Datenethikkommission der Bundesregierung. (2019). *Gutachten der Datenethikkommission*. https://datenethikkommission.de/wp-content/uploads/191128_DEK_Gutachten_bf_b.pdf.

de Marchi, S., Gelpi, C., & Grynaviski, J. D. (2004). Untangeling neural nets. *American Political Science Review*, *98*(2), 371–378. https://doi.org/10.1017/S0003055404001200

DeMenno, M. B., & Büthe, T. (2022). Voice and influence in global governance: An analytical framework. In J. Pauwelyn, M. Maggetti, T. Büthe, & A. Berman (Eds.), *Rethinking participation in global governance: Voice and influence after stakeholder reforms in global finance and health* (pp. 31–70). Oxford University Press.

Ding, J. (2022). Feature translation: China AI venture capital data report (IT Juzi). *ChinAI Newsletter* #191.

Ding, J., & Dafoe, A. (2021). The logic of strategic assets: From oil to AI. *Security Studies*, *30*(2), 182–212. https://doi.org/10.1080/09636412.2021.1915583

DiNuovo, A. (2018, November 28). Here's how robots can fight loneliness and ageing. *World Economic Forum in Focus*. Retrieved September 10, 2022, from https://www.weforum.org/agenda/2018/11/robot-carers-could-help-lonely-seniors-they-re-cheering-humans-up-already.

Djeffal, C. (2019). AI, democracy, and the law. In A. Sudmann (Ed.), *The democratization of artificial intelligence. Net politics in the era of learning algorithms* (pp. 255–284). Transcript.

Djeffal, C. (2020). Artificial intelligence and public governance. Normative guidelines for artificial intelligence in government and public administration. In T. Wischmeyer & T. Rademacher (Eds.), *Regulating artificial intelligence* (pp. 277–293). Springer.

Djeffal, C., Siewert, M. B., & Wurster, S. (2022). Role of the state and responsibility in governing artificial intelligence: A comparative analysis of AI strategies. *Journal of European Public Policy*, *29*(11). https://doi.org/10.1080/13501763.2022.2094987

Drezner, D. W. (2019). Technological change and international relations. *International Relations*, *33*(2), 286–303. https://doi.org/10.1177/0047117819834629

Dryzek, J. S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J. N., Felicetti, A., Fishkin, J. S., Farrell, D. M., Fung, A., Gutmann, A., Landemore, H., Mansbridge, J., Marien, S., Neblo, M. A., Niemeyer, S., Setälä, M., Slothuus, R., Suiter, J., Thompson, D., & Warren, M. E. (2019). The crisis of democracy and the science of deliberation: Citizens can avoid polarization and make sound decisions. *Science*, *363*(6432), 1144–1146. https://doi.org/10.1126/science.aaw2694

Ehret, S. (2022). Public preferences for governing AI technology: Comparative Evidence. *Journal of European Public Policy*, *29*(11), 1779–1798. https://doi.org/10.1080/13501763.2022.2094988

Eisenberg, J. A. (1992). *The limits of reason*. Transaction Publishers.

Elstub, S., & Escobar, O. (2019). *Handbook of democratic innovation and governance*. Edward Elgar.

European Commission. (2018). *European Commission digital strategy: A digitally transformed, user-focused and data-driven commission (C(2018) 7118 final)*.

European Commission. (2020). *On artificial intelligence – A European approach to excellence and trust*. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts (COM(2021) 206 final)*.

European Parliament. (2020). *Resolution of 20 October with recommendations to the Commission on civil liability regimes for artificial intelligence (2020/2014 (INL)), Official Journal of the European Union C 404/107*.

European Parliament, Committee on Industry, Research and Energy. (2022a). *Draft opinion, 2021/0106(COD)*. https://www.europarl.europa.eu/doceo/document/JURI-PA-719827_EN.pdf.

European Parliament, Committee on Legal Affairs. (2022b). *Draft opinion, 2021/0106 (COD)*. https://www.europarl.europa.eu/doceo/document/JURI-PA-719827_EN.pdf.

European Parliament. (2022c). *Artificial Intelligence in a digital age. European Parliament resolution of 3 May 2022 on artificial intelligence in a digital age (2020/2266(INI))* (P9_TA(2022) 0140).

Feldstein, S. (2019). The road to digital unfreedom: How artificial intelligence is reshaping repression. *Journal of Democracy*, *30*(1), 40–52. https://doi.org/10.1353/jod.2019.0003

Fenwick, M., Kaal, W. A., & Vermeulen, E. P. M. (2017). Regulation tomorrow: What happens when technology is faster than the law? *American University Business Law Review*, *6*(3), 561–594. https://doi.org/10.2139/ssrn.2834531

Fishkin, J. S. (2018). *Democracy when the people are thinking: Revitalizing our politics through public deliberation*. Oxford University Press.

Flonk, D. (2021). Emerging illiberal norms: Russia and China as promoters of internet content control. *International Affairs*, *97*(6), 1925–1944. http://doi.org/10.1093/ia/iiab146

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In S. Carta (Ed.), *Machine learning and the city: Applications in architecture and urban design* (pp. 535–545). John Wiley & Sons.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fung, A., & Warren, M. E. (2011). The participedia project: An introduction. *International Public Management Journal*, *14*(3), 341–362. https://doi.org/10.1080/10967494.2011.618309

Furman, J., & Seamans, R. (2018). *AI and the economy*. NBER Working Paper Series no.24689.

Gasser, U., Budish, R., & Ashar, A. (2018). *Artificial Intelligence (AI) for Development: Module on setting the stage for AI governance – Interfaces, infrastructures, and*

institutions for policymakers and regulators. Retrieved September 10, 2022, from https://www.itu.int/en/ITU-D/Conferences/GSR/Documents/GSR2018/documents/AISeries_GovernanceModule_GSR18.pdf

Gill, A. S. (2019). Artificial intelligence and international security: The long view. *Ethics and International Affairs*, *33*(2), 169–179. https://doi.org/10.1017/S0892679419000145

Grabs, J., Auld, G., & Cashore, B. (2021). Private regulation, public policy, and the perils of adverse ontological selection. *Regulation and Governance*, *15*(4), 1183–1208. https://doi.org/10.1111/rego.12354

Gravett, W. H. (2020). Digital coloniser? China and artificial intelligence in Africa. *Survival*, *62*(6), 153–178. https://doi.org/10.1080/00396338.2020.1851098

Green, J. F. (2014). *Rethinking private authority: Agents and entrepreneurs in global environmental governance*. Princeton University Press.

Grigorescu, A. (2015). *Democratic international organizations? Normative pressures and decision-making rules*. Cambridge University Press.

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2*(4), e61. https://doi.org/10.1002/ail2.61

Guo, J., & Li, B. (2018). The application of medical artificial intelligence technology in rural areas of developing countries. *Health Equity*, *2*(1), 174–191. https://doi.org/10.1089/heq.2018.0037

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, *69* (Supplement), 36–40. https://doi.org/10.1016/j.metabol.2017.01.011

Hasselbalch, J. A. (2018). Innovation assessment: Governing through periods of disruptive technological change. *Journal of European Public Policy*, *25*(12), 1855–1873. https://doi.org/10.1080/13501763.2017.1363805

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

Hine, E., & Floridi, L. (2022, forthcoming). Artificial intelligence with American values and Chinese characteristics: A comparative analysis of American and Chinese governmental AI policies. *AI & Society*. https://doi.org/10.1007/s00146-022-01499-8

Hintze, O. (1975). Military organization and the organization of the state. In F. Gilbert (Ed.), *The historical essays of Otto Hintze* (pp. 178–215). Oxford University Press. First published in 1906.

HLEG-AI: EU High-level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission.

Hofmann, J. (2018). Digitalisierung und demokratischer Wandel als Spiegelbilder? In F. Martinsen (Ed.), *Wissen – Macht – Meinung. Demokratie und Digitalisierung: Die 20. Hannah-Arendt-Tage 2017* (pp. 14–21). Velbrück Wissenschaft.

Hong, S.-H., Lee, J., Jang, S., & Hwang, H. (2022). Making regulation flexible for the governance of disruptive innovation: A comparative study of AVs regulation in the United Kingdom and South Korea. *Journal of European Public Policy*, *29*(11). https://doi.org/10.1080/13501763.2022.2096101

Horowitz, M. C. (2016, July 29). Who'll want artificially intelligent weapons? ISIS, democracies, or autocracies? *Bulletin of the Atomic Scientists*. http://thebulletin.org/who'llwant-artificially-intelligent-weapons-isis-democracies-orautocracies9692.

IEEE. (2019). *Ethically aligned design – a vision for prioritizing human well-being with autonomous and intelligent systems*. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

Inglehart, R. (1990). *Culture shift in advanced industrialized society*. Princeton University Press.

ITU: International Telecommunication Union. (2021). AI for good – accelerating the United Nations sustainable development goals. https://aiforgood.itu.int/

Jervis, R. C. (1997). *System effects: Complexity in political and social life*. Princeton University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Katznelson, I. (1996). *Liberalism's Crooked Circle*. Princeton University Press.

Kerry, C. F., Meltzer, J. P., Renda, A., Engler, A. C., & Fanni, R. (2021, October). *Strengthening International Cooperation on AI: Progress report*. Report from the Brookings Institution and Centre for European Policy Studies Forum for Cooperation on Artificial Intelligence.

Khan, S. (2022, March 1). How can AI support diversity, equity and inclusion? *World Economic Forum*.

Kinniburgh, F., Selin, H., Selin, N. E., & Schreurs, M. (2022, forthcoming). When private governance impedes multilateralism: The case of international pesticide governance. *Regulation & Governance*. https://doi.org/10.1111/rego.12463

Kiron, D., & Unruh, D. (2019). Even If AI can cure loneliness — should It? *MIT Sloan Management Review*, *60*(2).

König, P. D., & Wenzelburger, G. (2019). Why parties take up digitization in their manifestos: An empirical analysis of eight western European economies. *Journal of European Public Policy*, *26*(11), 1678–1695. https://doi.org/10.1080/13501763.2018.1544268

König, P. D., Wurster, S., & Siewert, M. B. (2022). Consumers are willing to pay a price for explainable, but not for green AI: Evidence from a choice-based conjoint analysis. *Big Data and Society*, *9*(1), 1–13. https://doi.org/10.1177/20539517211069632

Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics*, *160*(2), 377–392. https://doi.org/10.1007/s10551-019-04204-w

Lütge, C. (2019). *White paper on AI ethics and governance: Building a connected, intelligent and ethical world*. https://ieai.mcts.tum.de/wp-content/uploads/2020/04/White-Paper_AI-Ethics-and-Governance-_March-20201.pdf.

Lütge, C. (2020). *AI ethics and governance: Building a connected, intelligent and ethical world*. https://ieai.mcts.tum.de/wp-content/uploads/2020/04/White-Paper_AI-Ethics-and-Governance-_March-20201.pdf.

Lütge, C., Poszler, F., Joaquin Acosta, A., Danks, D., Gottehrer, G., Mihet-Popa, N. L., & Naseer, A. (2021). AI4People-ethical guidelines for the automotive sector: Fundamental requirements and practical recommendations. *International Journal of Technoethics*, *12*(1), 101–125. https://doi.org/10.4018/IJT.20210101.oa2

Maasen, S., & Sutter, B. (2020). Die Neuerfindung der Soziologie in einer, für eine und mit einer sich digitalisierende(n) Gesellschaft. In S. Maasen & J.-H. Passoth (Eds.), *Soziale Welt Sonderband: Vol. 23. Soziologie des Digitalen - Digitale Soziologie?* (pp. 73–90). Nomos. https://doi.org/10.5771/9783845295008

Mann, M. (1993). *The sources of social power, vol.2: The rise of classes and nation states, 1760–1914*. Cambridge University Press.

McCarthy, J. (1989). Review of the question of artificial intelligence edited by Brian Bloomfield. *Annals of the History of Computing*. http://www-formal.stanford.edu/jmc/reviews/bloomfield.pdf.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*. Retrieved March 31, 2017, from http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Meinecke, F. (1948). *Schaffender Spiegel: Studien zur Deutschen Geschichtschreibung und Geschichtsauffassung*. Stuttgart.

Meltzer, J. (2021, June). *The role of international standards in AI and the geopolitical implications*. Brookings. Unpublished manuscript (on file with the corresponding author).

Meltzer, J. P., & Tielemans, A. (2022, May). *The European Union AI Act: Next steps and issues for building international cooperation*. Brookings Institution Policy Brief.

Mertes, S., Huber, T., Weitz, K., Heimerl, A., & André, E. (2022). GANterfactual – counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in Artificial Intelligences*, *5*(4). art. 825565. https://doi.org/10.3389/frai.2022.825565

Milner, H. V. (2006). The digital divide: The role of political institutions in technology diffusion. *Comparative Political Studies*, *39*(2), 176–199. https://doi.org/10.1177/0010414005282983

Milner, H. V., & Solstad, S. U. (2021). Technological change and the international system. *World Politics*, *73*(3), 545–589. https://doi.org/10.1017/S0043887121000010

Molnar, C. (2022). *A guide for making black box models explainable*. Lulu Publishers.

Moravcsik, A. (1997). Taking preferences seriously: A liberal theory of international politics. *International Organization*, *51*(4), 513–553. https://doi.org/10.1162/002081897550447

Mosley, L. (2009). Private governance for the public good? Exploring private sector participation in global financial regulation. In H. V. Milner & A. Moravcsik (Eds.), *Power, interdependence, and non-state actors in world politics* (pp. 126–146). Princeton University Press.

Munck, G. L. (2004). Tools for qualitative research. In H. E. Brady & D. Collier (Eds.), *Rethinking social inquiry: Diverse tools* (1st ed., pp. 105–121). Rowman & Littlefield.

Nield, D. (2019, November 21). A promising solar energy breakthrough just achieved 1,000-degree heat from sunlight. *ScienceAlert*, Retrieved September 10, 2022, from https://www.sciencealert.com/ai-plus-sunlight-equals-hotter-solar-ovens-and-no-need-for-fossil-fuels.

Nitzberg, M., & Zysman, J. (2022). Algorithms, data, and platforms: The diverse challenges of governing AI. *Journal of European Public Policy*, *29*(11). https://doi.org/10.1080/13501763.2022.2096668

NSCAI: National Security Commission on Artificial Intelligence [of the United States]. (2021). *Final Report*. https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

OECD. (2019). *OECD principles on AI*. https://www.oecd.org/going-digital/ai/principles/.

OECD.AI. (2021, March 1). *Database of national AI policies*. https://oecd.ai/en/dashboards.

Owen, R., Bessant, J., & Heintz, M. (2013). *Responsible Innovation: Managing the responsible emergence of science and innovation in society*. Wiley.

Pagallo, U., Aurucci, P., Casanovas, P., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Schafer, B., & Valcke, P. (2019). *On Good AI Governance: 14 priority*

actions, a S.M.A.R.T. model of governance, and a regulatory toolbox. AI4People. AI4Peoples-Report-on-Good-AI-Governance_compressed.pdf (eismd.eu).

Pauwelyn, J., Maggetti, M., Büthe, T., & Berman, A. (2022). *Rethinking Participation in Global Governance: Voice and Influence after Stakeholder Reforms in Global Finance and Health*. Oxford University Press.

Renda, A. (2019). *Artificial intelligence: Ethics, governance and policy challenges*. Centre for European Policy Studies.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI and Society*, *36*(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2

Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer. (State-of-the-art survey, 11700).

Sartori, G. (1970). Concept misformation in comparative politics. *American Political Science Review*, *64*(4), 1033–1053.

Schleifer, P., Fiorini, M., & Fransen, L. (2019). Missing the bigger picture: A population-level analysis of transnational private governance organizations active in the global south. *Ecological Economics*, *164*(art.106362). https://doi.org/10.1016/j.ecolecon.2019.106362

Schönmann, M., Bodenschatz, A., Uhl, M., & Walkowitz, G. (2022). The care-dependent are less averse to care robots: Comparing intuitions of the affected and the non-affected. *Munich Papers in Political Economy*, *24*.

Shadbolt, N., Van Kleek, M., & Binns, R. (2016). The rise of social machines. *IEEE Consumer Electronics Magazine*, *5*(2), 106–111. https://doi.org/10.1109/MCE.2016.2516179

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*(Feb). art.102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Siegmann, C., & Anderljung, M. (2022). *The Brussels effect and artificial intelligence: How EU regulation will impact the global AI market*. Report from the Centre for the Governance of AI.

Stephan, B. (2018). Im toten Winkel. Süddeutsche Zeitung Magazin 51/2018.

TCS: TATA Consultancy Services. (2017). *Getting smarter by the day: How AI is elevating the performance of global companies, TCS global trend study: Part 1*. Retrieved September 10, 2022. https://www.tcs.com/content/dam/tcs/pdf/Industries/global-trend-studies/ai/TCS-GTS-how-AI-elevating-performance-global-companies.pdf.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Unruh, C. F., Haid, C., Fottner, J., & Büthe, T. (2022). Human autonomy in algorithmic management. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 753–762). https://doi.org/10.1145/3514094.3534168

Vincent, C. J., Niezen, G., O'Kane, A. A., & Stawarz, K. (2015). Can standards and regulations keep up with health technology? *JMIR Mhealth and Uhealth*, *3*(2), e64. https://doi.org/10.2196/mhealth.3918

Vogel, D. (2005). *The market for virtue: The potential and limits of corporate social responsibility*. Brookings Institution Press.

European Commission, Directorate-General for Communication, & von der Leyen, U. (2019). *A union that strives for more: My agenda for Europe: Political guidelines for the next European Commission 2019-2024*. Publications Office.

Von Ingersleben-Seip, N. (2022). *Competition and cooperation in artificial intelligence standard setting: Explaining emergent patterns*. Unpublished manuscript.

Von Ingersleben-Seip, N., & Georgieva, Z. (forthcoming). Old tools for the new economy? Counterfactual causation in foreclosure assessment and choice of remedies on data-driven markets. *Journal of Antitrust Enforcement*.

Walker, R. (2021, June 7). Germany warns: AI arms race already underway. *DW News*.

Weitz, K., Hassan, T., Schmid, U., & Garbas, J.-U. (2019). Deep-Learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. *Technisches Messen*, *86*(7–8), 404–412. https://doi.org/10.1515/teme-2019-0024

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 177–192.

Zysman, J., & Kenney, M. (2017). Intelligent tools and digital platforms: Implications for work and employment. *Intereconomics*, *52*(6), 329–334. https://doi.org/10.1007/s10272-017-0699-y

Zysman, J., & Newman, A. L. (2006). *How revolutionary was the digital revolution? National responses, market transitions, and global technology*. Stanford University Press.

Zysman, J., & Nitzberg, M. (2020). Governing AI: Understanding the limits, possibilities, and risks of AI in an era of intelligent tools and systems. *Wilson Center Science and Technology Innovation Program Report*.