Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence

Alexander Buhmann
Christian Fieseler
BI Norwegian Business School

Responsible innovation in artificial intelligence (AI) calls for public deliberation: well-informed "deep democratic" debate that involves actors from the public, private, and civil society sectors in joint efforts to critically address the goals and means of AI. Adopting such an approach constitutes a challenge, however, due to the opacity of AI and strong knowledge boundaries between experts and citizens. This undermines trust in AI and undercuts key conditions for deliberation. We approach this challenge as a problem of situating the knowledge of actors from the AI industry within a deliberative system. We develop a new framework of responsibilities for AI innovation as well as a deliberative governance approach for enacting these responsibilities. In elucidating this approach, we show how actors from the AI industry can most effectively engage with experts and nonexperts in different social venues to facilitate well-informed judgments on opaque AI systems and thus effectuate their democratic governance.

Key words: artificial intelligence (AI), AI ethics, AI governance, responsible innovation, political corporate social responsibility (PCSR), deliberative democracy

Paradigmatic advances in machine learning techniques have greatly expanded the capabilities of artificial intelligence (AI). These systems mimic functions typically associated with human attributes and augment them at scale via software, including the functions not only of vision and speech but also of language processing, learning, and problem solving. On the basis of these burgeoning capabilities, AIs can exercise an ever-increasing degree of autonomy in decision-making in crucial spheres, including in government (Coglianese & Lehr, 2016), health care (Norgeot, Glicksberg, & Butte, 2019), management (Kellogg, Valentine, & Christin, 2020), and policing (Kaufmann, Egbert, & Leese, 2019). Despite their many upside promises, AI systems can fail—like humans—to achieve their intended goals, either because the training data they use may be biased or because their recommendations, decisions, and actions yield unintended and negative consequences (Crawford & Calo, 2016; Martin, 2019; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). Through such failures, AIs can have wide-ranging adverse effects on public goods, such as justice, equity, and privacy, even potentially undermining the processes of fair democratic elections

Business Ethics Quarterly (2022), pp. 1–34. DOI:10.1017/beq.2021.42
Published by Cambridge University Press on behalf of the Society for Business Ethics.

(Calo, 2017; Eubanks, 2018; Tutt, 2016; Zarsky, 2016). Therefore most governments have now declared a commitment to addressing innovations in AI as global challenges to the safeguarding of public goods (Cath, Wachter, Mittelstadt, Taddeo, & Floridi, 2018).

Many of the challenges entailed in seeking to establish responsible innovation in AI are not all entirely new, as they closely resemble issues in other fields, such as bioethics (Floridi et al., 2018). The literature on responsible AI has identified and continues to discuss, however, the unique role of *epistemic challenges* ensuing from the poor "traceability" (Mittelstadt et al., 2016) and "explicability" (Floridi et al., 2018) of "opaque" (Burrell, 2016) AI systems. Broadly speaking, such epistemic challenges arise from the self-learning capacities of algorithms and the autonomy of AI systems that results from these capacities. This can make it difficult even for AI developers themselves to forecast or reconstruct how data inputs are handled within such systems, how decisions are made, and how these decisions impact domains of application in the long term (Mittelstadt et al., 2016). This in turn imposes significant limitations on the effectiveness of government regulations to protect societies and the environment from the harmful impacts of AI (Buhmann & Fieseler, 2021b; Morley, Elhalal, Garcia, Kinsey, Mökander, & Floridi, 2021). Attempts to address this problem have led to a surge in the issuance of guidelines for "ethical AI" over the past five years, authored by governments, nongovernmental organizations, and corporations (Schiff, Borenstein, Biddle, & Laas, 2021). Recent scholarship has endeavored to synthesize these guidelines within a meta-framework of principles for ethical AI (Floridi & Cowls, 2019) and to move beyond principles (or "what" questions) to the creation of translational tools (or "how" questions) for tackling ethical challenges in practice, that is, within the process of AI design (Morley, Floridi, Kinsey, & Elhalal, 2020). This discussion to date is directed mostly at AI practitioners, such as designers, engineers, and controllers, and focused on making principles applicable for the diagnosis of ethical issues in specific microcontexts. Less attention has so far been paid to linking such principles and translational tools with questions of societal and corporate governance (Morley et al., 2021). Although most principles and translational tools currently being developed envisage active and collaborative involvement on the part of the AI industry, and specifically those organizations that develop and employ semi-autonomous systems, with actors from the public, private, and civil society sectors as a means of overcoming the limitations of government regulations (Buhmann & Fieseler, 2021b; Buhmann, Paßmann, & Fieseler, 2020; Morley et al., 2020; Rahwan, 2018; Veale & Binns, 2017), the matter of which specific actors to involve in solutions and how precisely to involve these actors is rarely elaborated in detail. This raises the question, what should be the role of actors from within the AI industry in contributing to the governance of responsible AI innovation, specifically in addressing both the need for the *collaborative involvement* of the AI industry and the need to tackle the *epistemic challenges* pertaining to the governance of AI? This question highlights several further open questions in the ethical AI and AI governance literatures.

The first of these outstanding questions is how and under which conditions societal and corporate governance structures can gainfully interact with translational tools for ethical AI (and with the principles on which these tools are based). As Morley et al. (2021: 241) observed, "there is, as of yet, little evidence that the use of any of these translational tools/methods has an impact on the governability of algorithmic systems." This unresolved question highlights the fact that the governability of systems is ultimately a matter to be decided in the context of concrete models of governance. In turn, this implies that any further discussion of tools for ethical AI needs to address their application at the levels of both system design *and* governance, that is, clarify not only their "technical implementation" along the AI development pipeline but also their "administrative implementation" within mechanisms of societal and organizational decision-making.

A second question is which specific form of governance would best help actors in the AI industry to identify and implement legitimate approaches to responsible innovation while at the same time allowing for and fostering technically and economically efficient processes of AI innovation. Addressing this question thus calls for the development of steering mechanisms that would allow the AI industry to innovate while also taking societal needs and fears into due consideration. For example, such consideration would involve balancing conflicting pressures between harnessing the potential for accuracy of AI systems (including their power to do good) against the need for these systems to be accountable (Goodman & Flaxman, 2017).

A third issue to be addressed is that implementing ethical AI entails a realistic appraisal of the prospects for and challenges involved in bringing about the active and collaborative engagement of the AI industry in the process of responsible innovation. This in turn calls for a problematization of the power imbalances between different stakeholders, including the epistemic challenges and knowledge inequalities between AI experts and the general public, further calling into question the arguably ambivalent role of the AI industry in gatekeeping such endeavors.

And as a fourth and final question related to all the preceding questions, what are the most appropriate political visions and values for the governance of responsible AI innovation? This question highlights the need for feeding macro-ethical considerations into current micro-ethical discussions that focus on the design specifications of algorithms and the AI development process. Although this macro–micro connection is currently explored in relation to data ethics more broadly (Taddeo & Floridi, 2016; Tsamados et al., 2021), it has rarely been examined with a business ethics focus in mind (Häußermann & Lütge, 2021).

As Whittaker et al. (2018: 4) succinctly concluded in their *AI Now Report* of 2018, "the AI industry urgently needs new approaches to governance." Responding to this call and the questions outlined earlier, we will develop our argument as follows. First, building on the work of Mittelstadt et al. (2016) and Tsamados et al. (2021), we identify three types of ethical concerns specific to AI innovation, that is, evidence, outcome, and epistemic concerns. We then interrelate these concerns with a recent normative concept of responsible innovation (Voegtlin & Scherer, 2017) to propose a new framework of responsibilities for innovation in AI. In developing this

framework, we foreground the importance of facilitating governance that addresses epistemic concerns as a meta-responsibility. Second, we discuss the rationale and possibilities for involving the AI industry in broader collective efforts to enact such governance. Here we argue from the perspective of political corporate social responsibility (PCSR) (Scherer & Palazzo, 2007, 2011), focusing on the potential of deliberation for addressing questions of legitimation, contributions to collective goals, and organizational learning, and we outline the challenges in applying this perspective to responsible AI innovation. Subsequently, we set forth the prospects of a "distributed deliberation" approach as a means of overcoming these challenges. We elaborate this approach by proposing a model of distributed deliberation for responsible innovation in AI, identifying different venues of deliberation and specifying the role and responsibilities of the AI industry in these different fora. Finally, we discuss prospects and challenges of the proposed approach and model and highlight avenues for further research on deliberation and the governance of responsible innovation in AI.

TOWARD A FRAMEWORK OF RESPONSIBILITIES FOR THE INNOVATION OF ARTIFICIAL INTELLIGENCE

Three Sets of Challenges for Responsible Innovation in AI

Broadly speaking, responsible innovation refers to the exercise of collective care for the future by way of stewardship of innovation in the present (Owen, Bessant, & Heintz, 2013: 36). Such stewardship calls for informed anticipation of key challenges and concerns regarding the purposes, processes, and outcomes of innovation (Barben, Fisher, Selin, & Guston, 2008; Stilgoe, Owen, & Macnaghten, 2013). From the ongoing debate on the ethics of AI and algorithms (Mittelstadt et al., 2016; Tsamados et al., 2021), three sets of challenges can be summarized (see similarly Buhmann et al., 2020).

Evidence concerns relate to the mechanisms by which self-learning systems transform massive quantities of data into "insights" that inform an AI system's decisions, recommendations, and actions. Such concerns arise because AIs are designed to reach conclusions on the basis of *probabilities* rather than conclusive evidence of certain outcomes. These probabilities are derived from seemingly meaningful patterns detected within vast collections of data, often involving inferences of causality based on mere correlations within such data. The decisions reached by AIs may be based on misguided evidence, moreover, as when algorithmic conclusions rely on incomplete and incorrect data or when decisions are based on unethical or otherwise inadequate inputs (Mittelstadt et al., 2016; Tsamados et al., 2021; Veale & Binns, 2017). In short, flawed AI decisions can arise both from poor-quality data and also (intended or unintended) properties of data sets, models, or entire systems.

Outcome concerns relate to the potentially adverse consequences of decisions reached by AI systems, including both directly and indirectly harmful outcomes. Directly harmful outcomes may take the form of discrimination against certain entities or groups of people, as, for example, when data-driven decision-support systems serve to perpetuate existing injustices related to ethnicity or gender, either

because these systems are biased in their design or because human biases are picked up in the training data used for algorithms (Tufekci, 2015). Poorly designed AI may further generate feedback loops that reinforce inequalities, as in the case of predictive policing (Kaufmann, Egbert, & Leese, 2019), for example, or in predictions of creditworthiness that render it difficult for individuals to escape vicious cycles of poverty (O'Neill, 2014). Indirectly harmful outcomes of AIs can arise from the application of AI technologies more generally, often with long-term consequences, such as large-scale technological unemployment (Korinek & Stiglitz, 2018). Such outcomes can also take the form of so-called latent, secondary, and transformative effects (Mittelstadt et al., 2016) that occur when AI outcomes change the ways that people perceive situations, as, for example, in the case of profiling algorithms that powerfully ontologize the world in particular ways and trigger new patterns of behavior (Pasquale, 2015), though these effects are also evident in the ways that content curation and news recommendation algorithms lead to people being unwittingly socialized in "filter bubbles" (Berman & Katona, 2020).

Epistemic concerns relate to issues stemming from the "opacity" of AI (Burrell, 2016), including both the inscrutability of algorithmic inputs and their processing and the poor traceability of potentially latent and long-term harmful consequences of AIs.[1] These concerns arise when AIs are not readily open to explication and scrutiny and when the outcomes of their application are not relatable in any straightforward way to the vast sets of data on which AIs draw to reach their conclusions (Miller & Record, 2013). Harmful outcomes may be difficult to trace to a particular system's operations, moreover, because of the fluid and diffuse, that is, networked, ways in which such systems evolve (Sandvig, Hamilton, Karahalios, & Langbort, 2014). As software artifacts applied in data processing, AIs give rise to ethical issues that are incorporated into their very design as well as the data used to test and train models (Mittelstadt et al., 2016). Epistemic concerns thus relate to all technical and sociotechnical factors that render it difficult to detect the potential harm caused by algorithms and to identify the causes and responsibilities for such harm. Indeed, epistemic concerns are arguably what truly set AI innovation apart from other ethically complex fields, such as biotechnology, and pronounces it as a "grand challenge" (Buhmann & Fieseler, 2021a), especially in the case of AIs that are "truly opaque" (Floridi et al., 2018), as we argue next.

Types of Epistemic Concerns about AI

Epistemic concerns can be further differentiated in relation to three broad categories of strategic, expert, and true opacity. *Strategic opacity* refers to inscrutability and poor traceability resulting from deliberate intent on the part of the designers of a

[1] Although specific types of inscrutability and traceability are sometimes discussed separately in review articles, primarily for the purpose of highlighting different *causes* of AI opacity (Mittelstadt et al., 2016; Tsamados et al., 2021), in this review, we have grouped such epistemic concerns to highlight their role as a distinct cluster of "meta-concerns" related to AI. See also the further elaboration of this cluster of concerns in the following section and their discussion as a distinct dimension within the framework of responsibilities for AI innovation that is developed in the section thereafter.

certain AI. In this case, algorithms that might otherwise be interpretable and whose effects might be traceable are intentionally kept secret, obfuscated, or "black-boxed." Typical motives for strategic opacity include relatively noncontroversial aims like optimizing the functionality of an AI, ensuring its competitiveness, or protecting the privacy of user data (Ananny & Crawford, 2018; Glenn & Monteith, 2014; Leese, 2014; Stark & Fins, 2013) but also the motives of avoiding accountability and evading regulations (Ananny & Crawford, 2018; Martin, 2019).

Epistemic concerns regarding *expert opacity* relate to the issue of "popular comprehensibility." Whereas the design, development, and outcomes of an AI may be explicable and interpretable among experts, these aspects of AI remain widely inscrutable, uninterpretable, and untraceable for laypeople. Expert opacity can thus be described in broad terms as arising at the intersection of system complexity and "technical literacy" (Burrell, 2016). Common themes identified in the literature on expert opacity include so-called epistemic vices, such as AI "gullibility," "dogmatism," and "automation bias" (Tsamados et al., 2021). Expert opacity can also arise inadvertently through attempts at disclosure and transparency that overwhelm citizens on account of the sheer volume and complexity of information made available to them (Ananny & Crawford, 2018), though here it should be noted that any *intentional* obfuscation by such disclosure "overload" would rather constitute an element of strategic opacity (Aïvodji, Arai, Fortineau, Gambs, Hara, & Tapp, 2019).

The third group of epistemic concerns relates to AI processes and outcomes that are difficult to scrutinize and trace not only for laypeople but also for AI experts and developers themselves. We refer to this as *true opacity*, which arises from the ways in which AIs are developed and evolve as emergent phenomena in practice, since AIs and algorithms do not simply comprise mathematical entities but further constitute "technology in action." Together with the fact that AI developers often reuse and repurpose code from libraries, thereby leading to the wide dispersion and therefore obfuscation of responsibilities for particular code and outcomes, the perpetually evolving aspect of AI leads even software designers to "regularly treat part of their work as black boxes" (Mittelstadt et al., 2016: 15). Such *true opacity* is especially problematic in that it is not merely a matter of insufficient popular comprehension and technical literacy that could potentially be addressed directly through explanation and training. In the face of true opacity, AIs can only be understood by way of an iterative process and not merely through studying an AI system's properties and mathematical ontology (Burrell, 2016).

True opacity can relate to evidence, outcomes, or both. At the level of evidence, for example, such opacity can take the form of uncertainty in identifying potentially problematic and sensitive variables used by AIs (Veale & Binns, 2017). At the level of outcomes, meanwhile, examples of true opacity include uncertainty about the latent impacts of AIs (Sandvig et al., 2014) and the appropriateness of extant social evaluation of these impacts (Baum, 2020; Buhmann et al., 2020). In relation to evidence and outcomes in combination, true opacity can take the form of uncertainty about the allocation of responsibilities across vast and poorly transparent networks of human, software, and hybrid agents (Floridi, 2016) or uncertainty about the norms

incorporated within automated systems that are thereby excluded from the sphere of social reflexivity (D'Agostino & Durante, 2018). As such, the term *true opacity* is not an ontologization but rather denotes phenomena that AI experts themselves refer to as "opaque."

A Framework of Responsibilities for AI Innovation

For all the numerous guidelines that have been published on "ethical AI" by governments, private corporations, and nongovernmental organizations, especially over the past five years (Schiff et al., 2021), the lack of consensus still surrounding key areas threatens to delay the development of a clear model of governance to ensure the responsible design, development, and deployment of AI (Jobin, Ienca, & Vayena, 2019). More promisingly, however, some recent research has started to offer meta-analyses, with growing agreement apparently emerging around a five-dimensional framework of principles for ethical AI. This framework considers *beneficence* (AI that benefits and respects people and the environment), *nonmaleficence* (AI that is cautious, robust, and secure), *autonomy* (AI that conserves and furthers human values), *justice* (AI that is fair), and *explicability* (AI that is explainable, comprehensible, and accountable) (Floridi & Cowls, 2019). Nevertheless, these efforts to attain one common framework still include some inconsistencies. It remains unclear, for instance, why certain aspects of justice (such as "*avoiding* unfairness") or of autonomy (such as "*protecting* people's power to decide") are not simply subsumed within the dimension of *nonmalfeasance* and why other aspects of justice (such as "*promoting* diversity and inclusion") or autonomy (such as "*furthering* human autonomy") are not positioned as elements of *beneficence.* Furthermore, and more importantly, "explicability" appears in this framework both as a stand-alone dimension *and* as a necessary element of all other dimensions, because such explicability is necessary to *enable* AI beneficence, justice, and so on. In its current version, moreover, the framework appears to replicate a central omission in AI ethics guidelines concerning the role of governance: as shown by a recent study of twenty-two guidelines (Hagendorff, 2020), questions of governance are rarely addressed in codes and principles for ethical AI. The framework developed by Floridi and Cowls (2019), founded on a meta-review of such guidelines, likewise falls short of interrelating principles for ethical AI with principles for governance. Governance, however, is key for responsible processes of innovation (Jordan, 2008).

To address these issues, we suggest working toward a framework of responsibilities that interrelates the three sets of challenges reviewed earlier with a normative concept of responsible innovation (Voegtlin & Scherer, 2017), which involves three basic types of responsibilities: 1) responsibilities to do no harm, 2) responsibilities to do good, and 3) responsibilities for governance that enables the first two dimensions. This three-dimensional setup has recently been applied to principles in ethical AI (Buhmann & Fieseler, 2021b), closely corresponds to the basic distinction between AI risks versus opportunities used earlier by Floridi et al. (2018), and, more importantly, adds the key dimension of governance. Thus we propose a matrix consisting of three responsibility dimensions that are further operationalized by three constitutive responsibilities that each address one of the challenges reviewed earlier.

(See Table 1 for an overview with examples from the current AI ethics literature for illustration.)

The dimension of *AI nonmalfeasance* (avoiding harm from AI) refers to responsibilities for managing risks and controlling for potentially harmful consequences. These include the evidence responsibility for avoiding harm by using the right data and using data in the right way so as to ensure robustness of evidence and the protection of security, safety, and integrity in algorithmic processing; the outcome responsibility for avoiding harm by protecting human autonomy and avoiding discriminatory effects like biases; and the epistemic responsibility for avoiding harm by identifying any inconclusiveness and fallibilities of AI systems and creating awareness and knowledge regarding any negative impacts of AI.

The dimension of *AI beneficence* (doing good with AI) refers to responsibilities for the improvement of living conditions in accordance with agreed principles or aims, such as the United Nations Sustainable Development Goals (SDGs). These include the evidence responsibility for doing good by assessing data and their algorithmic processing according to their potential to promote fairness, justice, and well-being for people and the environment; the outcome responsibility for doing good by furthering justice through AI and applying AI for achieving agreed aims, such as the SDGs, and tackling "grand challenges"; and the epistemic responsibility for doing good through building knowledge and trust to maximize the social utility potential of AI and prevent the "underuse" of AI systems owing to fear or ignorance.

The dimension of *responsible AI governance* refers to responsibilities for the development and support of institutions, structures, and mechanisms aimed at facilitating responsible innovation in AI. Specifically, this entails enabling and enacting governance of the evidence responsibility for preventing the use of potentially inconclusive and misguided evidence in algorithmic processing, governance of the outcome responsibility for monitoring the direct and indirect effects of AI, and governance of the epistemic responsibility for scrutinizing algorithmic processes and enabling traceability of AI.

Responsible AI governance must be addressed at two levels in parallel: at the technological level of *AI design* and at the level of *translational tools* that are supposed to operationalize responsible AI design by enhancing the evaluation, understanding, and legitimation of AI. In other words, translational tools (including their development and implementation) need to be explained and justified *together* with the technology they are supposed to help govern (Morley et al., 2020). Furthermore, the governance of evidence, outcome, and epistemic responsibilities merits particular attention in that it constitutes the key dimension of responsible AI innovation. This is because, as a governance responsibility, it operates at a meta-level, meaning it facilitates responsible innovation on the other two dimensions (Voegtlin & Scherer, 2017). Within responsible AI governance, the *governance of epistemic responsibilities* plays a pivotal role, for two main reasons. First, among the three meta-responsibilities, it operates itself on a meta-level, as the detection and governance of potential harm as well as opportunities to do good on the levels of evidence and outcomes *rely* on scrutable and traceable systems. In other words, epistemic challenges like poor scrutability and traceability may significantly hinder

Table 1: A Framework of Responsibilities for the Innovation of Artificial Intelligence

Operationalization of responsibilities[a]	Dimensions of responsible innovation in artificial intelligence[b]		
	AI nonmalfeasance: "avoiding harm from AI"	**AI beneficence: "doing good with AI"**	**Responsible AI governance**
Evidence responsibility	**Avoiding the use of inconclusive or misguided evidence** Using the right data and using data right; protecting the ability of AI to make good decisions (Floridi & Cowls, 2019); fostering the robustness, security, and safety of evidence (Organisation for Economic Co-operation and Development, 2021)	**Using conclusive evidence based on adequate inputs** Choosing data and algorithmic processing based on their potential to promote fairness, justice, and benefits to people and the environment (Veale & Binns, 2017)	**Enabling and enacting governance directed at evidence responsibilities to avoid harm and do good**
Outcome responsibility	**Avoiding direct and indirect harmful impact** Respecting human autonomy and preventing harm (European Commission, 2019); mitigating negative impacts; minimizing discrimination and bias (Beijing Academy of Artificial Intelligence, 2019)	**Using AI to achieve beneficial direct and indirect impact** AI that promotes justice and fairness (European Commission, 2019); fosters inclusive growth, sustainable development, and well-being (Organisation for Economic Co-operation and Development, 2019); promotes diversity, and serves humanity by furthering human values, including freedom and autonomy (Beijing Academy of Artificial Intelligence, 2019)	**Enabling and enacting governance directed at outcome responsibilities to avoid harm and do good**
Epistemic responsibility	**Identifying harmful design, development, and deployment of AI** Labeling and recognizing points of inconclusiveness and the fallibility of systems (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016); raising awareness of negative impacts (Beijing Academy of Artificial Intelligence, 2019); AI monitoring and algorithm auditing (Rahwan, 2018); clarifying tensions between the risks of being wrong and epistemic responsibilities (Miller & Record, 2013)	**Identifying beneficial design, development, and deployment of AI** Building knowledge and trust to prevent the "underuse" of AI below its full potential due to fear or ignorance (Floridi et al., 2018); balancing the pressure for interpretability and control with the need to harness the potential for creativity and accuracy of systems (Ecoffet et al., 2020; Goodman & Flaxman, 2017)	**Enabling and enacting governance directed at epistemic responsibilities to avoid harm and do good**

[a] Based on three sets of challenges for responsible innovation in AI. [b]Adapted from Voegtlin and Scherer (2017) and Floridi and Cowls (2019).

effective apprehension of the purposes, processes, and outcomes of AI. Second, the high demand for transparency that *results* from prevalent epistemic concerns can potentially divert resources away from important advances in AI performance and accuracy (Ananny & Crawford, 2018), which means the governance of epistemic responsibility needs to support business and society in seeking legitimate solutions to prevalent tensions in AI. Addressing these tensions includes balancing the pressure for interpretability, accountability, and control of AI systems with the need to avoid hindering the potential of AI systems for greater creativity and accuracy (Ecoffet, Clune, & Lehman, 2020; Goodman & Flaxman, 2017).

ENACTING RESPONSIBLE AI GOVERNANCE: A POLITICAL CORPORATE SOCIAL RESPONSIBILITY APPROACH

The Prospects of Political Corporate Social Responsibility

The framework of responsibilities for the innovation of artificial intelligence described in the preceding sections accentuates responsible AI governance as a meta-responsibility. Within this dimension, we have pointed to the particular importance of enacting governance directed at epistemic responsibilities. As we argue subsequently, both these emphases in the responsible innovation of AI point toward the prospects of deliberation for governing AI innovation.

In following common frameworks on responsible innovation (Owen, Bessant, & Heintz, 2013; Stilgoe et al., 2013), we argue that responsible AI governance needs to be enacted through a deliberative control process. This entails open and well-informed "deep democratic" debate (Michelman, 1997) aimed at generating broadly agreed-upon opinions and decisions (Chambers, 2003). Such deliberation for responsible innovation necessitates "structures at various levels (e.g., global, societal, corporate) that facilitate an inclusive process of collective will formation on the goals and means and the societal acceptability of innovation" (Scherer & Voegtlin, 2020: 184). Recent scholarship on responsible innovation within the management and business ethics literatures has discussed the capacity of different corporate governance models for responsible innovation and explored the prospects of approaches that address nonstate entities like corporations as political actors (Brand & Blok, 2019; Scherer & Voegtlin, 2020). Rather than focusing corporate responsibilities on shareholders or stakeholders, this scholarship has developed a program of PCSR that tasks nonstate actors with an active role in the collaborative endeavor of producing and protecting public goods (Scherer & Palazzo, 2007, 2011).[2] For this, PCSR builds on ideals of deliberative democracy (Habermas, 1998; Thompson, 2008), foregrounding the collaborative engagement of state and nonstate actors in collective decisions through a rational process of principled

[2] The literature on PCSR has proliferated to become an extensive field of research (cf. Rajwani & Liedong, 2015; Rasche, 2015; Scherer, 2018; Scherer & Palazzo, 2011). However, here we mostly follow the key contributions made by Scherer and Palazzo (2007, 2011) and related work on responsible innovation (Voegtlin & Scherer, 2017) that places a normative emphasis on social responsibility and the proactive engagement of nongovernmental actors.

communication that "draws in" the diverse knowledges and perspectives of all those potentially affected by such decisions.

From a PCSR perspective, achieving responsible innovation is understood as a challenge embedded in complex and globalized business environments that requires the involvement of nonstate actors as active participants in public governance to support deliberation aimed at alleviating institutional deficits (Voegtlin & Scherer, 2017). This perspective has strong similarities with the debate on responsible innovation, especially in the fundamental importance it places on deliberative democracy (Brand & Blok, 2019; Scherer, 2018; Scherer & Voegtlin, 2020). We see three main ways in which the PCSR approach is particularly well suited to tackle the challenges involved in achieving responsible innovation in AI. First, widespread outcome concerns about the potential negative impacts of AI, together with epistemic concerns related to this technology, constitute a relevant context and basis for considering organizations in the AI industry as public actors with a responsibility for social well-being and the collective good. PCSR's focus on innovation as a "political activity" and its positioning of nongovernmental actors as subject to democratic governance resonate directly with calls for politicizing the debate on responsible AI innovation (Green & Viljoen, 2020; Helbing et al., 2019; Wong, 2020; Yun, Lee, Ahn, Park, & Yigitcanlar, 2016). These calls highlight the need for a clear connection to be drawn between discussions about AI governance and questions related to the public good, including the duties and contributions of the AI industry to the public good (Hartley, Pearce, & Taylor, 2017; Wong, 2020).

Second, we believe that by highlighting and addressing the limitations of merely formal compliance with legal regulations and social expectations (Scherer & Palazzo, 2007, 2011), the PCSR approach takes into account the challenges that arise from the opacity of AI. This opacity means that corporate AI developers cannot rely merely on extant laws and regulations for legitimation and accountability but also need to consider communicative and discursive strategies. In particular, true opacity as well as expert opacity constitute a permanent concern for the AI industry in terms of the industry's legitimacy and reputation, especially as the industry may struggle to give immediate explanations and provide satisfactory accounts when critical stakeholders demand information and transparency (Buhmann et al., 2020). In conditions of unclear (external) demands related to opaque information systems, the kind of "discursive engagement" advocated in the PCSR approach for facilitating legitimate outcomes (Scherer, Palazzo, & Seidl, 2013) is highly relevant and appropriate (Mingers & Walsham, 2010). This is because important knowledge about the workings of AI systems and their wide-ranging ramifications does not reside exclusively with AI industry actors but must emerge from open deliberation with other actors that use and are affected by these systems (Lubit, 2001).

Third, by emphasizing the role of organizational learning (Scherer & Palazzo, 2010), the PCSR approach takes account of the dynamic nature of AI and the related potential for corporate routines, goals, and governance structures to be revised and shifted over time, either to achieve competitive (first-mover) advantages in AI (Horowitz, 2018) or as a means of proactively managing compliance, accountability, and reputation in the AI industry (Buhmann et al., 2020). Such concerns about

organizational learning may serve to push AI industry actors toward discursive approaches and compel them to enter into proactive deliberative debates. In practice, however, the impossibility of attaining complete "AI transparency" can be used as an excuse for organizations not to fulfill ethical duties to deliver conventional explanations and straightforward accounts based on fixed legal frameworks. In this regard, PCSR highlights not only the necessity of managing reputation and facilitating learning but also the ethical obligation of organizations to enable and participate actively in joint deliberation with other actors from government and civil society to mitigate the impediments to responsible AI innovation that arise based on expert and true opacity. As a governance approach, PCSR is thus highly compatible with current work on translational tools for ethical design that aim to compensate for the limits of hard regulation by proposing mechanisms for effectively opening up AI design and development to social scrutiny (Morley et al., 2020; Rahwan, 2018; Veale & Binns, 2017). In the following section, we discuss key challenges related to deliberation and PCSR in opening AI design and development to social scrutiny, and based on this discussion, we argue for "distributed deliberation" as an approach to help offset these challenges.

The Challenges Related to Deliberation in Governing Responsible AI Innovation

Specific limits to deliberation involving AI industry actors can be demonstrated based on the following operational principles of deliberation—see especially Nanz and Steffek (2005) and Steenbergen, Bächtiger, Sporndli, and Steiner (2003) or, for a discussion and application of these principles in the AI ethics literature, Buhmann et al. (2020). The first principle relates to *participation* and the imperative that subjects who potentially suffer negative effects should have equal access to communicative fora that aim to spotlight potential issues and facilitate argumentation with the goal of reaching broadly acceptable decisions. The second relates to *comprehension* and the principle that participants should have access to all necessary information about the issues at stake as well as proposed solutions, including the ramifications of such solutions. The principle of *multivocality*, meanwhile, means that participants need to have a chance to voice their concerns and exchange arguments freely, including the opportunity to revise their positions based on stronger and more informed arguments.

In terms of widening participation to achieve responsible innovation in AI, the challenge here lies not only in access itself but also in ensuring sufficient permanency of access, especially where systems evolve in dynamic and fluid ways (Sandvig et al., 2014; Buhmann & Fieseler, 2021a; Buhmann et al., 2020). Indeed, one of the key challenges evident in the literature on principled AI and translational tools is how to go beyond currently prevalent "one-off" approaches to ethical AI, because these approaches lack sufficient continuity of validation, verification, and evaluation of systems (Morley et al., 2021). While some scholars have suggested cooperative and procedural audits of algorithms to address this issue (Mittelstadt et al., 2016; Sandvig et al., 2014), the focus of such scholarship has so far been mostly on expert settings. Although such approaches would enable developers, engineers, and other "industry insiders" to diagnose ethical issues, these solutions

lack mechanisms to ensure the inclusion of external actors and stakeholders to "plug in" social views and evaluations from outside of the industry. Moreover, in those studies that do explicitly envisage external evaluation (e.g., Rahwan, 2018; Veale & Binns, 2017), there is a tendency to treat laypeople—or "the public"—as a monolithic entity, without addressing ways to augment public and private engagement. No consideration is given, for example, of the possibility of establishing fora or venues for involving actors in deliberation based on different kinds of knowledge and expertise. Without such ties, translational tools will remain limited to a decontextualized technical exercise that potentially distorts or neglects social injustices (Wong, 2020). This issue of participation exists as much *within* such approaches and translational tools as it does *for* them. For while practical applications proposed for ethical AI do incorporate and promote standards for assessing algorithmic practices, there is rarely any discussion of ways to subject these tools to evaluation themselves (Fazelpour & Lipton, 2020). In the absence of any such meta-evaluation, the choice of translational tools is left to developers, increasing the likelihood of convenient rather than ethical solutions, that is, approaches that favor the functionality and accuracy of for-profit systems over pro-ethical systems that bolster explicability and control in support of societal needs (Morley et al., 2021).

Lack of social evaluation is not an issue merely of participation but also of comprehension. The ideal scenario whereby informed citizens' judgments should have a critical bearing on AI design and regulation (Kemper & Kolkman, 2019) seems to be undermined not only in the most fundamental sense by the way in which AIs are developed and evolve as emergent phenomena in practice (true opacity) but also by steep knowledge inequalities between AI industry actors, policy makers, and citizens (expert opacity). While issues of expert opacity can be addressed at least in part through efforts aimed at replacing black-box models with interpretable ones (Rudin, 2019), these solutions do not address important tensions related to expert and civic engagement in deliberation. This is because such efforts do not take place in a social vacuum but in specific cultural and organizational settings (Felzmann, Villaronga, Lutz, & Tamò-Larrieux, 2019; Kemper & Kolkman, 2019; Miller, 2019), meaning they are performative and may have unintended consequences and downsides (Albu & Flyverbom, 2019). This is the case, for example, in attempts at AI explicability that actually serve to obfuscate further through disclosure (Aïvodji et al., 2019; Ananny & Crawford, 2018).

Limits to comprehension are apparent not only at the level of laypeople, moreover, because even AI experts and industry insiders themselves necessarily lack insights into latent interests, newly arising issues, and tensions between public goals. Such insights are vital for understanding and managing sensitive variables in the processing of algorithmic evidence and the latent impacts of AI. This limitation impedes the ability of experts to reflect and reconfigure their approaches (Dryzek & Pickering, 2017). The only way in which this can be compensated for is arguably through the development of a diverse knowledge base through citizen participation (Meadowcroft & Steurer, 2018).

Finally, the limitations of deliberation involving the AI industry are also evident in the often limited means available to citizens for formulating and deliberating their

concerns. For example, although public code repositories may foster open access (participation) and provide extensive information on codes and interpretable models (comprehension), actual engagement via such platforms remains hierarchical and dominated by experts (Buhmann et al., 2020). This makes it especially difficult to render accountable all the "unknowns" of algorithmic actions (Paßmann & Boersma, 2017), because exploring different forms of opacity requires inclusive observation and debate. Arguably, such lack of multivocality can be ameliorated to some extent by journalistic media, as in the case of "data journalism," for example (Diakopoulos, 2019). Such media-backed public scrutiny of AI is only possible, however, in instances of sufficient magnitude to attract the attention of "watchdog" journalism (on "criminal justice algorithms," see the discussed examples in Buhmann et al., 2020).

Although scholars have highlighted the challenges of ethical AI in relation to technical opacity and obfuscation arising from efforts to provide explanations, as well as the need for "citizen insight" (Tsamados et al., 2021), studies rarely address the question of what exactly should be disclosed to whom, which actors should be engaged and how, and where the boundaries to particular discussions and information should be drawn to enable "bigger picture" governance of AI. User participation and comprehension have so far been discussed largely as a "micro issue" in the form of technical tools or procedures with which to test and audit systems. Such approaches thus fall short of envisaging ways to increase comprehension *across* different expert and citizen fora and venues to enable the kind of broader deliberative process needed to facilitate a *socially situated* traceability and explicability of AI systems.

A MODEL OF DISTRIBUTED DELIBERATION FOR GOVERNING RESPONSIBLE AI INNOVATION

A Systems Perspective: The Prospects of Distributed Deliberation

Discussions on knowledge inequalities in deliberation (Moore, 2016) suggest that decisions and actions that result from processes in which there are strong boundaries between experts and nonexperts undermine trust in deliberative processes of governing AI for several important reasons, including 1) the inability of citizens to comprehend the *content* matters being discussed, 2) their inability to trace and evaluate the *internal process* through which AI experts reach decisions and recommendations, and 3) *epistemic deficits* that arise because diverse assessments and evaluations are not sufficiently "fed into" deliberation. The preceding considerations regarding participation, comprehension, and multivocality indicate a need to address the tensions around knowledge inequalities between the (expertise of the) AI industry and other actors—see similarly also the discussions in Stirling (2008) on the governance of technology and in Dryzek and Pickering (2017) on environmental governance, as well as the subsequent uptake of these discussions in the corporate governance literature in Scherer and Voegtlin (2020). Such knowledge inequalities are inherent in any process of analyzing, regulating, and managing complex technological and societal problems (Mansbridge et al., 2012).

Viewed from a systems perspective, knowledge inequalities are distributed across various venues, including AI expert committees, civil society organizations, public fora, and individual contemplations and reflections about AI and its governance. Examples of such venues include initiatives like the Ethics and Governance of Artificial Intelligence Initiative launched by MIT's Media Lab and Harvard University's Berkman Klein Center, professional association initiatives like the Institute of Electrical and Electronics Engineers' Global Initiative on Ethics of Autonomous and Intelligent Systems, open source activist initiatives like the Open Ethics Initiative, and corporate forays like Google's recent efforts to help its customers better understand and interpret the predictions of its machine learning models. Each of these venues could in principle assume different functions in "interacting" (Thompson, 2008: 515) single deliberative moments in support of wider public judgment. Here the idea is that although none of these venues by themselves can fully enact the deliberative virtues of participation, comprehension, and multivocality, they can still support public reasoning at large by fulfilling a distinct function in a wider network of deliberation (Parkinson, 2006; Thompson, 2008: 515). This view recognizes the inevitability of a division of labor in the deliberative process as a result of the differing types of expertise among actors in the AI industry and those outside the industry. Instead of a focus on "true" single-actor venues for deliberation, this approach emphasizes that different parts of a system can be *complementary* in supporting "deliberative rationality" for the governance of responsible AI innovation. What is most important about the judgments or outputs of deliberative venues is not so much whether they are conducive to a truly rational process "within" but whether that venue's particular discourse leads to a useful output that can be further "processed" by other venues.

On the basis of the important arguments advanced by Alfred Moore (2016) on deliberative democracy and epistemic inequalities, we hold that citizens outside of a particular expert venue for deliberating AI and its governance *need* to be able to exercise judgment on the closed deliberations of AI developers and other experts on the inside. However, as Moore (2016) discussed, meeting this need presents a complex challenge in that outsiders can be expected neither to possess the knowledge required to trace and follow the subjects deliberated upon in closed AI expert discourse nor to be able to corroborate whether this discourse follows a process of fair and principled deliberation. Furthermore, the sharpness of the boundary between expert and nonexpert venues points to an important difference in reasoning within these venues: whereas experts reason among themselves to deliver evaluations of the design, development, and impacts of particular systems or to decide on proposals for translational tools and policies to govern AI (as outputs of their deliberation), nonexpert "outsiders" of these venues need to form well-informed opinions on whether to accept (trust) or reject (resist/contest) these outputs. Moreover, the reasons that outsiders might have for accepting experts' evaluations and decisions may be quite different from what first led these experts to develop and support these outputs (Moore, 2016).

If the internal reasoning of AI expert discourse is detached in this way from the concerns and reasoning of citizens, expert venues would consequently be both

secluded from public oversight and scrutiny (a *legitimation issue*) and cut off from public feedback as an important source of creative fact finding and articulation of new public issues emerging from AI in practice (an *epistemic sourcing issue*). For the division of labor to function effectively in support of wider public reasoning, there needs to be both a real possibility of *contesting and withdrawing* legitimation and a real opportunity to *influence* the content of expert deliberation (Moore, 2016). In the following section, we follow these important reflections by Moore to discuss further how this division of labor might be gainfully distributed among different deliberative venues and across the whole AI innovation pipeline.

Modeling Distributed Deliberation across the AI Innovation Pipeline

The AI innovation pipeline consists of an iterative process that ranges from *design* (e.g., business and use-case development, where problems are defined and uses of AI are proposed), to *development* (e.g., data procurement, programming, and turning business cases into concrete design requirements applied to training data sets) and *deployment* (where AI "goes live," it is used, and its performance is monitored in practice) (Saltz & Dewar, 2019). This pipeline is a useful reference point in that various ethical challenges and related translational tools for ethical AI can be better understood by plotting them across this pipeline and addressing their implications in relation to these different stages (Morley et al., 2020). Similarly, the role of different deliberative venues and their potential for mitigating the epistemic challenges entailed in responsible AI innovation can best be elucidated by discussing them vis-à-vis this process (see Figure 1).

The Role of AI Expert Discourse

From a systems perspective, the core function of AI expert discourse is to deliver assessments and evaluations of AI systems to the wider public and to suggest new translational tools and policies for AI governance. Possible venues for such discourse include AI committees, commissions, and councils. Insofar as participation in such expert venues is based on competence, however, these venues fall short of the virtues of inclusive and open participation. The mode of engagement here is "technical expertise" (Fung, 2006), and actors from the AI industry should have an active and structured role in such expert deliberations. Selective access to these venues, combined with this proactive role of the AI industry, inevitably limits the focus of deliberation in AI expert discourse. Unlike broader public deliberation, these venues lack the ability to relate concerns about AI to broader questions of moral norms and the public good. Instead, the efforts of experts are more usefully focused on rather "narrow" technical issues and judgments (see "deliberative output 1" in the model presented in Figure 1). Such issues could include addressing system limitations due to evidence concerns about AI (e.g., trade-offs between reliability and costs, i.e., quantifying risks) or considering ways of making systems more intelligible by linking data inputs to conclusions to afford a better understanding of outcomes in relation to data, that is, *what* data are used (their scope, quality, etc.) and *how* data points are used for learning (Miller & Record, 2013). Among experts and system developers, the focus should be primarily on "how" explanations, that is,

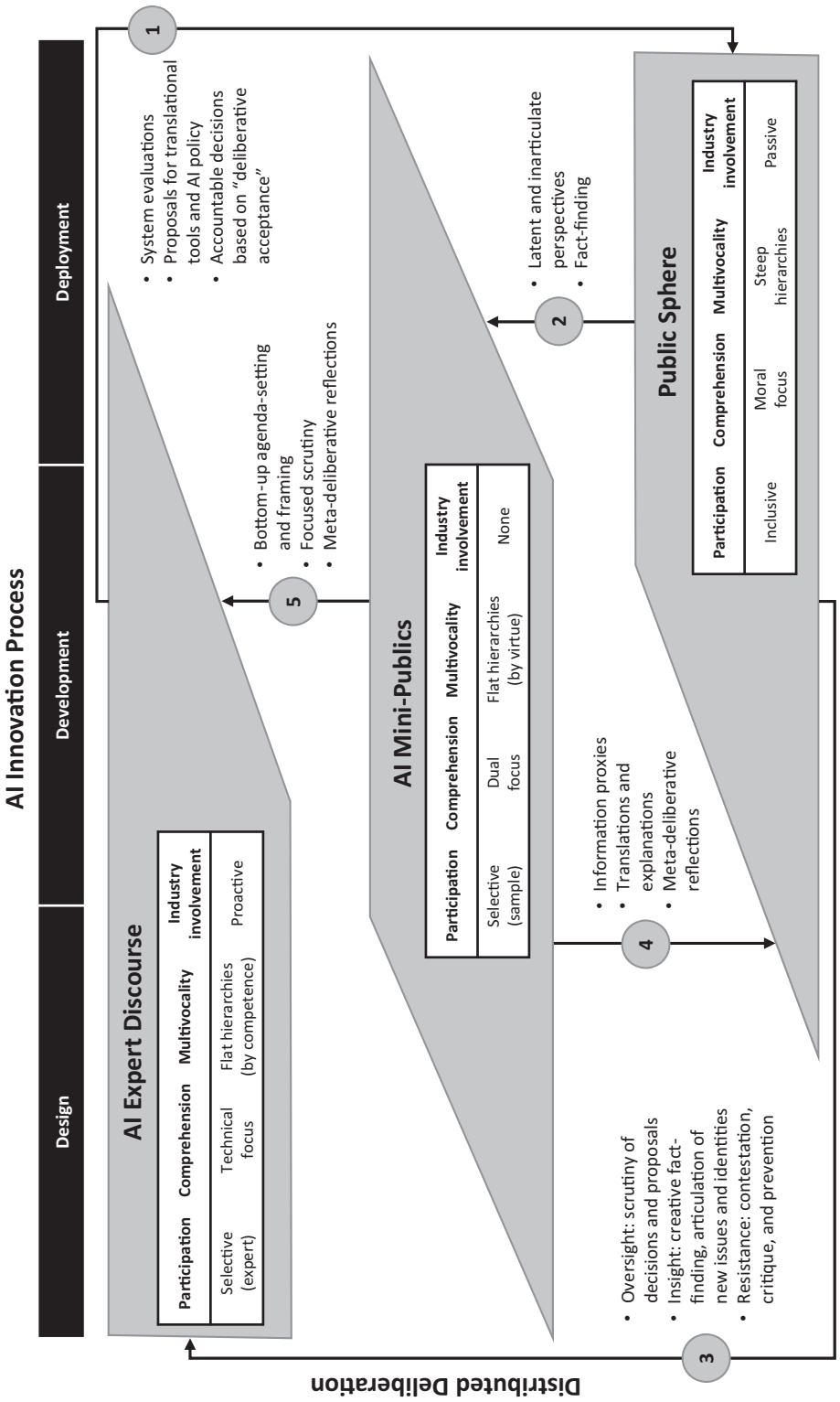

Figure 1: A Model of Distributed Deliberation for Responsible Innovation in Artificial Intelligence

on the interpretability of systems, qualitatively assessing whether they meet other desiderata, such as fairness, privacy, reliability, robustness, causality, usability, and trust (Doshi-Velez & Kim, 2017: 3). In principle, such *how* explanations can be both prospective and retrospective, that is, "How will the system operate and use data?" and "How and why were decisions reached?" (Preece, 2018). Toward outsiders, meanwhile, the focus should be on *why* explanations, that is, on providing reasons to end users as to why a particular course of action was taken (Dwivedi et al., 2019; Rahwan, 2018). The focus of AI expert discourse accordingly tends to be more on the front end of the AI development pipeline, where a larger share of issues can be meaningfully addressed at a technical level. Toward the back end of the pipeline, however, where issues emerge in the form of long-term transformative effects that transpire during system deployment, the focus of discourse needs to be on moral judgments and decisions regarding the public good. For this, AI expert discourse is a lot less conductive as inclusive participation is compromised. In AI expert discourse, the virtue focus is strongly on comprehension and multivocality. Unlike in wider public debate, this is possible because secluded deliberations among experts afford good opportunities for achieving a shared commitment to the equality of participants and for facilitating diverse reason giving, because knowledge about AI is equally distributed in these venues. This allows for a strong and transparent relationship to be established between the multitude of arguments weighed against each other and the decisions reached as an outcome of such deliberation (Moore, 2016). These decisions may take the form of expert consensus arising from unanimity of beliefs and evaluations (cf. "scientific consensus" in Turner, 2003) or as an equally collective decision that spans other differences and incorporates a willing suspension of possible disagreement (cf. "active consensus" in Beatty & Moore, 2010). In the latter case, the assumption is that potential disagreements can more readily be suspended based precisely on the propensity to flat hierarchies (based on the mutually high competence among participants) and strong multivocality amid experts.

Such informed suspension of disagreement in deliberation in turn provides people outside of secluded AI expert discourse with good reasons and confidence to accept expert system evaluations and proposals for translational tools and policy. To support this kind of trust in expertise, Moore (2016) proposes that expert discourse should work toward "deliberative acceptance" by signaling the deliberative quality of the expert venue to the outside. This could be achieved, for example, by experts taking a vote that must not only secure a majority but also a confirmation from those who disagree that their concerns and criticisms have been appropriately taken into account and that they have had ample opportunity to challenge and prevent the final expert decision.

The Public Sphere as the Main Venue for AI Scrutiny

Although most citizens and everyday users of AI and algorithms lack sufficient formal knowledge and qualifications to participate in AI expert discourse, they can nonetheless be considered as potential "citizen experts" (Fischer, 2000) inasmuch as they have experiential knowledge accumulated from varied and particular contexts

of AI systems in use. Such citizen expertise must thus be considered an important part of the overall process and dynamic of mitigating epistemic challenges in AI through distributed deliberation. For the public sphere tasked with deliberating on "AI in practice" and scrutinizing the system evaluations and AI policy proposals produced by expert venues, the focus should primarily be on normative concerns about AI that require broad judgments regarding moral norms and the public good. (This corresponds to "deliberative output 2" in our model, as an input for AI "mini-publics," which will be introduced as a separate venue later.) While such public scrutiny of AI may be of limited value for certain, more technical evaluations regarding design and development, it is indispensable at the back end of the AI innovation pipeline, that is, in the stages of AI testing, deployment, and monitoring. This highlights the important role of the public sphere, especially in tackling what Mittelstadt et al. (2016: 5) term "strictly ethical"—as opposed to more technical—problems with AI and in assessing the "observer-dependent fairness of the (AI) action and its effects." Tackling these problems could include addressing topics of concern, such as the potentially biased outcomes of AI applications and identifying and weighing up the potential long-term transformative effects of applying AI in particular social domains. While public-sphere deliberation should thus tend toward a greater emphasis on such back-end concerns, it is also needed to address certain aspects of the design and development stages, for example, through deliberations concerning potentially unethical and discriminatory variables (Veale & Binns, 2017).

In line with Moore's (2016) conclusions, the functions of the public sphere can be enacted in at least three important ways vis-à-vis AI expert discourse: 1) through overseeing and scrutinizing AI developers and experts based on "lifeworld-bound" perspectives, 2) through stimulating expert discourse by articulating new issues and identities that emerge from the everyday use of AI applications, and 3) through empowering and exercising resistance against problematic AI systems and policies. These functions correspond to "deliberative output 3" in our model in Figure 1.

AI Mini-publics as Mediating and Moderating Venues

On the basis of the concept of "mini-publics" (Setälä & Smith, 2018), we refer here to "AI mini-publics" as venues for deliberation that comprise a "sample of citizens" situated at the intersection of closed AI expert venues and the wider public sphere. Such venues may take the form of purposeful associations, citizen panels, AI think tanks, and interest groups. As Moore (2016) notes, the idea of "minipopuli" was proposed from early on as a way of bringing public judgment to bear on expert discourse. The literature on responsible innovation has described such venues as an important means for ensuring inclusion and for "upstreaming" public debate into the "technical parts" of governing innovation (Stilgoe et al., 2013: 1571). In agreement with arguments advanced by, for example, Moore (2016), Niemeyer (2011), and Brown (2009), we posit that AI mini-publics constitute a central mode for enabling and supporting rational public judgment of both AI systems and policy. Situated between the "expert layer" and lay citizens, AI mini-publics can concern themselves with all stages of the AI innovation pipeline, ranging from assessments of proposals

for AI use (during the design phase) to monitoring and evaluating the long-term transformative effects of AI in practice (during the deployment phase). As elucidated in what follows, mini-publics as venues of AI scrutiny serve at least three related functions.

First, AI mini-publics *mediate* between secluded AI expert venues and the public sphere by providing "palatable expertise," serving as "information proxies" (MacKenzie & Warren, 2012) that offer translations and explanations of poorly scrutable and traceable systems, AI policies, and expert arguments and decisions (corresponding to "deliberative output 4" in our model). In the case of efforts by AI industry actors to frame principles for AI governance (Schiff et al., 2021) or to help users and implementers understand their machine learning systems in action by providing interpretative tools (Mitchell et al., 2019), for example, these industry-led efforts could be gainfully contextualized and evaluated through the work of AI mini-publics. By supplementing these efforts with alternative assessments and explanations, AI mini-publics could strengthen users' understanding of otherwise mainly industry-based framings, thereby bolstering public resilience to potentially biased accounts. As such, AI mini-publics can augment the comparatively low capacity of public deliberation for comprehension and multivocality.

As "mediators," AI mini-publics need to have a dual focus of judgment to relate narrower technical issues with AI to broader normative concerns, and vice versa. In this sense, they are particularly well equipped to address "traceability issues" in AI, that is, to answer questions about the causes of these issues and the responsibilities of actors where broader questions about accountability relate to narrow technical issues in data sets or code design. As such, AI mini-publics are key long-term agents of traceability (Mittelstadt et al., 2016) and explicability (Floridi et al., 2018). This is because they can support what Morley et al. (2020) call "the development of a common language" beyond any expert community, linking terminologies and interpretations across diverse and dispersed deliberative venues with different knowledge, experiences, and virtue foci. This translating function seems especially important for enabling the public at large to reflect upon and stay alert to the long-term impacts and transformative effects of AI. This is crucial because such long-term effects, unlike the biases of specific applications or other more direct harmful outcomes of AI systems, can have much less obvious but wide-ranging harmful consequences.

Second, AI mini-publics can *moderate* the distributed efforts of deliberation on AI and its governance by providing reflections on the need for a division of deliberative labor to mitigate the epistemic challenges around AI. This is essential because for deliberative venues to interact constructively, their division of labor must itself be a potential object of justification through deliberation (Mansbridge et al., 2012). Epistemic inequalities resulting from AI developers' and other experts' knowledge and their deliberative distribution need to be accompanied by the possibility of "meta-deliberation" on the procedures and functional differentiation of the deliberative system itself. Such meta-deliberative reflection can be enacted though the work of mini-publics (Moore, 2016) and is thus included in our model in Figure 1 as outputs 4 and 5.

Third, AI mini-publics are important for generating and directing media attention to otherwise latent issues around AI. Such attention is often necessary for framing issues, widening popular mobilization, and deepening support for arguments or points of critique (Fung, 2003); it corresponds to "deliberative output 5" in our model. Media reporting on discrimination and unfairness, errors and mistakes, violations of social and legal norms, and human misuse of AI can be useful for exposing the contours of algorithmic power (Diakopoulos, 2019). Given that the media system is itself embedded in a hierarchical arena of communicative actors (Habermas, 2006), however, this system is difficult to penetrate, especially by unorganized civil society interests. Only a functioning media can, via a latent escalation potential, counteract the tendency within expert discourses to keep concerns latent and suppress public dissent (Moore, 2016).

In conclusion, we argue that venues currently being set up by associations like the Institute of Electrical and Electronics Engineers, the Royal Society, and groups hosted by the Future of Life Institute to discuss the workings and desiderata of autonomous systems can be positioned to perform the role and functions of AI mini-publics. These functions, as we have seen, include explaining, translating, and contextualizing the outputs of closed expert discourse for the public sphere. For example, this could involve developing contents and formats for "documentary procedures" by which to increase the transparency of systems, providing reflections through meta-deliberation on the necessity and particular "location" of boundaries between expert and nonexpert venues, and increasing the capacity for media attention to support the public scrutiny of otherwise latent issues. For AI mini-publics to exercise these roles, however, they need to exclude actors with organized particular interests, or what Moore (2016: 201) calls "partisans." This entails "cutting" the involvement of the AI industry wherever possible at this level.

DISCUSSION

The Promise and Peril of a Deliberatively Engaged AI Industry

We have started our article by proposing a new framework of responsibilities for innovation in AI. As an addition to extant frameworks for ethical AI (Floridi & Cowls, 2019), our framework highlights the importance of governance that draws on deliberation to address epistemic concerns as a meta-responsibility in AI innovation. On the basis of this framework, we have advocated for PCSR as an approach to such AI governance because it allows the foregrounding of the role of the AI industry in deliberative processes of principled communication and collective decision-making. Besides the discussed upsides of the PCSR approach and the challenges related to deliberation in responsible AI innovation (see section "Enacting Responsible AI Governance"), open participation and deliberation by AI industry actors can obviously create problems of agency, for instance, in cases when the disclosure and sharing of information lead to disproportionate advantages for competitors (Hippel & Van Krogh, 2003). For reasons of self-interest, therefore, including the desire to maintain power imbalances and information advantages, AI industry actors, just like other corporate actors (Hussain & Moriarty, 2018), would seem unlikely to be

willing to solve challenges deliberatively. In fact, the AI industry is often accused of disregarding participation and user consent in favor of "closed-door" decision-making and of prioritizing frictionless functionality in accordance with profit-driven business models (Campolo, Sanfilippo, Whittaker, & Crawford, 2017).

In addition to the poor scrutability and traceability of AI, the private interests and power of AI industry actors would seem to challenge the optimistic notion that the epistemic power of "deep democracy" could foster responsible AI governance by way of deliberatively engaging AI developers and other AI industry actors in a wider network of empowered actors from the public, private, and civil society sectors. Indeed, there are good reasons for rejecting proposals to extend the political role of businesses, not least on the basis that this could turn corporations into "supervising authorities" and thus lead rather to a democratic deficit than the desired increase in informed deliberation (Hussain & Moriarty, 2018). Such a normative approach to AI governance can further be criticized for shifting the focus of ethical expectations away from corporate conduct that is adaptive to external demands and concerns, instead proposing a discursive negotiation of ethical conduct that risks ultimately serving the interests of corporations and further suppressing already marginalized publics (Ehrnström-Fuentes, 2016; Whelan, 2012; Willke & Willke, 2008). For corporate actors, the prioritization of "mutual dialogue" to allegedly resolve issues may be a much *easier option* than changing or simply abandoning contested conduct (Banerjee, 2010). Normative stakeholder engagement in this case would thus constitute merely a means of deflection.

All this may inspire little optimism when it comes to a deliberatively engaged AI industry, not only as it relates to openness in addressing evidence, outcome, and epistemic concerns about AI, especially in the case of what we earlier labeled "strategic opacity," but also on a meta-level to the development of guidelines for soft governance. Consider, for instance, the recent criticism that the AI industry utilizes soft governance merely for purposes of "ethics washing" and for delaying regulation (Butcher & Beridze, 2019; Floridi, 2019b). However, we believe that our discussion on PCSR as an approach to enacting responsible AI governance and the related arguments regarding collective goals, legitimation, and organizational learning provide some promising grounds for positioning the role of the AI industry as less adversarial and more communicative than is often proposed to be the case with corporate innovators more generally (Brand, Blok, & Verweij, 2020; Hussain & Moriarty, 2018). In this article, we have presented and discussed instances in which fostering and participating in deliberation is not simply an "easier option" for the AI industry but the best available approach to manage responsible AI innovation in view of the poor scrutability and traceability of AI systems. Current instances of apparent "ethics washing" may well be part of an early stage on a longer path toward the substantive adoption and institutionalization of corporate social responsibility, which often starts with the adoption of ceremonial forms that may look like ethics washing (Haack, Martignoni, & Schoeneborn, 2020). In particular, the need to manage concerns about "expert" and "true" opacity in AI may eventually lead organizations to look less favorably on strategic approaches to risk management and merely instrumental stakeholder engagement (Van Huijstee & Glasbergen,

2008), leading them toward the adoption of more open, prosocial, and consensus-oriented approaches. (On this aspect, see Scherer & Palazzo, 2007, as well as Buhmann et al., 2020, in relation to AI developers more specifically.)

Augmenting Distributed Deliberation for Responsible Innovation in AI

Building on the PCSR perspective and a discussion of the challenges entailed in deliberation for responsible AI innovation, we have argued for the prospects of a "distributed deliberation" approach as a means of overcoming said challenges. And we have, subsequently, considered ways in which different venues can reach deliberative conclusions supportive of a broader, distributed process of deliberative governance for responsible AI innovation. In addition to the need for further empirical exploration of how different internal procedures for reaching decisions within AI expert discourse and AI mini-publics (on concrete AI systems and translational tools) can support the ability of citizens to make informed judgments in accepting or rejecting expert decisions, there are important theoretical questions to be considered.

First and foremost among these is the concern that the approach of distributed deliberation we have proposed is only convincing to the extent that AI mini-publics do indeed *supplement* rather than *replace* the critical public sphere and its judgments. This important criticism has previously been leveled against the concept of mini-publics on the basis that it could lead to "deliberative elitism" (Lafont, 2015), effectively displacing important instances of public-sphere scrutiny, including social movements. From this critical perspective, the outcomes of deliberation by AI mini-publics are seen not as vital information proxies directed at the broader public sphere in support of rational public discourse but rather as dominant elite recommendations that undermine public-sphere rationality (see the discussion of this critique in Moore, 2016). If this were the case in practice, distributed deliberation mediated by AI mini-publics would indeed serve to sharpen epistemic inequalities and further exacerbate ethical problems related to AI opacity. In agreement with other proponents of mini-publics (e.g., Brown, 2009; Fisher, 2000; Fung, 2003; Moore, 2016), however, we hold that AI mini-publics can meaningfully supplement and enhance public-sphere-level judgments on AI systems and policies. Accordingly, we propose that further research on responsible AI innovation and governance should include empirical exploration of how and to what extent this happens in practice. More specifically, this calls for a closer look at how the efforts of mini-publics can support public-sphere deliberations on the creation and evaluation of translational tools, especially because the creation of such tools has been proposed as a key focus for the machine learning expert community (Morley et al., 2020). These efforts could be explored at the level of professional associations, think tanks, advocacy groups, and more loose and time-bound workshop groups, conferences, and collaborations, such as those that produced the Asilomar AI Principles and the Montréal Declaration for a Responsible Development of AI.[3]

[3] See https://futureoflife.org/ai-principles as well as https://monoskop.org/images/b/b2/Report_Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf.

Second, we have suggested that AI expert discourse constitutes a venue type with a narrow focus on technical judgments rather than on moral norms or the common good. Although this seems an apt conceptualization of the work and dynamics of most closed AI expert venues, such as the Organisation for Economic Co-operation and Development's Expert Group on AI or the European Union's High-Level Expert Group on Artificial Intelligence, at least one important type of AI expert venue does not match this definition in any straightforward manner. This is the case with AI ethics councils, which, for example, although these are distinctly expert deliberative venues, also necessarily have an orientation toward questions of moral norms and the common good. The special role of AI ethics councils thus warrants particular theoretical and empirical attention in further developing the proposed distributed approach. (For related discussions on the role of ethics councils, see Wynne, 2001; Moore, 2010.)

Third, adopting a deliberative democracy approach takes on board this literature's strong reliance on present possibilities for human actors to argue and debate. This makes deliberation controversial when potentially "ethically affected" entities cannot participate in deliberation themselves—as, for example, in the case of animals, infants, the environment, or future generations (O'Neill, 2001). With this present-day and anthropocentric focus, our deliberative approach potentially fails to recognize important perspectives and concerns in AI ethics that are not directly linked to deliberating actors; such perspectives could include future generations or AI as an actor in its own right. Further developments based on our proposed approach may further specify how such perspectives and concerns can be addressed within deliberative solutions. Starting points for this can be the fundamental issues discussed early on by Habermas (1988, 1990) and subsequent operational discussions on governance (Tonn, 1996) or on the challenges of representation in deliberative democracy (O'Neill, 2001).

Future Research at the Intersection of Political Corporate Social Responsibility and Responsible AI

Notwithstanding these critical reflections, our arguments contribute in two parallel ways to advancing research on PCSR in relation to the "AI domain," and particularly so to scholarship focused on the role of nongovernmental actors in public governance. First, our work responds to the need flagged up by Scherer, Rasche, Palazzo, and Spicer (2016) for research to consider different contexts in which "governance gaps" require attention and involvement from politically engaged organizations, thus extending the scope of inquiry beyond the typical focus in this literature on governance gaps that come with the globalized context of multinational corporations —such as studies of fragile states (Scherer & Palazzo, 2007, 2011). Indeed, our arguments can extend recent work from the PCSR literature on governing innovation more generally (Scherer & Voegtlin, 2020; Voegtlin & Scherer, 2017). For although the challenges we discuss are particularly pervasive and pronounced in AI innovation, especially in the case of epistemic challenges, our proposed model for "distributing expertise" may also be applicable to other contexts of governing innovation in which steep knowledge inequalities can impede responsible governance (see, e.g., Dryzek & Pickering, 2017; Stirling, 2008).

Second, our discussions have addressed the gap identified by Scherer et al. (2016) in the PCSR literature regarding big data technologies and their impact on business and society (for a recent exception, see Rasche, Morsing, & Wetter, 2021). Here we specifically position our contribution in relation to the current "external evaluation gap" in ethical AI (Morley et al., 2021) by exploring processes to enable broader and more inclusive societal-level deliberation for responsible AI innovation. Analogous to Scherer and Voegtlin (2020), we argue here that, without active investment in deliberative governance, the AI industry runs the risk of relying on overly hierarchical decision-making procedures that may undermine social acceptance of AI. This could decouple the design and development of AI systems not only from collective goals for human well-being and the environment but also from what would be economically and technically the most efficient way to develop semiautonomous systems. Future work seeking to relate this discussion to operational-level translational tools should follow Scherer, Baumann-Pauly, and Schneider (2013) in further addressing the question of how to implement and internalize deliberative mechanisms within corporate organizational actors in the AI industry to support the deliberative governance of AI as a whole. Salient questions could include the following: How does the implementation of PCSR work *within* corporations that develop and apply AI? How do concrete responses to evidence, outcome, and epistemic responsibilities relate to the ways in which different business departments interact or how data scientists engage with general managers? How do AI developers make sense of responsibilities for deliberative governance in light of conflicting pressures between AI accountability and accuracy?

Related investigations could also further explore the relationship between the political role of nongovernmental actors and behavior at the individual level (cf. Frynas & Stephens, 2015; Scherer, 2018). Unlike most current discussion on applied AI ethics that adopts a microperspective on methods and principles for explainable and accountable AI geared toward AI practitioners (Morley et al., 2020), our work provides the basis for investigations into precisely this relationship between nongovernmental actors and individual-level behavior. Research questions along this line could explore the implications of "discursively shared responsibility" among corporations that develop AI and 1) individual end users and/or 2) programmers and data engineers. In this regard, another question is how the AI industry and its AI developers, programmers, engineers, and controllers can work together to be "discourse ready" for engaging in deliberation to achieve ethical AI outcomes—a line of inquiry analogous to that pursued by Constantinescu and Kaptein (2015). Developers are increasingly called upon to treat AI design choices as *political choices* and thus to reflect on their own professional roles in relation to the common good, which "should occur through open discussion and deliberation" (Green & Viljoen, 2020: 26). Here our work provides orientation on how this could be achieved in practice, what could motivate actors to engage in deliberations on AI systems, who to deliberate with, and which venues and "venue arrangements" would be most conducive to interactive deliberations.

Finally, the discussion we propose should apply a special focus on the challenges of "deliberative continuity." To date, many translational tools have been

implemented merely as tick-box tests of single steps or instances in AI design and development (Morley et al., 2021). Such a limited approach is especially inadequate given the procedural nature of algorithms, potentially leading to a form of checklist compliance that hinders the detection and solution of issues—as has been shown, for example, in the case of auditing procedures (LeBaron & Lister, 2016). Needed instead is a continuous process of validation, verification, and evaluation that addresses the following questions respectively: Is the right system being developed? Is the system being developed in the right way? Are there any emergent issues during deployment that require system revision or improvement? (Floridi, 2019a). Here it is worth reprinting the following point recently made by Morley et al. (2021: 244):

> Unless ethical evaluation becomes an integral part of a system's operation, there is no guarantee that pro-ethical translational tools will have any positive impact on the ethical implications of AI systems. Indeed, they could have a negative impact by fostering a false sense of security and consequential complacency.

The model we have proposed for distributed deliberation mediated by mini-publics suggests ways of iterating ethical evaluations across a deliberative system that can feed social evaluations into the development and implementation of translational tools while at the same time securing citizens' trust in AI innovation and its governance. Crucially, this distributed model means that citizens do not need to have the expert capacity to undertake independent examination of algorithmic systems or even to comprehend the rationale behind the tools and policies proposed to govern these systems.

CONCLUSION

As an emerging and fluid technology, AI tends to perpetuate an "institutional void" (Hajer, 2003) bereft of any agreed-upon rules and structures. Addressing this challenge calls for decentralized and open-ended governance in which deliberative settings or venues serve an important role for all actors involved. For this endeavor, the perspective of PCSR and the proposed model of distributed deliberation show strong potential. This is because they allow for the address of both the need for a politically engaged AI industry in governing innovation and the need to tackle the poor traceability and explicability of AI within such a governance approach. Although there are many legitimate concerns about the potential of politically engaged businesses to pose a threat to already fragile deliberative settings, our discussion in this article has highlighted the limits of top-down regulation and rigid accountability frameworks, suggesting that a politically *disengaged* AI industry may pose a much greater threat to the prospects of responsible innovation in AI.

Acknowledgments

This work was financially supported by the Norwegian Research Council as part of its Algorithmic Accountability: Designing Governance for Responsible Digital Transformations project (grant 299178).

REFERENCES

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. 2019. Fairwashing: The risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*: 161–70. Long Beach, CA: PMLR 97. https://arxiv.org/pdf/1901.09749.pdf.

Albu, O. B., & Flyverbom, M. 2019. Organizational transparency: Conceptualizations, conditions, and consequences. *Business and Society*, 58(2): 268–97.

Ananny, M., & Crawford, K. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, 20(3): 973–89.

Banerjee, S. B. 2010. Governing the global corporation: A critical perspective. *Business Ethics Quarterly*, 20(2): 265–74.

Barben, D., Fisher, E., Selin, C., & Guston, D. H. 2008. Anticipatory governance of nanotechnology: Foresight, engagement, and integration. In E. J. Hackett & O. Amsterdamska (Eds.), *The handbook of science and technology studies*: 979–1000. Cambridge, MA: MIT Press.

Baum, S. D. 2020. Social choice ethics in artificial intelligence. *AI and Society*, 35(1): 165–76.

Beatty, J., & Moore, A. 2010. Should we aim for consensus? *Episteme: A Journal of Social Epistemology*, 7(3): 198–214.

Beijing Academy of Artificial Intelligence. 2019. *Beijing AI principles.* https://www.baai.ac.cn/blog/beijing-ai-principles.

Berman, R., & Katona, Z. 2020. Curation algorithms and filter bubbles in social networks. *Marketing Science*, 39(2): 296–316.

Brand, T., & Blok, V. 2019. Responsible innovation in business: A critical reflection on deliberative engagement as a central governance mechanism. *Journal of Responsible Innovation*, 6(1): 4–24.

Brand, T., Blok, V., & Verweij, M. 2020. Stakeholder dialogue as agonistic deliberation: Exploring the role of conflict and self-interest in business-NGO interaction. *Business Ethics Quarterly*, 30(1): 3–30.

Brown, M. 2009. *Science in democracy: Expertise, institutions and representation*. Cambridge, MA: MIT Press.

Buhmann, A., & Fieseler, C. 2021a. Tackling the grand challenge of algorithmic opacity through principled robust action. *Morals and Machines*, 1(1): 74–85.

Buhmann, A., & Fieseler, C. 2021b. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, 64: 101475.

Buhmann, A., Paßmann, J., & Fieseler, C. (2020). Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse. *Journal of Business Ethics*, 163(2): 265–80.

Burrell, J. 2016. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data and Society*, 3(1): 1–12.

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

Calo, R. 2017. Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review*, 51(2): 399–435.

Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. 2017. *AI Now 2017 report*. New York: AI Now Institute at New York University. https://experts.illinois.edu/en/publications/ai-now-2017-report.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. 2018. Artificial intelligence and the "good society": The US, EU, and UK approach. *Science and Engineering Ethics*, 24(2): 505–28.

Chambers, S. 2003. Deliberative democratic theory. *Annual Review of Political Science*, 6: 307–26.

Coglianese, C., & Lehr, D. 2016. Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal*, 105(5): 1147–224.

Constantinescu, M., & Kaptein, M. 2015. Mutually enhancing responsibility: A theoretical exploration of the interaction mechanisms between individual and corporate moral responsibility. *Journal of Business Ethics*, 129(2): 325–39.

Crawford, K., & Calo, R. 2016. There is a blind spot in AI research. *Nature*, 538(7625): 311–13.

D'Agostino, M., & Durante, M. 2018. The governance of algorithms. *Philosophy and Technology*, 31(4): 499–505.

Diakopoulos, N. 2019. *Automating the news: How algorithms are rewriting the media*. Cambridge, MA: Harvard University Press.

Doshi-Velez, F., & Kim, B. 2017. *Towards a rigorous science of interpretable machine learning*. https://arxiv.org/abs/1702.08608.

Dryzek, J. S., & Pickering, J. 2017. Deliberation as a catalyst for reflexive environmental governance. *Ecological Economics*, 131(C): 353–60.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Vigneswara Ilavarasank, P., Janssen, M., Jones, P., Kumar Kar, A., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Kumar Sharma, S., Bahadur Singh, J., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., & Williams, M. D. 2019. Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57: 101994.

Ecoffet, A., Clune, J., & Lehman, J. 2020. Open questions in creating safe open-ended AI: Tensions between control and creativity. In *ALIFE 2020: The 2020 Conference on Artificial Life*: 27–35. Cambridge, MA: MIT Press.

Ehrnström-Fuentes, M. 2016. Delinking legitimacies: A pluriversal perspective on political CSR. *Journal of Management Studies*, 53(3): 433–62.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Fazelpour, S., & Lipton, Z. C. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*: 57–63. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. 2019. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data and Society*, 6(1): 1–14.

Fischer, F. 2000. *Citizens, experts and the environment: The politics of local knowledge*. Durham, NC: Duke University Press.

Floridi, L. 2016. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A*, 374 (2083): 20160112.

Floridi, L. 2019a. *The logic of information: A theory of philosophy as conceptual design* (1st ed.). New York: Oxford University Press.

Floridi, L. 2019b. Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy and Technology*, 32(2): 185–93.

Floridi, L., & Cowls, J. 2019. A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Scahfer, B., Valcke, P., & Vayena, E. 2018. AI4People—An ethical framework for a good AI Society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28: 689–707.

Frynas, J. G., & Stephens, S. 2015. Political corporate social responsibility: Reviewing theories and setting new agendas. *International Journal of Management Reviews*, 17(4): 483–509.

Fung, A. 2003. Recipes for public spheres: Eight institutional design choices and their consequences. *Journal of Political Philosophy*, 11(3): 338–67.

Fung, A. 2006. Varieties of participation in complex governance. *Public Administration Review*, 66: 66–75.

Glenn, T., & Monteith, S. 2014. Privacy in the digital world: Medical and health data outside of HIPAA protections. *Current Psychiatry Reports*, 16(494): 1–11.

Goodman, B., & Flaxman, S. 2017. European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine*, 38(3): 50–57.

Green, B., & Viljoen, S. 2020. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 19–31. New York: Association for Computing Machinery.

Haack, P., Martignoni, D., & Schoeneborn, D. 2020. A bait-and-switch model of corporate social responsibility. *Academy of Management Review*, 46(3): 440–64.

Habermas, J. 1988. Morality and ethical life: Does Hegel's critique of Kant apply to discourse ethics? *Northwestern University Law Review*, 83(1/2): 38–53.

Habermas, J. 1990. *Moral consciousness and communicative action*. Malden, MA: Polity Press.

Habermas, J. 1998. *On the pragmatics of communication*. Cambridge, MA: MIT Press.

Habermas, J. 2006. Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. *Communication Theory*, 16(4): 411–26.

Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1): 99–120.

Hajer, M. 2003. Policy without polity? Policy analysis and the institutional void. *Policy Sciences*, 36(2): 175–95.

Hartley, S., Pearce, W., & Taylor, A. 2017. Against the tide of depoliticisation: The politics of research governance. *Policy and Politics*, 45(3): 361–77.

Häußermann, J. J., & Lütge, C. 2021. Community-in-the-loop: Towards pluralistic value creation in AI, or—why AI needs business ethics. *AI Ethics.* DOI: 10.1007/s43681-021-00047-2.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Van Den Hoven, J., Zicar, R. V., & Zwitter, A. 2019. Will democracy survive Big Data and artificial intelligence? In D. Helbing (Ed.), *Towards digital enlightenment*: *Essays on the dark and light sides of the digital revolution*: 73–98. Cham, Switzerland: Springer.

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics guidelines for trustworthy AI.* Brussels: European Commission. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

Hippel, E. V., & Van Krogh, G. V. 2003. Open source software and the "private-collective" innovation model: Issues for organization science. *Organization Science*, 14(2): 209–23.

Horowitz, M. C. 2018. Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, 1(3): 36–57.

Hussain, W., & Moriarty, J. 2018. Accountable to whom? Rethinking the role of corporations in political CSR. *Journal of Business Ethics*, 149(3): 519–34.

Jobin, A., Ienca, M., & Vayena, E. 2019. The global landscape of ethics guidelines. *Nature Machine Intelligence*, 1: 389–99.

Jordan, A. 2008. The governance of sustainable development: Taking stock and looking forwards. *Environment and Planning C*, 26(1): 17–33.

Kaufmann, M., Egbert, S., & Leese, M. 2019. Predictive policing and the politics of patterns. *British Journal of Criminology*, 59(3): 674–92.

Kellogg, K. C., Valentine, M. A., & Christin, A. 2020. Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1): 366–410.

Kemper, J., & Kolkman, D. 2019. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication, and Society*, 22(14): 2081–96.

Korinek, A., & Stiglitz, J. 2018. Artificial intelligence and its implications for income distribution and unemployment. In A. Ajay, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda*: 349–90. Chicago: University of Chicago Press.

Lafont, C. 2015. Deliberation, participation, and democratic legitimacy: Should deliberative mini-publics shape public policy? *Journal of Political Philosophy*, 23(1): 40–63.

LeBaron, G., & Lister, J. 2016. Ethical audits and the supply chains of global corporations (Global Political Economy Brief No. 1). Sheffield, UK: Sheffield Political Economy Research Institute, University of Sheffield. https://speri.dept.shef.ac.uk/wp-content/uploads/2018/11/Global-Brief-1-Ethical-Audits-and-the-Supply-Chains-of-Global-Corporations.pdf.

Leese, M. 2014. The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue*, 45(5): 494–511.

Lubit, R. 2001. The keys to sustainable competitive advantage: Tacit knowledge and knowledge management. *Organizational Dynamics*, 29(3): 164–78.

MacKenzie, M. M., & Warren, M. E. 2012. Two trust-based uses of minipublics in democratic systems. In J. Parkinson & J. Mansbridge (Eds.), *Deliberative systems: Deliberative democracy at the large scale*: 95–124. Cambridge: Cambridge University Press.

Mansbridge, J., Bohman, J., Chambers, S., Christiano, T., Fung, A., Parkinson, J., Thompson, D., & Warren, M. 2012. A systemic approach to deliberative democracy. In J. Parkinson & J. Mansbridge (Eds.), *Deliberative systems: Deliberative democracy at the large scale*: 1–26. Cambridge: Cambridge University Press.

Martin, K. 2019. Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160(4): 835–50.

Meadowcroft, J., & Steurer, R. 2018. Assessment practices in the policy and politics cycles: A contribution to reflexive governance for sustainable development? *Journal of Environmental Policy and Planning*, 20(6): 734–51.

Michelman, F. I. 1997. How can the people ever make the law? A critique of deliberative democracy. *Modern Schoolman*, 74(4): 311–30.

Miller, B., & Record, I. 2013. Justified belief in a digital age: On the epistemic implications of secret Internet technologies. *Episteme*, 10(2): 117–34.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267(February): 1–38.

Mingers, J., & Walsham, G. 2010. Toward ethical information systems: The contribution of discourse ethics. *MIS Quarterly*, 34(4): 833–85.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. B., & Gebru, T. 2019. Model cards for model reporting. In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA*: 220–29. New York: Association for Computing Machinery.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data and Society*, 3(2): 1–21.

Moore, A. 2010. Public bioethics and deliberative democracy. *Political Studies*, 58(4): 715–30.

Moore, A. 2016. Deliberative elitism? Distributed deliberation and the organization of epistemic inequality. *Critical Policy Studies*, 10(2): 191–208.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. 2021. Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds and Machines*, 31(2): 239–56.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. 2020. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4): 2141–68.

Nanz, P., & Steffek, J. 2005. Assessing the democratic quality of deliberation in international governance: Criteria and research strategies. *Acta Politica*, 40(3): 368–83.

Niemeyer, S. (2011). The emancipatory effect of deliberation: Empirical lessons from mini-publics. *Politics and Society*, 39(1): 103–40.

Norgeot, B., Glicksberg, B. S., & Butte, A. J. 2019. A call for deep-learning healthcare. *Nature Medicine*, 25(1): 14–15.

O'Neill, J. 2001. Representing people, representing nature, representing the world. *Environment and Planning C*, 19(4): 483–500.

O'Neill, O. 2014. Trust, trustworthiness and accountability. In N. Morris & D. Vines (Eds.), *Capital failure: Rebuilding trust in financial services*: 172–89. Oxford: Oxford University Press.

Organisation for Economic Co-operation and Development. 2021. *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

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

Parkinson, J. 2006. *Deliberating in the real world*. Oxford: Oxford University Press.

Pasquale, F. 2015. *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Paßmann, J., & Boersma A. 2017. Unknowing algorithms: On transparency of unopenable black boxes. In M. T. Schäfer & K. van Es (Eds.), *The datafied society: Studying culture through data*: 139–46. Amsterdam: Amsterdam University Press.

Preece, A. 2018. Asking "why" in AI: Explainability of intelligent systems—perspectives and challenges. *Intelligent Systems in Accounting, Finance, and Management*, 25(2): 63–72.

Rahwan, I. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1): 5–14.

Rajwani, T., & Liedong, T. A. 2015. Political activity and firm performance within non-market research: A review and international comparative assessment. *Journal of World Business*, 50(2): 273–83.

Rasche, A. 2015. The corporation as a political actor: European and North American perspectives. *European Management Journal*, 33(1): 4–8.

Rasche, A., Morsing, M., & Wetter, E. 2021. Assessing the legitimacy of "open" and "closed" data partnerships for sustainable development. *Business and Society*, 60(3): 547–81.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–15.

Saltz, J. S., & Dewar, N. 2019. Data science ethical considerations: A systematic literature review and proposed project framework. *Ethics and Information Technology*, 21(3): 197–208.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. 2014. An algorithm audit. In S. P. Gangadharan (Ed.), *Data and discrimination: Collected essays*: 6–10. Washington, DC: New America Foundation.

Scherer, A. G. 2018. Theory assessment and agenda setting in political CSR: A critical theory perspective. *International Journal of Management Reviews*, 20(2): 387–410.

Scherer, A. G., Baumann-Pauly, D., & Schneider, A. 2013. Democratizing corporate governance: Compensating for the democratic deficit of corporate political activity and corporate citizenship. *Business and Society*, 52(3): 473–514.

Scherer, A. G., & Palazzo, G. 2007. Toward a political conception of corporate responsibility: Business and society seen from a Habermasian perspective. *Academy of Management Review*, 32(4): 1096–120.

Scherer, A. G., & Palazzo, G. 2010. The UN Global Compact as a learning approach. In A. Rasche & G. Kell (Eds.), *The United Nations Global Compact: Achievements, trends and challenges*: 234–47. Cambridge: Cambridge University Press.

Scherer, A. G., & Palazzo, G. 2011. The new political role of business in a globalized world: A review of a new perspective on CSR and its implications for the firm, governance and democracy. *Journal of Management Studies*, 48(4): 899–931.

Scherer, A. G., Palazzo, G., & Seidl, D. 2013. Managing legitimacy in complex and heterogeneous environments: Sustainable development in a globalized world. *Journal of Management Studies*, 50(2): 259–84.

Scherer, A. G., Rasche, A., Palazzo, G., & Spicer, A. 2016. Managing for political corporate social responsibility: New challenges and directions for PCSR 2.0. *Journal of Management Studies*, 53: 273–98.

Scherer, A. G., & Voegtlin, C. 2020. Corporate governance for responsible innovation: Approaches to corporate governance and their implications for sustainable development. *Academy of Management Perspectives*, 34(2): 182–208.

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. 2021. AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, 2(1): 31–42.

Setälä, M., & Smith, G. 2018. Mini-publics and deliberative democracy. In A. Bächtiger, J. S. Dryzek, J. Mansbridge, & M. Warren (Eds.), *The Oxford handbook of deliberative democracy*: 300–314. Oxford: Oxford University Press.

Stark, M., & Fins, J. J. 2013. What's not being shared in shared decision-making? *Hastings Center Report*, 43(4): 13–16.

Steenbergen, M. R., Bächtiger, A., Sporndli, M., & Steiner, J. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1): 21–48.

Stilgoe, J., Owen, R., & Macnaghten, P. 2013. Developing a framework for responsible innovation. *Research Policy*, 42(9): 1568–80.

Stirling, A. 2008. "Opening up" and "closing down": Power, participation, and pluralism in the social appraisal of technology. *Science, Technology, and Human Values*, 33(2): 262–94.

Taddeo, M., & Floridi, L. 2016. The debate on the moral responsibilities of online service providers. *Science and Engineering Ethics*, 22(6): 1575–603.

Thompson, D. F. 2008. Deliberative democratic theory and empirical political science. *Annual Review of Political Science*, 11: 497–520.

Tonn, B. 1996. A design for future-oriented government. *Futures*, 28(5): 413–31.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. 2021. The ethics of algorithms: Key problems and solutions. *AI and Society*. DOI: 10.1007/s00146-021-01154-8.

Tufekci, Z. 2015. Facebook said its algorithms do help form echo chambers, and the tech press missed it. *New Perspectives Quarterly*, 32(3): 9–12.

Turner, S. P. 2003. *Liberal democracy 3.0*. London: Sage.

Tutt, A. 2016. An FDA for algorithms. *Administrative Law Review*, 69: 83–123.

Van Huijstee, M., & Glasbergen, P. 2008. The practice of stakeholder dialogue between multinationals and NGOs. *Corporate Social Responsibility and Environmental Management*, 15(5): 298–310.

Veale, M., & Binns, R. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data and Society*, 4(2): 1–17.

Voegtlin, C., & Scherer, A. G. 2017. Responsible innovation and the innovation of responsibility: Governing sustainable development in a globalized world. *Journal of Business Ethics*, 143(2): 227–43.

Whelan, G. 2012. The political perspective of corporate social responsibility: A critical research agenda. *Business Ethics Quarterly*, 22(4): 709–37.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. 2018. *AI Now report 2018*. New York: AI Now Institute at New York University. https://ainowinstitute.org/AI_Now_2018_Report.pdf.

Willke, H., & Willke, G. 2008. Corporate moral legitimacy and the legitimacy of morals: A critique of Palazzo/Scherer's communicative framework. *Journal of Business Ethics*, 81(1), 27–38.

Wong, P.-H. 2020. Democratizing algorithmic fairness. *Philosophy and Technology*, 33: 225–44.

Wynne, B. 2001. Creating public alienation: Expert cultures of risk and ethics on GMOs. *Science as Culture*, 10(4): 445–81.

Yun, J., Lee, D., Ahn, H., Park, K., & Yigitcanlar, T. 2016. Not deep learning but autonomous learning of open innovation for sustainable artificial intelligence. *Sustainability*, 8(8): 1–20.

Zarsky, T. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, and Human Values*, 41(1): 118–32.

. . .

ALEXANDER BUHMANN (alexander.buhmann@bi.no, corresponding author) is associate professor of corporate communication at BI Norwegian Business School and director of the Nordic Alliance for Communication and Management. His current research interests evolve around reputation management, digitalization, and the accountability and governance of artificial intelligence. Buhmann received his PhD in communication studies from the University of Fribourg, Switzerland.

CHRISTIAN FIESELER is professor for communication management at BI Norwegian Business School and a director of the Nordic Centre for Internet and Society. His research is focused on the question of how individuals and organizations adapt to the shift brought by new, digital media and how to design participative and inclusive spaces in this new media regime. Fieseler received his PhD in management and economics from the University of St. Gallen, Switzerland.