Technology in Society 64 (2021) 101475

Contents lists available at ScienceDirect

Technology in Society

journal homepage: http://www.elsevier.com/locate/techsoc

Towards a deliberative framework for responsible innovation in artificial intelligence

Alexander Buhmann *, Christian Fieseler

BI Norwegian Business School, Department of Communication and Culture, NO-0442, Oslo, Norway

ARTICLE INFO

Keywords:
Artificial intelligence
AI governance
Accountability
Transparency
Deliberation

ABSTRACT

The rapid innovation in artificial intelligence (AI) is raising concerns regarding human autonomy, agency, fairness, and justice. While responsible stewardship of innovation calls for public engagement, inclusiveness, and informed discourse, AI seemingly challenges such informed discourse by way of its opacity (poor transparency, explainability, and accountability). We apply a deliberative approach to propose a framework for responsible innovation in AI. This framework foregrounds discourse principles geared to help offset these opacity challenges. To support better public governance, we consider the mutual roles and dependencies of organizations that develop and apply AI, as well as civil society actors, and investigative media in exploring pathways for responsible AI innovation.

1. Introduction

Technological innovation can produce both benefits and unforeseen, harmful consequences. Examples range from nuclear power and bioengineering to genetically modified foods [1–3]. Innovation in Artificial Intelligence (AI) is no different. On the one hand, AI is proving to deliver increased organisational performance and decisions [4]. Trough machine learning techniques and deep neural networks, algorithms are able to learn and succeed at solving more complex tasks than ever before [5]. On the other hand, like humans, AIs can fail their intended goals, either because the training data they use may be biased or because their recommendations and decisions may yield unintended and negative consequences [6,7]. And since AIs have been suggested to affect, e.g., justice, privacy, stock and commodity trading, labour markets, and even the outcomes of democratic elections [8–12], governments around the world have started to recognize AI as a significant and global threat to safeguarding social goods, justice, and welfare [13]. The pivotal question thus arises: How can we foster sustainable AI that is not harmful but beneficial to human life? [14].

Recent works in both the business and technology ethics literatures have emphasized and discussed the value of deliberative engagement for shaping responsible innovation [3,15–17]. With private enterprises often initiating and steering technological innovations, arguably there is a need for forms of "extended corporate citizenship" [18] in which ethical businesses engage with local actors, governments, and civil society to foster responsible processes for innovation [19,20]. These works widely—implicitly and explicitly—draw on concepts of "deep democracy" [21,22], highlighting the epistemic potential of open engagement processes rooted in principled communication among multiple stakeholders. Broadly speaking, this literature proposes that innovators, as proactive participants of a wider public debate and discourse, can contribute to responsible processes of innovation. To harness the potential of such public engagement in finding optimal solutions, this process should satisfy conditions of inclusiveness to various stakeholders, ensure the reciprocity of actors, and facilitate diverse and well-informed arguments and viewpoints.

However, with artificial intelligences, such as machine learning, we are currently witnessing innovations that seemingly *undermine* these very deliberative conditions meant to facilitate responsible innovation. While the proponents of deliberative models for responsible innovation would suggest to involve AI innovators from the private sector in the proactive facilitation of rational public debate (cf., e.g., [3,15–17,23], the often problematized opacity of AI and autonomous systems [25] may severely undermine the very conditions necessary for safeguarding "deliberative rationality". While new technologies are always to some extent perceived as being opaque, modern algorithmic decision systems seem to come with a special kind of opacity—as compared, e.g., to the global proliferation and societal penetration of earlier technologies such as the car, electricity, and the telephone. This is because machine-learning algorithms are not only a set of rules defined by

* Corresponding author.
E-mail addresses: alexander.buhmann@bi.no (A. Buhmann), christian.fieseler@bi.no (C. Fieseler).

https://doi.org/10.1016/j.techsoc.2020.101475
Received 15 June 2019; Received in revised form 1 December 2020; Accepted 4 December 2020
Available online 18 January 2021

programmers but also contain algorithmically self-produced rules of learning [25]. These procedures may for practical purposes be structurally inaccessible and incomprehensible not only to laypersons but oftentimes also, at least in everyday practice, to the organizations that own and employ them, and even to system programmers and specialists [25,26]. The core societal issue concerning algorithmic decision systems, therefore, is that they may not routinely and easily be accessed for public scrutiny. This is not only because they are proprietary entities of the organizations who own or license them and who may keep them private to ensure functionality, competitiveness, and the confidentiality of data [27–30] but more importantly due to *technical* and *procedural* factors. This raises the question of how exactly, under such conditions, deliberation may contribute to responsible AI governance and policy.

In this article we build on the recent discussion in the business and technology ethics literature that calls for deliberatively engaged innovators for fostering responsible innovation [23,31,95]. Such frameworks need not only discuss new forms of governance and regulation that encompass the contributions of non-state actors, such private-sector innovators, but also new types of innovations, such as AI, that may pose unique challenges to society and the proposed deliberative frameworks in particular. While the business and technology ethics literatures have started to address the general question of whether and how deliberative engagement is suitable for fostering responsible innovation [15], in this article we take up the specific discussion about the prospects and challenges of deliberation for responsible innovation in AI. In particular, we address the role and functions of public fora, to explore pathways to participation in technology design [32], to suggest how a society may meaningfully debate, and eventually agree on, systemic compromises for the governance of AI [33].

As such, our paper relates and contributes to two current fields of debate. First, we add to the literature on deliberative models of "deep democracy" to conceptualize new forms of governance that emphasize the role of corporate actors, such as AI developers [34,35,96], and that has started, more recently, to influence the discussion on frameworks for responsible innovation [15,23,36,95]. We extend on this literature to propose a deliberative framework for responsible innovation in artificial intelligence. Specifically, we do so by a) discussing the ideal requirements for deliberation to the intricate conditions of AI opacity—which seemingly contest deliberative process for fostering responsible innovation—and b) relating the roles of key actors (corporate AI developers, inquisitive media and journalism, as well as an engaged civil society) in deliberative exploration and evaluation of responsible implementations of AI.

Second, we add to the current discussion on AI ethics [37,38,98] by going beyond the common 'micro perspective' on methods and principles for explainable and accountable AI, on the one hand, and the concentration on government regulation on the other, adding to this a *macro layer* that relates these approaches to broader societal processes of engagement and legitimation and discusses the role of AI developers therein—that is, in relation to other key actors within public deliberation about AI. Specifically, we propose that, together with discursive contributions from corporations and 'fluid observation' facilitated through quality media, the bottom-up identification, interpretation, and problematization of AI in practice achieved by a critical civil society can mount a deliberative framework for responsible innovation in AI.

2. AI and the need for responsible innovation

2.1. AI and responsible innovation

From a business perspective, innovation is a "multi-stage process whereby organizations transform ideas into new/improved products, service or processes, in order to advance, compete and differentiate themselves successfully in their marketplace" ([39]: 1334). AI innovation, herein, refers to an organization's endeavour to train algorithms to mimic functions typically associated with human attributes such as vision and speech but also language processing, learning, and problem-solving, and to scale these functions via software [40]. Broadly speaking, algorithms are "encoded procedures for transforming input data into a desired output, based on specified calculations" ([41]: 167). Hence, algorithms are not always software and they can be found in any culture with sufficiently developed mathematical procedures. Yet, their rapid proliferation is a consequence of digitization.

Responsible innovation encompasses three main dimensions that are directly reflected also in the current discourse on AI ethics ([23]; cf. Table 1). First, the *responsibility to avoid harm*, which refers, e.g., to risk management approaches supposed to control for potentially harmful consequences. In recent principles for ethical AI we see this dimension reflected in calls for non-maleficence of AI [38] or AI robustness, security, and safety [42]. Second, the *responsibility to do good* refers to the improvement of living conditions, such as in accordance with the sustainable development goals. Calls for AI to promote prosperity [43] or serve humanity by furthering human values [44] reflect principles for responsible AI on this dimension. Finally, *governance responsibility* refers to the responsibility to create and support global governance structures that can facilitate the former two responsibilities. On this dimension, recent principles for ethical AI have addressed, for instance, tensions between the need to predict AI impact on the one hand and the inability to draw boundaries around fluid technology on the other [45].

With its focus on public value, the ethos underlying responsible innovation necessitates not only to "place a premium on inclusive participation" ([17]: 754) but, more specifically on "substantive processes of inclusive reflection and deliberative democracy" ([17]: 755).

2.2. The challenges of AI explainability

Emerging AI legislation increasingly pursues the idea of a "right to explanation" [51]. As the ways in which AI reaches decisions become unforeseeable, demands for greater transparency come to the fore. These are typically divided into prospective and retrospective transparency

Table 1
Dimensions of responsible innovation and emerging principles for ethical AI.

Avoid harm	'Do good'	Ethical governance
• Non-Maleficence [38]	• Beneficence, justice [38]	• Regulations against possible future development which could be detrimental to human and societal values [46]
• Protect autonomy and ability to make good decisions [43]	• AI must be beneficial to humanity and promote prosperity [43]	• Tension between the need to predict AI impact and inability to draw boundaries around this highly dynamic technology [45].
• Respect for human autonomy, prevention of harm [47]	• Fairness [47]	• Mechanisms of governance that minimize risks and potential pitfalls [48,49].
• Robustness, security and safety [42]	• Inclusive growth, sustainable development and well-being [42]	• Embedding ethical principles into AI systems to ensure that they act morally [50].
• Awareness and mitigation of negative impacts; ensure data security and AI safety; minimizing discrimination and bias [44]	• Promote human society and the environment; diversity (benefit as many people as possible); serve humanity by furthering human values including freedom and autonomy [44]	• AI ethics require continuous study of moral beliefs and behavior to ensure reasonable and well-founded policy.

[52,53]. Prospective transparency informs users about the data processing and working of the system upfront, describing how the AI system reaches decisions in general [54]. Retrospective transparency, on the other hand, refers to post hoc explanations and rationales [55], revealing how and why a certain decision was reached in a specific case, describing the data processing step by step. Such 'explicability' is seen as a foundational principle of vital importance in the ethical-AI community because, to a certain extent, it lays the foundation for developing and ensuring ethical AI in the first place [43]. There is a difference between "how" explanations, which are useful for AI system developers, and "why" explanations, which are most helpful to end-users [49]: The first, concerning the interpretability of systems, can assist in qualitatively ascertaining whether other desiderata are met, such as fairness, privacy, reliability, robustness, causality, usability, and trust [56]. The second refers to providing an explanation to outside parties as to why a particular course of action was taken.

Felzmann and colleagues (2019: 8) argue that 'such practices do not take place in a social vacuum but in specific cultural and organizational settings, and that transparency can intentionally occlude, for example, when so much information is strategically disclosed that it is impractical or impossible to sift through ('needle in the haystack' problem)' (based on Albu, & Flyverbom [58]; Ananny and Crawford [27]. It is thus crucial to consider the information literacy of the data subjects as well [52,59].

2.3. The challenges of AI accountability

Going beyond mere explanation, then leads into challenges for accountability—i.e., the justification of decisions or judgements to an evaluative audience [60]. This extends to managerial accountability within organizations but also requires the existence of effective external systems of accountability [52]. Achieving accountability might rely "on democratic or corporate forms of governance, or on legal, financial or professional forms" ([61]: 177). Whereas for most technologies, "purification work" [62] between a company and the public, technology and its users, the inside and the outside, can be accomplished quite easily, in constellations of distributive information and opacity production such purification is, at least on first inspection, counter-productive (cf. [63]: 3). Large arrays of involved actors, and the manner in which they associate and interact, may further exacerbate complexity in that they affect people beyond the immediate reach of relevant organizations, since organizations are in need of assessment beyond their own boundaries. AI unfolds not within a single organisation but at field level, where actors and actions are more diverse, and thus more difficult to govern than within organizations. Any understanding of a shared issue is likely to be continuously (re)negotiated [64]. As such, AI opacity often cannot simply be 'tackled' by demanding that organizations 'make their algorithms transparent' based on a fixed standard or framework [98]. In other words, there is no straightforward way to address poorly transparent and highly fluid algorithmic processes, and organizations may struggle to simply 'deliver accounts' of these technologies. Rather, such processes need to be addressed by organizations in a participative and discursive process together with their stakeholders [98], i.e. by adopting the 'pragmatic treatment' that Ferraro et al. [63] have proposed for grand challenges [65]. Along these lines Kemper and Kolkman ([57]: 2083) argue that the "transparency of algorithms can only be attained by virtue of an interested critical audience." Similarly, recent work on ethics in AI and information systems has proposed building on normative concepts of communicative action and discourse ethics to develop principles for multi-stakeholder engagement and procedural norms for public communication as a means to enable 'collective truth tracking' and safeguard the accountability of complex systems [36,66,98].

3. Communicative principles for responsible AI

In the face of unintended negative consequences of AI and the seeming opacity of this technology, communicative and deliberative approaches may offer fitting solutions as a) the far-reaching societal ramifications of AIs and their rapid proliferations in all public and private spheres of human life make them a central object of broad political concern; and b) what is needed for opaque AI systems especially is the 'epistemic power' of deliberation to improve knowledge and feedback through self-correcting learning processes among empowered actors.

In the following, we focus on normative requirements for deliberation and discourse as well as recent work on AI accountability to first propose a number of basic communicative principles meant to foster discursive spaces for responsible AI [98]. We then build on this work by elaborating on how key actors from the spheres of private business, (quality) media and journalism, as well as the wider civil society may, based on such principles, mutually work towards more responsible governance of AI innovation.

The discourse-ethical approach implies that public debate about algorithms has the capacity to enable actors to collectively mitigate the opacity challenges posed by AI. However, the rational potential of discourses can only be harnessed if basic communicative principles are met. Numerous approaches have been proposed to arrive at a comprehensive set of dimensions with which to assess whether factual discourses measure up to the ideal communicative principles [67]. Below we adhere closely to a set of four principles proposed by Nanz and Steffek [68] in the context of international governance. Their approach fits well for working towards a deliberative framework for AI accountability and governance because the principles are developed with a focus on practical discourses involving complex and contentious issues and are designed with empirical measures in mind that would allow for operationalization using empirical measures to assess discourse quality (see [98] for a detailed discussion of these four principles).

The first of these four principles is that the intricate issues around algorithmic accountability need to be discussed in an open forum in which every subject with the competence to speak and act is provided an opportunity to take part in the debate. Specifically, all those who potentially suffer the negative effects of the processes and decisions of algorithmic systems should have equal access to a forum and a communicative process that aims to spotlight potential issues and facilitate argumentation with the aim of arriving at broadly acceptable decisions. Stakeholders need institutionalised *access to deliberative settings* to ensure they have a chance to voice their concerns, opinions, and arguments.

Second, all those who participate in the deliberative process need to have as much information as possible about the issues at stake, the various suggestions for their solution, and the ramifications of these proposed solutions. This principle points directly to a fundamental challenge in accounting for complex algorithms: "While datasets may be extremely large but possible to comprehend and code may be written with clarity, the interplay between the two in the mechanism of the algorithm is what yields the complexity (and thus opacity)" ([25]: 5). This puts the emphasis on *enabling comprehension* of the joint and procedural dynamic of data and code in a given algorithmic context.

Third, in addition to the inclusion of informed stakeholders, the inclusion of all arguments (*multivocality*) is a central principle for enabling rational discourse and deliberation. This is because participants should have the opportunity to see and assess an issue from all possible viewpoints. All of those who are possibly affected should have a chance to communicate all their concerns. While participation and access to information are preconditions for a process of deliberation to take place, it is this inclusion of all arguments that constitutes the main precondition of the rationality of that process.

Finally, the three principles above are meaningless if the different concerns and suggestions put forth by various stakeholders regarding algorithmic systems are not adequately taken into account and are unable to influence recommendations or decisions in practice as a result of the discourse [68]. The process needs to be clearly *responsive* to these suggestions and concerns.

The opacity challenge of algorithms and its 'tackling' via, above all,

communicative means, highlights the 'middle way' proposed by Habermas [69] for securing social progress through technological advancement. As Dai and Hao ([70]: 10), have recently put it: "we should combine technological advancements with public concerns for questions about the good life and discussions of values, and by constantly re-evaluating the applications of technology to society, mitigate its negative effects". As such, the normative principles outlined above help not only a) to detect and describe the extent to which specific cases may lack or have the potential to resolve accountability issues through discourse, but also b) to clarify that the 'opacity challenge' found in contemporary debate is also a common (and recurring) diagnosis about new technologies. This requires the integration of these issues in the long-established normative frameworks of democratic societies rather than an adjustment of these frameworks to the alleged specificity of algorithmic opacity that merely dispenses of it as an object of ethical concern [98].

4. Towards a deliberative framework for responsible innovation in AI: challenges and prospects

Our argument until now has related ideal requirements for deliberation to the intricate conditions of AI opacity. Herein deliberation appears as both a necessary but, seemingly, also technically contested process for fostering responsible innovation in AI. Our investigation highlighted the role of principled multi-stakeholder engagement. To develop this argument further towards a deliberative framework for AI governance, we extend on the above principles to relate the roles of key actors, i.e., corporate AI developers, inquisitive media and journalism as well as an engaged civil society for exploring and deciding on responsible implementations of AI. We shall propose that, together with discursive contributions from corporations and 'fluid observation' facilitated through quality media, the bottom-up identification, interpretation, and problematization of AI in practice achieved by a critical civil society can mount a basic deliberative framework for responsible innovation in AI.

For developing operational propositions of how deliberation may contribute to responsible AI governance and policy, we propose to regard and specify our argument in relation to three main actor perspectives: The perspective of the *corporation* addresses the important role of developers and proprietors of AI and their engagement in collective action and deliberation. To explicate this role in context, we then further address the perspectives of media and civil society respectively. Specifically, the *media* is addressed in its role to link public sphere communication with civil society on the one hand and with the state and social functional systems on the other [22], and the perspective of *civil society* emphasizes the contribution of informal, unorganized, and 'weak' publics that critically serve the identification, interpretation, and problematization of the social ramifications of AI in practice.

This combined focus of the contribution of the corporation on the one hand with the related role of civil society and media actors on the other resonates with those normative requirements that have been given particular emphasis in recent works on deliberative democracy more generally, as these highlight the need to further explore the central role of empowered quality media and civil society in offsetting deliberative deficits [22] – an anrgument which we extend here towards the prospects and challenges of deliberation for responsible AI in particular. This idealized, unified space, in practice, is of course rather dispersed and prone to 'irrational' interference by corporate and otherwise organized interests [99]. 'True' deliberative engagement is easily skewed through strategic interference by strategic actors. Sceptically viewed, a deliberative role of AI innovators seems unlikely even if poor explainability and accountability of AI systems were not an issue. Organizational contexts arguably tend precisely to *block* communicative action [71] and corporations are often claimed to be unable to abstain from self-interest and strategic action to the extent necessary to foster 'true' deliberation and discourse [72,73]. In relation to responsible innovation in AI, such practical concerns are most obvious for the proprietary nature of these systems and organizations' inclination to keep them secret in order to ensure functionality and competitiveness [27–30]. In these instances, AI developers often exert significant control over how systems are described, interpreted, and evaluated. This shows, e.g., in their production of own algorithm validation studies with questionable informational value [74].

4.1. The corporation: towards engaged AI innovators

For reasons of self-interest, power imbalances, and information advantages [75], the corporate sector is seemingly unlikely to solve AI challenges deliberatively. However, we propose that ethical business could be part of the solution, not least from motives of enlightened self-interest. Arguably, open participation and deliberation can create agency problems for corporate actors when disclosure and sharing create disproportionate advantages for competitors [76]. However, corporations can also profit through options of gathering new and critical information, facilitating learning as well as building reputation and safeguarding against criticism directed at irresponsible conduct (and thus realise reputation and risk management) [23]. Especially in the case of AI, we suggest that such *epistemological and reputational concerns* may serve as key drivers for business firms to engage in deliberative processes for responsible innovation, for two main reasons: First, as particular developers and applicants of algorithms do not necessarily possess privileged knowledge for assessing these systems, stakeholders need to play an active part in assessing AI. Second, the very opacity of AI constitutes a permanent and lingering reputational concern for developers, proprietors, and users alike, because when critical stakeholders demand information and transparency these actors will inevitably struggle to give explanations and provide satisfactory accounts to a critical forum [98].

As such, epistemological and reputational concerns in AI may pressure organizations towards accountability and compel them to enter proactive debates. The practically often unattainable 'AI transparency' produces limitations on the ethical duties of organizations to deliver conventional explanations and accounts. Here, we see not only a pragmatic necessity for managing reputation but also an ethical obligation for organizations to enable and take part in open, dialogical, and rational discourse with their stakeholders. This necessitates that AI innovators consider a discursive approach to stakeholder engagement akin to 'Habermasian approaches' to moral legitimation [77] or deliberatively engaged public relations [101]. Such an approach may help to facilitate legitimate outcomes, *especially* under conditions of unclear (external) demands related to opaque information systems [66], where knowledge about the workings and ramifications of AI does not reside exclusively within an organisation but must emerge from open deliberation with actors in the organisation's environment who are affected by these systems [78].

As such, epistemological and reputational concerns provide motives to consider the role of AI innovators as potentially less adversarial and more communicative than often proposed for corporate actors more generally [75,79]. Strategic intentions of risk management may then move more easily from instrumental approaches of stakeholder engagement to open, pro-social, and consensus-oriented approaches [96]. Engagement by business would not only be geared towards the objectives of compliance and accountability but would also be aimed, more broadly, at gathering insights and enabling diverse and informative feedback on AI in practice, as well as for the continuous development and improvement of systems.

For such deliberative engagement to work to full effect, however, in terms of mitigating epistemological and reputational concerns, the deliberative role of the firm needs to be complemented with both empowered media and citizenry.

4.2. The media: empowered quality journalism for fluid observation

Both traditional and new forms of journalism serve as central accountability mechanism in public deliberation [80,97]—also for AI systems and algorithms [81]. Looking at available approaches and measures to support principled deliberation on AI innovation, the journalistic system appears as the central mediator in supporting participation, comprehension, and multivocality. The pivotal role of the journalistic system is highlighted by the inability to fully govern fluid AI technology. Rigid certification processes, for instance, would not be able to do justice to the speed at which most complex algorithmic systems change.

With this set of principles quality journalism is well-positioned to enquire into AI systems, developers, and policy to hold actors accountable—both in the sense of investing the resources and skills needed to enquire into the kind of information fed into algorithms, but also as watchdogs of the algorithms that feed information into the discursive ecosystem of a society [82]. Journalism assumes curiosity and a desire to understand and query the functions of the world in general, and ideally also of the algorithmic world. It also assumes the role of a 'translator and explainer' to general audiences. Journalistic means of scrutinizing and reporting on discrimination and unfairness, errors and mistakes, social and legal norm violations and human misuse of AI can be instrumental to publicly elucidate the contours of algorithmic power [82]. Data journalism, for instance, as an emerging field, sets aside the time to reengineer algorithms, scrape, collect and connect data, file freedom of information requests, and weigh the inputs, outputs, outcomes and effects of these exposed autonomous systems. Coming from an understanding of data and computation, journalists inform about the 'noisy nature' of data and the uncertainty that comes with predictions, and creates narratives of how AI systems operate. It is against this media-bolstered scrutiny that members of AI and expert discourse are then often challenged with specific objections to their expert authority. A functioning media counteracts, via a latent escalation potential, the tendency within expert discourses to simply keep concerns latent and supress public dissent [83]. To further prevent this, proactive reasons for expert deliberation are necessary, that is public justifications of the necessity to, at times, seclude expert venues from public deliberation (e.g., for scrutinizing systems or developing policy proposals for AI governance).

4.3. The civil society: empowered citizenry in un- and semi-organized spaces

The proposed deliberative framework also foregrounds the role of interactions with 'ordinary citizens', as opposed to formalized encounters with more organized stakeholders. To adequately theorize the place and role of AI innovators in deliberative settings in relation to more unorganized citizenry, it is necessary to address explicitly the institutions of "background justice of democratic deliberation" [84], i.e. to explicate the role of 'weak' publics constituted of lifeworld-bound associations and informal private actors central to the identification and interpretation of social problems. Such specification is necessary also as to not simply imply (and potentially overestimate) the ability of civil society to hold AI innovaotrs accountable [73].

Fundamentally, though, the emphasis on an empowered citizenry aligns with recent work on AI that specifically stresses the importance of fostering critical and informed publics and of engaging ordinary citizens for greater transparency and robust opinion formation and judgement in relation to AI technology [57]. Viewed form this 'citizenry perspective', processes and decisions related to AI governance and policy need to be under the continuous observation of a *critical and commenting* public. However, such critical commenting and input hinges not only on ordinary citizens' commitment to engagement and reciprocity, but also on their *capacities* for public reason-giving [21] in the specific domain of AI. Here, AI explainability and accountability may appear as central challenges. Oftentimes, the crucial information simply cannot be accessed by laypeople, and even in cases where it can be accessed it may not be comprehensible in any sense that can serve as meaningful basis for public debate.

Taking this 'citizenry perspective' in AI further elucidates that the current discussion in deliberative theorizing on the practical conditions that tend to render deliberative processes insufficient leaves out important points that pertain to AI explainability and accountability specifically: This discussion tends to focus either on: a) the social or cultural marginalization of groups as well as (technological or issue-based) segmentation of public discourse and thereby a group-wise exclusion from public debate [22,85,86]; or b) a colonialization of the public sphere by market imperatives and thereby a commodification of public debate [22,87]. Much less emphasis is given, however, to the *practical means* citizens have at their disposal to enlighten their experiences and debate. We hold that, especially for responsible AI innovation, these above discussions on the organization and formats of public communication in mass media and the colonization of political debate by strategic interests of marked actors, need to be supplemented by a focus on tools, approaches, and methods of making AI technology 'experienceable', explainable in context, and accountable to and by ordinary citizens. This is an argument similar to that recently put forth by Morley et al. [38]; who, in order to allow for greater explainability and accountability in AI design decisions, list over a hundred tools and methods and relate them specifically to stages of AI development (from design to testing and monitoring) and ethical concerns they can help to address (e.g., explicability, interpretability, justice, and autonomy).

From the perspective of a deliberative framework for responsible innovation in AI – and its focus on the role of civil society engagement in particular – the current discussion about tools and methods for AI explainability, interpretability, and accountability [38] needs to be extended to the question of how these methods and approaches are conducive to support and broaden people's lived experience with AI and improve their means of expression. As a basis for reasoned public dialogue made up of individuals' participation in a polyphonic discourse, these dialogues need to meaningfully intersect with participants' lived experience with AI. The essential claims of knowledge and value that emerge from the civic basis of the communicative arena need to be bolstered by individuals' 'AI literacy' [88], where the development of mental models of AI algorithms is experimentally' grounded [89] and ties in with the unorganized realm of everyday experience. Such local spheres of experimentation are crucial for the provision of a robust diversity of perspectives in deliberation, which means not just diverse opinions but diverse *viewpoints* [90]. At local level, experiential sources of opinion-formation can empower "mini publics" that support and build up otherwise latent perspectives on AI that are especially disadvantaged in public communication, which tends to privilege the perspectives of powerful groups and actors [91,92]. This would serve to counterbalance otherwise dominant claims and valuations from AI experts and political elites, and is an essential deliberative component in the identification and interpretation of social problems with AI in practice.

Finally, next to the level of lived experience in the unorganized realm of the lifeworld, the process of deliberation calls for the identified issues with AI to enter arenas of public communication. This crucial link of civic conversation and public debate has been problematized in relation to the inability of marginalized communities to apply public relations and other instrumental and professional means of communication to shape public debate [93,100]. Here, as in other contexts involving deliberations over complex issues [70], the enabling role of semi-organized civil associations as well as NGOs and civil society organization merits further exploration for responsible innovation in AI.

5. Conclusion

One of the grand challenges of our time is the fostering of sustainable

AI that is not harmful but beneficial to human life. Solutions to this challenge come, indeed, with established procedures for participatory technology design and public fora to meaningfully debate, and eventually agree on, systemic compromises for the governance of AI. Viewed from the perspective of deliberation for responsible innovation, however, an important variant of the above challenge is: how can we offset the poor transparency, explainability, and accountability of AI to enable public reasoning for responsible AI governance? In this paper we have argued that technical opacity of AI cannot serve as a general excuse to resist scrutiny; for while the practical opacity of algorithms may absolve AI innovators and proprietors of some of the typical duties of account-giving, at the same time this opacity charges them with additional duties to facilitate ongoing discourses about algorithms, as technologies for understanding algorithmic action are developed further. Just as in other domains in which simply the right to transparency will not create fairness [94], the ethical solution lies in the creation of discursive processes and fora. In the context of practical opacity, what is needed to ensure accountable and responsible AI are not merely reporting standards but standards for accountability *discourses* [98].

Here, we argue, deliberation serves an important function for both epistemic as well as moral justification in AI by highlighting particular tensions between common and ideal requirements for public reasoning on the one hand and particular challenges related to AI explainability and accountability on the other. A combined focus on the contribution of corporations with the related role of civil society and media actors resonates with those normative requirements that have been given particular emphasis in recent works on deliberative democracy more generally, as these highlight the need to further explore the central role of empowered quality media and civil society in offsetting deliberative deficits [22]. Together with discursive contributions from corporations and 'fluid observation' facilitated through the media, the bottom-up identification, interpretation, and problematization of AI in practice achieved by a critical civil society can mount a deliberative framework for responsible innovation in AI. Whether or not such deliberative multi-stakeholder arrangements can indeed be 'deep' in that they harness the rationalizing force of true discursive engagement, hinges on the adherence to normative principles of participation, comprehensibility, multivocality, and responsiveness.

Acknowledgement

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

References

[1] C. Groves, Technological futures and non-reciprocal responsibility, Int. J. Humanit. 4 (2) (2006) 57–61.
[2] A. Irwin, The politics of talk: coming to terms with the 'new' scientific governance, Soc. Stud. Sci. 36 (2) (2006) 299–320.
[3] J. Stilgoe, R. Owen, P. Macnaghten, Developing a framework for responsible innovation, Res. Pol. 42 (9) (2013) 1568–1580.
[4] T.H. Davenport, R. Ronanki, Artificial intelligence for the real world, Harv. Bus. Rev. 96 (1) (2018) 108–116.
[5] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidgeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglu, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.
[6] K. Crawford, R. Calo, There is a blind spot in AI research, Nature 538 (7625) (2016) 311–313.
[7] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: mapping the debate, Big Data Soc. 3 (2) (2016) 1–21.
[8] B.A. Barnet, Idiomedia: the rise of personalized, aggregated content', Continuum 23 (1) (2009) 93–99.
[9] R. Calo, Artificial Intelligence policy: a primer and roadmap, UC Davis Law Rev. 51 (2) (2017) 399–435.
[10] V. Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, St. Martin's Press, New York, 2018.
[11] A. Tutt, 'An FDA for algorithms', Adm. Law Rev. 69 (1) (2016). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2747994. (Accessed 6 October 2018).
[12] T. Zarsky, The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making, Sci. Technol. Hum. Val. 41 (1) (2016) 118–132.
[13] C. Cath, Governing artificial intelligence: ethical, legal and technical opportunities and challenges, Phil. Trans. Ser. A, Math. Phys. Eng. Sci. 376 (2133) (2018).
[14] J. Yun, D. Lee, H. Ahn, K. Park, T. Yigitcanlar, Not deep learning but autonomous learning of open innovation for sustainable artificial intelligence, Sustainability 8 (8) (2016) 797.
[15] T. Brand, V. Blok, Responsible innovation in business: a critical reflection on deliberative engagement as a central governance mechanism, J. Responsible Innovat. 6 (1) (2019) 4–24.
[16] R. Lubberink, V. Blok, J. Van Ophem, O. Omta, Lessons for responsible innovation in the business context: a systematic literature review of responsible, social and sustainable innovation practices, Sustainability 9 (5) (2017) 1–31, https://doi.org/10.3390/su9050721.
[17] R. Owen, P. Macnaghten, J. Stilgoe, Responsible research and innovation: from science in society to science for society, with society, Sci. Publ. Pol. 39 (6) (2012) 751–760.
[18] D. Matten, A. Crane, Corporate citizenship: toward an extended theoretical conceptualization, Acad. Manag. Rev. 30 (1) (2005) 166–179.
[19] R. Adams, J. Bessant, R. Phelps, Innovation management measurement: a review, Int. J. Manag. Rev. 8 (1) (2006) 21–47.
[20] M. Stafford-Smith, D. Griggs, O. Gaffney, F. Ullah, B. Reyers, N. Kanie, B. Stigson, P. Shrivastava, M. Leach, D. O'Connell, Integration: the key to implementing the sustainable development goals, Sustain. Sci. 12 (6) (2017) 911–919.
[21] F. Michelman, How can the people ever make the laws? A critique of deliberative democracy, in: J. Bohman, W. Rehg (Eds.), Deliberative Democracy: Essays on Reason and Politics: 145–172, MIT Press, Cambridge, MA, 1997.
[22] J. Habermas, Political communication in media society: does democracy still have an epistemic dimension? The impact of normative theory on empirical research, in: J. Ha-bermas (Ed.), Europe: the Faltering Project (Trans. C. Cronin): 138—183, Poli-ty, Malden, MA, 2009.
[23] C. Voegtlin, A.G. Scherer, Responsible innovation and the innovation of responsibility: governing sustainable development in a globalized world, J. Bus. Ethics 143 (2) (2017) 227–243.
[25] J. Burrell, 'How the machine 'thinks': understanding opacity in machine learning algorithms', Big Data Soc. 3 (1) (2016) 1–17.
[26] M. Ananny, Toward an ethics of algorithms: convening, observation, probability, and timeliness, Sci. Technol. Hum. Val. 41 (1) (2016) 93–117.
[27] M. Ananny, K. Crawford, Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability, New Media Soc. 2 (3) (2018) 1–17.
[28] T. Glenn, S. Monteith, Privacy in the digital world: medical and health data outside of HIPAA protections, Curr. Psychiatr. Rep. 16 (11) (2014) 1–11, 494.
[29] M. Leese, The new profiling: algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union, Secur. Dialog. 45 (5) (2014) 494–511.
[30] M. Stark, J.J. Fins, What's not being shared in shared decision-making? Hastings Cent. Rep. 43 (4) (2013) 13–16.
[31] A. Kerr, R.L. Hill, C. Till, The limits of responsible innovation: exploring care, vulnerability and precision medicine, Technol. Soc. 52 (2018) 24–31.
[32] C. Selin, R. Hudson, Envisioning nanotechnology: new media and future-oriented stakeholder dialogue, Technol. Soc. 32 (3) (2010) 173–182.
[33] C. Voinea, Designing for conviviality, Technol. Soc. 52 (2018) 70–78.
[34] A.G. Scherer, G. Palazzo, 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, J. Manag. Stud. 48 (4) (2011) 899–931.
[35] G. Palazzo, A.G. Scherer, Corporate legitimacy as deliberation: a communicative framework, J. Bus. Ethics 66 (1) (2006) 71–88.
[36] H. Berkowitz, Meta-organizing firms' capabilities for sustainable innovation: a conceptual framework, J. Clean. Prod. 175 (2018) 420–430.
[37] K. Martin, Ethical implications and accountability of algorithms, J. Bus. Ethics (2018), https://doi.org/10.1007/s10551-018-3921-3. Avaliable at:.
[38] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices, Sci. Eng. Ethics (2019) 1–28, https://doi.org/10.1007/s11948-019-00165-5.
[39] A. Baregheh, J. Rowley, S. Sambrook, Towards a multidisciplinary definition of innovation, Manag. Decis. 47 (8) (2009) 1323–1339.
[40] S.J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Pearson Education Limited, Malaysia, 2016.
[41] T. Gillespie, The relevance of algorithms, in: T. Gillespie, P.J. Boczkowski, K. A. Foot (Eds.), Media Technologies. Essays on Communication, Materiality, and Society, MIT Press, Cambridge, MA, 2014, pp. 167–194.
[42] OECD, Recommendation of the Council on Artificial Intelligence, 2019. Retrieved from, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. (Accessed 12 January 2020).
[43] L. Floridi, T. Clement-Jones, The five principles key to any ethical framework for AI, Tech New Statesman (2019). Retrieved from, https://tech.newstatesman.com/policy/ai-ethics-framework. (Accessed 12 January 2020).
[44] Beijing AI Principles, Retrieved from, https://www.baai.ac.cn/blog/beijing-ai-principles, 2019. (Accessed 12 January 2020).

[45] B.W. Wirtz, J.C. Weyerer, C. Geyer, Artificial intelligence and the public Sector—applications and challenges, Int. J. Publ. Adm. 42 (7) (2019) 596–615.
[46] Y. Duan, J.S. Edwards, Y.K. Dwivedi, Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda, Int. J. Inf. Manag. 48 (2019) 63–71.
[47] European Commission, Ethics Guidelines for Trustworthy AI, 2019. Retrieved from, https://ec.europa.eu/futurium/en/ai-alliance-consultation. (Accessed 12 January 2020).
[48] U. Gasser, V.A.F. Almeida, A layered model for AI governance, IEEE Internet Comput. 21 (6) (2017) 58–62, https://doi.org/10.1109/MIC.2017.4180835.
[49] I. Rahwan, Society-in-the-loop: programming the algorithmic social contract, Ethics Inf. Technol. 20 (1) (2018) 5–14.
[50] M. Anderson, S.L. Anderson (Eds.), Machine Ethics, Cambridge University Press, New York, NY, 2011.
[51] Council of Europe, Guidelines on the Protection of Individuals with Regard to the Processing of Personal Data in a World of Big Data, 2017. T-PD(1). Strasbourg.
[52] H. Felzmann, E.F. Villaronga, C. Lutz, A. Tamò-Larrieux, Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns, Big Data & Soc. 6 (1) (2019), 2053951719860542.
[53] A. Preece, Asking 'why' in AI: explainability of intelligent systems—perspectives and challenges, Intell. Syst. Account. Finance Manag. 25 (2) (2018) 63–72.
[54] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, Transparency in algorithmic and human decision-making: is there a double standard? Phil. Technol. 32 (4) (2018) 661–683.
[55] P. Paal, D. Pauly, Kommentar zur Datenschutzgrundverordnung und dem Bundesdatenschutzgesetz, C. H. Beck, Munich, 2018.
[56] F. Doshi-Velez, B. Kim, Towards a Rigorous Science of Interpretable Machine Learning, 2017 arXiv preprint arXiv:1702.08608.
[57] J. Kemper, D. Kolkman, Transparent to whom? No algorithmic accountability without a critical audience, Inf. Commun. Soc. 22 (14) (2019) 2081–2096.
[58] O.B. Albu, M. Flyverbom, Organizational transparency: conceptualizations, conditions, and consequences, Bus. Soc. 58 (2) (2019) 268–297.
[59] M. Bartsch, T. Dienlin, Control your Facebook: an analysis of online privacy literacy, Comput. Hum. Behav. 56 (2016) 147–154.
[60] M. Bovens, Two concepts of accountability: accountability as a virtue and as a mechanism, W. Eur. Polit. 33 (5) (2010) 946–967.
[61] O. O'Neill, Trust, trustworthiness and accountability, in: N. Morris, D. Vines (Eds.), Capital Failure: Rebuilding Trust in Financial Services: 172–189, Oxford University Press, Oxford, 2014.
[62] B. Latour, We Have Never Been Modern, Harvard University Press, Cambridge, MA, 1993. C. Porter (trans.).
[63] F. Ferraro, D. Etzion, J. Gehman, Tackling grand challenges pragmatically: robust action revisited, Organ. Stud. 36 (3) (2015) 363–390.
[64] S. Grodal, S. O'Mahony, How does a grand challenge become displaced? Explaining the duality of field mobilization, Acad. Manag. J. 60 (5) (2017) 1801–1827.
[65] G. George, J. Howard-Grenville, A. Joshi, L. Tihanyi, Understanding and tackling societal grand challenges through management research, Acad. Manag. J. 59 (6) (2016) 1880–1895.
[66] J. Mingers, G. Walsham, Toward ethical information systems: the contribution of discourse ethics, MIS Q. 34 (4) (2010) 833–885.
[67] M.R. Steenbergen, A. Bachtiger, M. Sporndli, J. Steiner, Measuring political deliberation: a discourse quality index, Comp. Eur. Polit. 1 (1) (2003) 21–48.
[68] P. Nanz, J. Steffek, Assessing the democratic quality of deliberation in international governance: criteria and research strategies, Acta Politic. 40 (3) (2005) 368–383.
[69] J. Habermas, in: J. Shapiro (Ed.), Toward a Rational Society, Beacon, Boston, 1970 trans.
[70] Y.X. Dai, S.T. Hao, Transcending the opposition between techno-utopianism and techno-dystopianism, Technol. Soc. 53 (2018) 9–13.
[71] M. Alvesson, A. Spicer, A stupidity-based theory of organizations, J. Manag. Stud. 49 (7) (2012) 1194–1220.
[72] J. Noland, R. Phillips, Stakeholder engagement, discourse ethics and strategic management, Int. J. Manag. Rev. 12 (1) (2010) 39–49.
[73] G. Whelan, The political perspective of corporate social responsibility: a critical research agenda, Bus. Ethics Q. 22 (4) (2012) 709–737.
[74] Electronic Privacy Information Center (EPIC), Algorithms in the criminal justice system, Available at: https://epic.org/algorithmic-transparency/crim-justice/, 2017. (Accessed 26 January 2020).
[75] W. Hussain, J. Moriarty, Accountable to whom? Rethinking the role of corporations in political CSR, J. Bus. Ethics 149 (3) (2018) 519–534.
[76] E.V. Hippel, G.V. Krogh, Open source software and the "private-collective" innovation model: issues for organization science, Organ. Sci. 14 (2) (2003) 209–223.
[77] A.G. Scherer, G. Palazzo, D. Seidl, Managing legitimacy in complex and heterogeneous environments: sustainable development in a globalized world, J. Manag. Stud. 50 (2013) 259–284.
[78] R. Lubit, The keys to sustainable competitive advantage: tacit knowledge and knowledge management, Organ. Dynam. 29 (3) (2001) 164–178.
[79] T. Brand, V. Blok, M. Verweij, Stakeholder dialogue as agonistic deliberation: exploring the role of conflict and self-interest in business-NGO interaction, Bus. Ethics Q. (2020) 1–28.
[80] Albert W. Dzur, Public journalism and deliberative democracy, Polity 34 (3) (2002) 313–336.
[81] N. Diakopoulos, Algorithmic accountability, Digital Journal. 3 (3) (2015) 398–415.
[82] N. Diakopoulos, Automating the News: How Algorithms Are Rewriting the Media, Harvard University Press, Cambridge, MA, 2019.
[83] A. Moore, Deliberative elitism? Distributed deliberation and the organization of epistemic inequality, Crit. Pol. Stud. 10 (2) (2016) 191–208.
[84] J. Mäkinen, A. Kourula, Pluralism in political corporate social responsibility, Bus. Ethics Q. 22 (4) (2012) 649–678.
[85] R.J. Dalton, Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies, CQ Press, Washington, 2006.
[86] T.T. Lee, Media effects on political disengagement revisited: a multiple-media approach, Journal. Mass Commun. Q. 82 (2) (2005) 416–433.
[87] C. Boggs, The great retreat: decline of the public sphere in late twentieth-century America, Theor. Soc. 26 (6) (1997) 741–780.
[88] P. Jandrić, The postdigital challenge of critical media literacy, Int. J. Crit. Media Literacy 1 (1) (2019) 26–37.
[89] P. Hingston, B. Combes, M. Masek, Teaching an undergraduate AI course with games and simulation, in: International Conference on Technologies for E-Learning and Digital Entertainment, Springer, Berlin, Heidelberg, 2006, pp. 494–506.
[90] G.A. Hauser, Vernacular discourse and the epistemic dimension of public opinion, Commun. Theor. 17 (4) (2007) 333–339.
[91] J. Bohman, Deliberative democracy and the epistemic benefits of diversity, Episteme 3 (3) (2006) 175–191.
[92] J. Bohman, Political communication and the epistemic value of diversity: deliberation and legitimation in media societies, Commun. Theor. 17 (4) (2007) 348–355.
[93] S.B. Banerjee, Governing the global corporation: a critical perspective, Bus. Ethics Q. 20 (2) (2010) 265–274.
[94] L. Edwards, M. Veale, 'Slave to the algorithm? Why a 'right to explanation' is probably not the remedy you are looking for', Duke Law Technol. Rev. 16 (1) (2017) 18–84, https://doi.org/10.2139/ssrn.2972855. Available at: SSRN: https://ssrn.com/abstract=2972855.
[95] A. Scherer, Theory assessment and agenda setting in political CSR: a critical theory perspective, Int. J. Manag. Rev. 20 (2) (2018) 387–410.
[96] A. Scherer, G. Palazzo, Toward a political conception of corporate responsibility: business and society seen from a Habermasian perspective, Acad. Manag. Rev. 32 (4) (2007) 1096–1120.
[97] P. Norris, Watchdog journalism. The Oxford Handbook of Public Accountability, Oxford University Press, 2014.
[98] A. Buhmann, J. Paßmann, C. Fieseler, Managing algorithmic accountability: balancing reputational concerns, engagement strategies, and the potential of rational discourse, J. Bus. Ethics (163) (2020) 265–280, https://doi.org/10.1007/s10551-019-04226-4.
[99] D. Thompson, Deliberative democratic theory and empirical political science, Annu. Rev. Polit. Sci. 11 (2008) 497–520.
[100] H. Willke, G. Willke, Corporate moral legitimacy and the legitimacy of morals: a critique of Palazzo/Scherer's communicative framework, J. Bus. Ethics 81 (1) (2008) 27–38.
[101] L. Edwards, The role of public relations in deliberative systems, J. Commun. 66 (1) (2016), https://doi.org/10.1111/jcom.12199.