DOI: http://dx.doi.org/xxxx

Buhmann, A., Paßmann, J. & Fieseler, C. Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse. *J Bus Ethics* **163,** 265–280 (2020). https://doi.org/10.1007/s10551-019-04226-4

This is a preprint of an article that is in press at the *Journal of Business Ethics*. Please cite as:

Buhmann, A., Paßmann, J., & Fieseler, C. (2019). Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse. *Journal of Business Ethics*, 1-16. https://doi.org/10.1007/s10551-019-04226-4

Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies and the Potential of Rational Discourse

Alexander Buhmann
BI Norwegian Business School
alexander.buhmann@bi.no

Johannes Paßmann
Siegen University

Christian Fieseler
BI Norwegian Business School

Abstract:

While organizations today make extensive use of complex algorithms, the notion of algorithmic accountability remains an elusive ideal due to the opacity and fluidity of algorithms. In this article, we develop a framework for managing algorithmic accountability that highlights three interrelated dimensions: reputational concerns, engagement strategies, and discourse principles. The framework clarifies: (a) that accountability processes for algorithms are driven by reputational concerns about the epistemic setup, opacity, and outcomes of algorithms; (b) that the way in which organizations practically engage with emergent expectations about algo-

rithms may be manipulative, adaptive, or moral; and (c) that when accountability relationships are heavily burdened by the opacity and fluidity of complex algorithmic systems, the emphasis of engagement should shift to a rational communication process through which a continuous and tentative assessment of the development, workings, and consequences of algorithms can be achieved over time. The degree to which such engagement is, in fact, rational can be as-sessed based on four discourse-ethical principles of participation, comprehension, multivocali-ty, and responsiveness. We conclude that the framework may help organizations and their environments to jointly work towards greater accountability for complex algorithms. It may further help organizations in reputational positioning surrounding accountability issues. The discourse-ethical principles introduced in this article are meant to elevate these positioning contests to extend beyond mere adaption or compliance and help guide organizations to find moral and forward-looking solutions to accountability issues.

1. Introduction

Different domains of organizational conduct have very different conditions for accountability. As organizational, technical and environmental complexity increases, the ongoing negotiation of accountability increasingly exposes organizations to reputational concerns (Scherer, Palazzo, and Seidl, 2013). One of the most current and pressing examples of this is the proliferation of algorithms as agents of complex computerized decision-making—with considerable social ramifications (Beer, 2009; Pasquale, 2015; Martin, 2018). Broadly speaking, algorithms are "encoded procedures for transforming input data into a desired output, based on specified calculations" (Gillespie, 2014, p. 167). That means, algorithms do not have to be software, and in many cases, they can theoretically be performed by humans. But only when performed by computers, they can proliferate as rational means of every-day decision making. Whereas

algorithms can be found in any culture with somewhat developed mathematical procedures, their rapid proliferation is a consequence of digitization.

The decisions algorithms make are often implicit and invisible. Yet, they yield intentional and unintentional consequences, e.g. for equality, privacy, stock and commodity exchange or even election outcomes, which increasingly makes them objects of public concern and scrutiny (Mittelstadt, Allo, Taddeo, Wachter, and Floridi 2016; Tutt, 2016). However, algorithms are also notorious for making such scrutiny nearly impossible, as they frequently remain "black boxes" (Graham, 2005; Pasquale, 2015) that create challenges for accountability in two main ways: The first is a *strategic* accountability challenge as many algorithms are proprietary entities. Organizations have strong incentives to keep them secret in order to ensure functionality, competitiveness, or the privacy of user data (Ananny and Crawford, 2016; Glenn and Monteith, 2014; Leese, 2014; Stark and Fins, 2013). Second, many algorithms pose *technical* accountability challenges. Their inner workings remain unclear to most stakeholders, and especially machine learning algorithms are deemed challenging to comprehend—even for specialists (Ananny, 2016, Burrell, 2016).

The growing public concern about algorithms as well as their black boxing, raise the question of how organizations are to account for their critical processes and consequences. While this has become a main focus in fields such as computer science (Datta, Sen and Zick, 2017), journalism studies (Diakopoulos, 2015), law (Doshi-Velez, and Kortz, 2017), and media studies (Sandvig et al., 2014a) research in business ethics has so far only sporadically addressed it. Early work on the subject pointed to the general need for "notice and choice" (Danna and Gandy, 2002)—stakeholders need to be informed of the ways in which information about them is used

by computerized systems. More recent research has stressed the obligation of algorithm designers to remain accountable for the ethical implications of the algorithms they develop (Martin, 2018). However, these works do not explain how the need for stakeholder information emerges as an expectation vis-à-vis distinct concerns with algorithmic systems, nor do they clarify how "notice and choice" can practically be realized within accountability processes given the opacity and fluidity of complex decision-making systems.

In this article, we argue that accountability relationships are mediated by reputational concerns, which act as a filter for external expectations and thus explain the varying degrees of interest in and intensity of accountability processes (Busuioc and Lodge, 2017). Reputational concerns, in this sense, go beyond a problem focus on critical events—such as instances of corporate misconduct or crisis (Coombs, 2013; Karpoff, 2012)—to include all actual or potential topics that can be considered a basis for general reputational contests. In a discussion of common reputational concerns related to algorithms, we show that concerns related to opacity are particularly obtrusive. We further argue that in the face of algorithmic opacity, expectations about accountability are highly fluid, and hence, accountability requires a constant, substantial engagement and active negotiation processes between organizations and their environment. More and more, reputation management also means algorithm management and vice versa. As algorithmic ramifications are not necessarily visible in a laboratory situation of designing and testing, but rather in the field of actual practice and engagement (Martin, 2018), the arena for this algorithm management is in large parts a public discourse. If simple accounts cannot be given by any one party, the emphasis should shift to an inclusive communication process through which a continuous and tentative assessment of the development, workings, and consequences of algorithmic technologies can be achieved over time. We argue that discourse-ethical principles are key to addressing accountability in the context of highly fluid and

constantly evolving information systems. Rather than a definite mechanism, we propose ethical principles that can be used to evaluate particular accountability practices and mechanisms as well as the effects that these mechanisms yield.

We intend for this paper to advance research in three related domains. We first add to the literature on corporate reputation by complementing work on reputation building through stakeholder engagement (Owen, Swift, Humphrey, and Bowerman, 2000; Romenti, 2010; Swift, 2001) with a framework of discourse-ethical principles for accountability in light of the poor transparency and proliferation of complex algorithmic systems. Second, we extend the literature on applied discourse ethics with a framework for corporate accountability that goes beyond the common focus on extant accountability standards (Gilbert and Rasche, 2007; Rasche and Esser, 2006). Instead we emphasize the mediating role of reputational concerns to address a new and problematic domain of corporate conduct where stakeholders must be continuously engaged in order to overcome highly restricted opportunities for account-giving and -holding. Third, we contribute to the literature on organizational accountability by focusing on a unique context in the balancing between reputational concerns and account-giving (Bovens, 2007; Busuioc and Lodge, 2017; Dubnick and Frederickson, 2010), in which the core reputational concerns reside in the fundamental inability to account for the practices in question.

2. Algorithms as an Accountability Issue

Complex algorithms are increasingly in charge of choices, operations, and decisions that have previously been left to human actors. They are used across a range of domains, from interpreting or predicting behaviour (Hildebrandt, 2008) to supporting and carrying out operations on behalf of their users (Kim et al., 2014). Algorithms drive recommendation and filtering systems that

curate personalized content (Barnet, 2009), support predictive policing systems (Zarsky, 2016), and identify—increasingly better than human observers—relationships and patterns across vast and distributed datasets. This is instrumental in a variety of scenarios, ranging from online shopping and equity trading to recommending medical diagnoses and treatments to physicians (Floridi, 2012).

A growing public unease about the massive societal ramifications of algorithms has perpetuated a public discourse that is increasingly concerned with their transparency and accountability, and has recently resulted in significant political interventions (Mittelstadt 2016; Pasquale, 2010; Tutt, 2016). In the broadest sense of the term, accountability refers to the requirement of a timely disclosure and justification of one's decisions or judgements to others (Hoos, Pruijssers, and Lander, 2017). More specifically, it "involves being answerable to an evaluative audience for performing up to the prescribed standards that are relevant to fulfilling obligations, duties, expectations and other charges" (DeZoort and Harrison, 2016, p. 3). In this sense, accountability encompasses both virtues and mechanisms (Bovens, 2010) that govern the relationship between an actor and a forum, which pressures to explain and justify conduct, formulates specific questions, and ultimately passes judgement (Bovens, 2007). The account-holding capacity of an "evaluative audience" and the appropriateness of "prescribed standards" hinges on somewhat informed stakeholders as well as more or less stable expectations and practices.

In light of the critical opacity and fluidity of many algorithmic systems, this is hardly attainable and has led to an ongoing effort to find general rules for the transparency of algorithms. As a result, calls for transparency have themselves become objects of criticism. Crawford (2016, p. 11), for example, calls them "disappointingly limited" and "doomed to fail" for two reasons:

First, many algorithms are the property of corporations, which do not want to lose competitive edge and do not want users to game their algorithms. Second, merely seeing mathematical operations does not make them meaningful or comprehensible. Why is that the case?

Why are Algorithms Opaque?

While the strong public concern around algorithmic accountability is still a fairly recent development, this phenomenon must be placed and understood in a longer tradition of problematizing software in general (Pasquale, 2015, Passig, 2017). It appears that the list of reasons for the opacity of software in general and algorithms in particular is not only long but also strikingly old: "[...] an algorithm", as David Marr writes in the late 1970s, "is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied.[...] Trying to understand it by reducing actions to lines of code would be [...] like trying to understand bird flight by studying only feathers" (Marr, 1982, p. 27). Like any technology in action, software and its algorithms have to be considered as emergent phenomena. This holds especially true for machine learning algorithms, which are in large part shaped by the training data they use, but also for digital data in general, as "Data have no value or meaning in isolation. They can be assets or liabilities or both. They exist within a knowledge infrastructure—an ecology of people, practices, technologies, institutions, material objects, and relationships. All parts of the infrastructure are in flux [...]" (Borgman, 2015, p. 27). Opacity is thus not only a result of technical complexity, but also of the fact that, in practice, these technologies are not simply reducible to its parts.

Much of the recent research on algorithms generalizes any kind of opacity to one finding: Many algorithms are and will remain opaque (Pasquale, 2015, Stalder, 2016). With that point of departure, researchers develop strategies for dealing with this opacity. The most frequent strategies are varieties of *reverse engineering* that study the subject by observing inputs and outputs rather than trying to understand flight by studying feathers. However, in scrutinizing those accounts, one can find quite different arguments as to why black boxes will or must remain closed or can theoretically be opened. From a business ethics perspective, a review of those arguments helps reveal that opacity is not always inevitable but rather a question of public discourse, technical development, and corporate strategy.

This argumentative field of objections may be summarized as addressing what Nissenbaum calls the "transparency paradox" (Nissenbaum, 2011, p. 36): More data does not mean more information. Quite on the contrary, the amount of data can become so large that the sheer mass of 'transparent information' produces intransparency. This means that transparency is not necessarily achieved through visibility. For example, every word in a standard form contract is perfectly readable, and nothing blocks users from reading it. Rather, it is length and complexity of the text itself that may foster opacity, and—most importantly—transparency and opacity cannot be understood as essential qualities of a given object. Rather, they have to be considered results of a practical construction and negotiation process, that every organization producing and publishing such objects is necessarily involved in.

Software Has Always Been Opaque

The opacity argument can be traced back at least as far as early research in cybernetics (cf. Passig, 2017): In his article "Some Moral and Technical Consequences of Automation",

Norbert Wiener states that "machines [...] develop unforeseen strategies at rates that baffle their programmers" (Wiener, 1960, p. 1355). Drawing on the example of checkers game-playing machines, he argues that although scientists might understand the inner logics of these machines, it would take so long to understand their "mode of performance" that "criticism may be ineffective until long after it is relevant" (ibid.).

Moreover, this opacity is a matter of not only time but also human cognition: huge sets of code and rules are extremely hard to inspect visually, especially when predictions involve complex combinations of probabilities (Van Otterlo, 2013). The difficulty is due not only to the sheer mass of code or pace of action but also to the code structure: poorly structured code remains intransparent, regardless of its accessibility (Mittelstadt, 2016, p. 4994). In a similar vein, computer scientist and sociologist Jenna Burrell outlines the field of "technical illiteracy", highlighting that writing and reading code is a specialized skill inaccessible to most people (Burrell, 2016, p. 4).

Additionally, many algorithms are fluid and ever-changing. As a famous article by Minsky (1967) argues early on (see Passig, 2017), many kinds of software are designed by larger teams with changing members. That means that an algorithm may gain in size to the point that no single person can perform oversight: "When a program grows in power by an evolution of partially-understood patches and fixes, the programmer begins to lose track of internal details and can no longer predict what will happen—and begins to hope instead of know, watching the program as though it were an individual of unpredictable behavior" (Minsky, 1967, np). Instead of trying to understand code, one can in many cases only test it. This limitation is caused by the ecology of software projects as well as common organizational work practices.

Of course, research and education on structured and understandable programming is developing and ever improving—in fact it is one of the most important parts of any programming training. Still, in practice, coding remains messy, large parts consist of third-party code from code libraries, frameworks and the like are copied and pasted from other projects (Kim et al., 2004) or cloned, which is in many cases "the only way to achieve a certain program behaviour" (Beller et al. 2017, p. 3). Programming is always a constant negotiation between well-developed principles for building algorithms (and software) on the one hand and their fluid, open-ended and ever-changing practice on the on the other hand. While programming standards are important and effectual, taking the open-endedness of digital objects seriously means that accountability cannot simply be delegated to given norms in software engineering. They constantly have to be re-negotiated and re-established in organizational practice. The copying and cloning techniques however, add to the strategic motivations to keep algorithms intransparent, as corporations veil plagiarized code. Further, even well-engineered algorithms can result in unexplained outcomes or errors, either because they contain bugs or because the conditions of their use change, thus invalidating assumptions on which the original design was based (ACM, 2017).

Opacity is thus not a novel ethical issue of algorithms. Rather, it appears to be a defining characteristic that has become stronger with the advancement of computing technologies that generally exceed the comprehension of single human beings. These issues, however, spawned their own solutions, such as visualization software that provides overviews by translating code into visual maps (or public discourse around the logics of certain kinds of algorithms). In other words, opacity can be reduced with the help of mediators, i.e., software or human collectives dividing and translating the interpretation work into digestible pieces. It cannot be considered

a *black box* that is principally unopenable. Rather, it requires a collective effort of socio-technical collectives.

Self-learning Algorithms as Fundamental Obstacles?

However, the roles and responsibilities in such collective efforts are obfuscated if this kind of *quantitative opacity* is not differentiated from somewhat more complicated cases. The last one or two decades have brought about new kinds of algorithms, *self-learning algorithms*, which pose accountability challenges. *Self-learning algorithms* are a set of rules defined not by programmers but by algorithmically produced rules of learning: "The internal decision logic of the algorithm is altered as it 'learns' on training data" (Burrell, 2016, p. 5). In other words: "Algorithms are used to program new algorithms" (Stalder, 2016, p. 178). As a result, they can be assessed only experimentally and not logically (cf. ibid., p. 179)—essentially a reiteration of Minsky's (1967) early observation.

That being the case, there are instances where the *practicality* of algorithmic opacity must be taken for granted. Machine learning systems, different from older rule-based algorithms, are often not conducive to or designed with human understanding in mind (Edwards and Veale, 2017). There are a number of reasons for this, chief among them being that corporations keep them secret for strategic reasons and that we do not (yet) have the socio-technical means to make them comprehensible for human collectives. However, the argumentative move to accept the *principle* of opacity as something that must be taken for granted seems unconvincing. There is no apparent reason to treat machine learning algorithms differently from their predecessors in terms of ethics: The presumed fact that we do not yet have the technologies to translate their actions into a humanly comprehensible language is not an excuse to diminish the accountability

duties of their owners or users. For example, if most recent algorithms were treated like research objects of molecular genetics because of their alleged fundamental opacity, as Passig (2017) postulates, this would discharge corporations from their liability to ensure accountability: A molecule has no obligation to make its functioning accountable. Corporations, however, do have accountability obligations for algorithms, regardless of current issues with their comprehensibility. As Martin (2018) argues, if the argument "too complicated to explain" would simply suffice, organizations would be incentivised to produce complicated systems precisely to avoid accountability.

Accountability Mechanisms for Algorithms

The scholarly discussions on algorithmic accountability gather around two main questions. The first is an epistemic question about the positivist notions of transparency. Ananny and Crawford (2016) argue that transparency has to be uncoupled from accountability, as it implies false expectations: "[…] making one part of an algorithmic system visible […] is not the same as holding the same assemblage accountable" (ibid., p. 12). Here, accountability serves as a better suited substitute for the ideal of transparency: As calls for transparency in this sense are not constructive, the focus of debate has to shift towards the accomplishment of accountability through concrete mechanisms.

The second strand of publications discusses how mechanisms for this kind of accountability can be developed. An accountability mechanism, in general, is an institutional arrangement (of a social, political, or administrative nature) in which an organization or person can be held accountable by a forum (Bovens, 2007). Such institutional arrangements may govern the relationship between and behaviour of the involved organization and the forum. Common

mechanisms for accountability are parliamentary hearings (Busuioc, 2013), performance reporting (Van de Walle and Cornelissen, 2014), or watchdog journalism (Norris, 2014).

Among the first and most prominent scholarly positions to address this issue for algorithms was Diakopoulos' (2013, 2015) work on several methods that help journalists uncover algorithmic agency (ibid.). Here, journalists produce accountability by exerting their "watchdog" function (Norris, 2014), especially by attempting to reverse-engineer the functioning of inaccessible algorithms.

Other authors have proposed "algorithm audits", an umbrella term for various methods for researching algorithms. Specific to these methods is that some function not via reverse engineering but via trusted third parties, such as researchers, that are provided access to otherwise secret code (Mittelstadt, 2016; Sandvig et al., 2014a; Sandvig et al., 2014b). The accountability mechanism here is thus not an "audit institution" (Posner and Shahan, 2014) but auditors representing neither commercial nor governmental interests. Other suggestions correspond more with the notion of audit institutions and propose regulatory agencies: "[…] certain classes of new algorithms should not be permitted to be distributed or sold without approval from a government agency designed along the lines of the FDA" (Tutt, 2016, p. 83).

On the one hand, these mechanisms address primarily algorithm owners and creators as actors who must be 'watch-dogged': Rather than asking what algorithm owners and creators can do to accomplish accountability, these approaches focus on how these owners and creators can be made accountable by others. On the other hand, these mechanisms do not provide normative standards. During the last year, several organizations have started to fill that gap.

The Association for Computing Machinery (ACM), for example, outlines seven principles for algorithmic accountability that are addressed at owners and producers of algorithms: "Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results" (ACM, 2017, p. 2). With these seven principles—which range from the *awareness* that organizations should create concerning their algorithms' agency to the *validation and testing* that they are supposed to perform on their technologies—the ACM outlines the responsibility that organizations should take, regardless of how opaque their algorithms may be.

In a similar—but less prescriptive—vein, the Utrecht Data School has developed the *Data Ethics Decision Aid* (DEDA). In contrast to the ACM's seven principles, DEDA names steps in a process of data production and use. Each step poses several ethical questions, ranging from "can you communicate how the algorithm works" to "how do you inform people that the data are used" (Schäfer and Franzke, 2017). That means that instead of positing explicit and general norms for responsibility, DEDA provides a process of raising ethical questions. These questions are, of course, implicitly normative. Nevertheless, they appeal to the ethos that an organization has and wants to maintain instead of postulating norms. Similar to ACM's principles, this also points to a public discourse that expects organizations to develop and manage an ethos of data and algorithm use.

An initiative by university and industry researchers—among them, Solon Barocas and Nicholas Diakopoulos—adds to the principles of Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) two further dimensions: First, in addition to expected principles such as

fairness and responsibility, they call for "Social Impact Statements" (Diakopoulos et al., 2018). That is, in the design, pre-launch and post-launch stages, algorithm creators should develop statements on the aforementioned criteria that are to be published when the system is launched (ibid.). The second dimension they add is especially relevant from a reputation perspective, as its consequences change the definition of reputation management in the context of algorithms. This dimension is explainable, accurate and audible algorithms (Diakopoulos et al., 2018). That means that instead of addressing how organizations should manage their existing algorithms towards a public, they focus on how these algorithms are coded. That implies that rather than wondering about how to handle opaque masses of code, organizations should make sure their technologies do not become opaque in the first place. The result of these principles is that part of algorithmic accountability would be more accountable coding practices, or from a reputation management perspective, coding itself would be an essential element of reputation building.

3. The Role of Reputational Concerns and Engagement Strategies in Algorithmic Accountability

Reputational Concerns and Accountability Relationships

Any accountability process is inherently linked with reputational concerns. Broadly speaking, reputational concerns refer to an organization's concerns with external expectations and the impact of its conduct on the perception of stakeholders (Suurmond, Swank, and Visser, 2004). Reputational concerns thus relate not to specific threats to reputation (cf. Coombs, 2013; Karpoff, 2012) but rather to the general realm of external expectations that organizations relate to in building and managing their reputation and legitimacy, thus securing and expanding their room to manoeuvre and their ability to acquire resources (Bernaz, 2013; Carmona, Donoso, and

Reckers, 2013; Eisenegger and Imhof, 2008). While legitimacy in this context is widely viewed as the generalized assumption that the current conduct of an organization is appropriate or acceptable, reputation is commonly defined as the prospect of an organization's future conduct, based on direct or mediated perceptions of its past and present behaviour (Deephouse and Carter, 2005; Rindova, Pollock, and Hayward, 2006). Hence, the constructs of reputation and legitimacy are similar in that they are based on judgements made by observers about an organization, and they have similar constitutive processes and antecedents; however, reputation can be distinguished from legitimacy by its stronger future orientation and its reliance on relative assessments, i.e., to other organizations, as a comparative mark (King and Whetten, 2008). In this sense, reputation 'encompasses' legitimacy: "first organizations meet standards in order to then be judged in relation to each other" (Bartlett, Pallas, and Frostenson, 2013, p. 531).

Current accountability research highlights that such expectations about future conduct and comparative judgements help account, in large part, for the interest in and intensity of accountability processes (e.g., Bovens, 2007; De Cremer and Barker, 2003; Dubnick and Frederickson, 2010). Accordingly, in this literature, reputational concerns are widely depicted as the central mediator in the relationship between account-givers and account-holders. More specifically, reputational concerns are said to "drive the way in which account-givers and account-holders relate to each other" as these concerns act as a filtering mechanism for external expectations (Busuioc and Lodge, 2017, p. 91). Hence, the enthusiasm (or reluctance) within organizations to account for their actions can be explained by their assumptions about their core reputation and the reputational implications of their conduct (Gilad, Maor, and Bloom, 2015).

In journalism, for instance, complaints about the use of poorly comprehensible algorithms in the creation of automated content (cf. Dörr and Hollnbuchner, 2017, Montal and Reich, 2017) get at the heart of the reputation of media organizations, as the general acknowledgment of the worthiness of these organizations relies on delivering reliable and trusted information (Bachmann, 2017). In this context, concerns about opaque algorithms become an opportunity for the account-giver to demonstrate their high journalistic standards. In practice, content-producing systems have been purposefully based on rather simple and straight-forward algorithms so that actions and decisions remain easily understandable and the "trusted journalist" remains the focal account-giver rather than delegating agency to a complex system (Fanta, 2017, p. 10). In these cases, the decisions are mediated by reputational concerns to the degree that they involve the anticipation of the audience's reaction to the purposefully limited use of algorithms in the production of news. Such decisions are much more likely to occur in media organizations, whose reputation builds upon the delivery of timely and reliable information (e.g. financial news) than for example in boulevard journalism.

Mapping Reputational Concerns About Algorithms

Concerns with external expectations regarding algorithmic practices are central not only for the creators of algorithms but also for the rapidly growing number of organizations that employ them. As more and more people interact on a constant basis with algorithms, the public perception of organizations increasingly depends upon them. Algorithms not only represent the organization that owns them and shape user experiences but also affect the reputations of organizations that rely on third-party algorithms as part of their value chain. As many organizations interconnect with the influential algorithms of Amazon, Google, Facebook and the like, their reputations also partly depend upon the algorithmic activities and accountability of these large players.

Based on recent discussions (cf. Mittelstadt et al., 2016 for an overview), three sets of concerns with algorithmic practices can be distinguished: evidence, outcome, and opacity concerns. First, algorithms pose *evidence concerns* because they may give inconclusive evidence by producing probable (not certain) outcomes, give inscrutable evidence when knowledge about input data and their use is limited, and give misguided evidence when the algorithms' conclusions rely on inadequate inputs (Mittelstadt et al., 2016). Such concerns are highly apparent in, e.g., patient assessment systems (PAS), which are used to project the success of medical treatment for fatal conditions and predict patients' deaths. Proprietary PAS (as offered by, e.g., *Aspire Health* or *23andMe*) use information on medical treatments, diagnoses of particular patients, and comparative patterns of common therapies. They neglect, however, central individual factors of personality and psyche, such as a patient's will to survive, that have been proven critical in treatment success. As doctors themselves often do not know the information basis and estimation procedures of proprietary PAS, these systems have recently become a topic of strong public concern (Beck, 2016).

Second, algorithms pose *outcome concerns* as they may produce unfair outcomes that are, e.g., discriminatory to a certain group of people or lead to secondary (i.e., transformative) effects when they change the way people perceive situations, as the case with profiling algorithms (Hildebrandt, 2008; Mittelstadt et al., 2016; Pasquale, 2015). Such outcome concerns are highly apparent, in the case of automated journalism mentioned above: Many news agencies use news robots to produce financial news (Fanta, 2017). Stock market data are automatically translated into text; this works precisely because they do not need human editors controlling them (ibid.). Any error in such outputs would obviously raise immense issues for the respective news agency.

Both evidence and outcome concerns can be considered *contingent concerns* as they are common but not necessarily linked to complex decision-making systems. However, the third set, *opacity concerns*, are qualitatively different in this regard. They are inherently *obtrusive* as all complex decision-making systems necessarily remain—at least in part—opaque. They make it excessively difficult, and in some cases impossible, to actually detect problems and identify causes. While issues related to evidence and outcomes may or may not arise, opacity concerns are inevitably tied to the technology and will sooner or later obtrude in accountability relationships between organizations and their stakeholders.

Algorithmic Accountability and Engagement Strategies

While reputational concerns filter external expectations and can help explain the varying degrees of interest in and intensity of accountability processes, the actual *quality* of this process is shaped by the specific way in which the organization then engages with the emergent external demands of its stakeholders (Greenwood, Raynard, Kodeih, Micelotta, and Lounsbury, 2011). Accordingly, engagement—as a practice undertaken by organizations to involve stakeholders—has been described as a way to achieve accountability (Gray, 2002; Van Buren, 2001). Moreover, in the case of algorithms, where external demands are often unclear and/or no clear-cut accountability standards are available, organizations need to engage with their various stakeholders to create such standards (Suchman, 1995). However, more engagement does not automatically mean more accountability; while some engagement practices may indeed be focused on listening and learning (Romenti 2010), others may aim mostly at creating an image of accountability (Swift, 2001), or even be outright deceptive (Greenwood, 2007). This raises the question: When reputational concerns arise, and no clear-cut accountability

mechanisms are in place, what general strategic options do organizations have to engage with the emergent external expectations? Following Scherer, Palazzo, and Seidl (2013), we argue that organizations have three fundamental strategic options: they can a) strategically manipulate expectations, b) adapt and conform to extant expectations in their environment, or c) engage in open public debate and reasoning over what *should* be expected. The *manipulative approach* describes the active attempt to shape and influence external expectations, e.g., through lobbying, public relations campaigns and other strategic communication instruments. This approach is guided not by adherence to external demands or institutional rights to information but rather by the solicitation of stakeholder views in a reputational contest for the sake of reputation, thus leading to 'soft accountability' (Owen et al., 2000; Swift, 2001).

The *adaptive approach* describes isomorphic behaviour aimed to conform with extant expectations through meeting the demands of powerful stakeholders or complying with established standards, e.g., leading to practices of reporting or performance review. Through a reputational lens, this emphasizes an outside-in approach beyond mere influence, where stakeholder partnerships facilitate organizational learning and the adjustment of main reputation drivers (Romenti, 2010). Of course, for this approach to work, external expectations have to be rather clear-cut and stable.

Finally, the *moral approach* builds on ethical principles that allow for open discourse between the organization and its stakeholders and free exchange of arguments that can lead to common and consensual outcomes in terms of what should be expected. As such, a reputational contest that builds on a moral approach helps facilitate a) legitimate outcomes under conditions of unclear external demands (Mingers and Walsham, 2010) and b) opportunities for competitive

advantage under conditions of ever-changing and fluid technology, where knowledge about its workings and ramifications does not reside exclusively within the organization but has to emerge from an open deliberation with actors in the organization's environment who are affected by it (Lubit, 2001).

For organizations, these three fundamental strategic options constitute parallel approaches rather than mutually exclusive strategies, and depending on the particular challenge at hand, they can be enacted simultaneously (Scherer et al., 2013).

4. Managing Reputational Concerns and Accountability Relationships: A Discourse-Ethical Approach for Opaque Algorithms

The growing public concern about algorithms on the one hand and the technical and strategic challenges for algorithmic accountability on the other hand represent a classical issue of discourse in democratic societies: the central ideal that issues of public concern can be addressed in open and critical discourse, which demands that actors give reasons for their actions. However, for opaque algorithms, there is no straightforward way to deliver accounts at any one point in time. Thus, there is both a pragmatic necessity and a normative obligation for organizations to take part in fora that allow for an inclusive debate on the workings and ramifications of the algorithms that they employ. The technical and strategic challenges for algorithmic accountability in particular point towards the necessity to structure such public debate not only within particular accountability mechanisms but, more fundamentally, based on ethical principles that can guide actors towards an open and rational process of debating accountability issues.

For this, we suggest following a discourse-ethical approach. In their most widely used form, discourse-ethical approaches draw on Habermas' (1999) work on discourse about competing validity claims, in which participants consider each other's arguments, give reasons for their position, and are ultimately willing to reassess and, if necessary, revise their original position. Such discourse leads to a deeper understanding of the problems, positions and concerns of the various actors as well as a greater mutual acceptance of all parties involved and the common (ideally consensual) decisions. However, the possibility of such positive outcomes hinges on the adherence to normative principles when debating the acceptance or rejection of particular validity claims, such as the principle of open and equal access to forums of discussion, the availability and transparency of information, and equal opportunities for all to introduce arguments into the debate. These communicative principles are to ensure that discourses are uncorrupted by power differences or strategic motivations (see, e.g., Niemi, 2008 for a concise summary of the approach).

In the context of business ethics, discourse-ethical approaches are widely used to apply these principles to analyse discursive settings and processes that are meant to, e.g., resolve intercultural business conflicts (French et al., 2001) or build corporate legitimacy (Palazzo and Scherer, 2006; Seele and Lock, 2015). Moreover, the approach has been applied as an ethical framework for corporate accountability (Gilbert and Rasche, 2007; Rasche and Esser, 2006).

In the context of algorithms, we suggest the discursive approach as a pathway to manage accountability relationships that are burdened by algorithmic opacity and fluidity. When organizations' poorly transparent and highly fluid algorithmic practices become the object of reputational concerns, these organizations cannot hope to merely "deliver" accounts. They need

to be prepared to participate in a discursive process together with their stakeholders in order to work towards good practices of account-giving and account-holding. As discourse principles place a strong emphasis on involving those affected by decisions, they are well suited for algorithmic accountability. Stakeholders need to be an active part of detecting and assessing the potential shortcomings of algorithms, as particular developers and applicants of algorithms do not necessarily hold a privileged position in assessing these issues. Accordingly, discourse principles are key to addressing accountability in the context of highly fluid and constantly evolving information systems (Mingers and Walsham, 2010). If simple accounts cannot be given by one party, the emphasis shifts to an inclusive communication *process* through which a *continuous and tentative assessment* of the development, workings, and consequences of algorithms can be achieved over time.

The discourse-ethical approach suggests that public debate about algorithms will allow actors to collectively mitigate the black box challenges they pose. However, the rational potential of discourses on competing validity claims can be harnessed only if basic ethical principles are met. Various approaches have been suggested based on Habermas' theory in order to arrive at a comprehensive but manageable set of dimensions to assess whether factual discourses measure up to the ideal (Steenbergen, Bachtiger, Sporndli, and Steiner, 2003). Below, we adhere closely to a set of principles developed by Nanz and Steffek (2005) in the context of international governance. Their approach fits well in this context because the principles are developed with a particular sensitivity for discourses involving highly complex and complicated issues. They also address the challenge to include citizens' concerns and highlight the importance of civil society organizations as central actors to make complicated technical issues accessible and understandable to a wider public.

To illustrate how the discourse principles allow to practically assess the extent to which engagement may harness the rational potential of discourse, we complement the following discussion of each principle with a case example (see boxes 1 to 5).

Box 1: The Case of Criminal Justice Algorithms
Criminal justice algorithms (CJAs) are tools that aim to predict future behavior of defendants and incarcerated persons. CJAs use information on, e.g., family background, socioeconomic status, employment status, and neighborhood crime to provide a prediction of an individual's criminal risk. In the US, CJAs are used in most states to determine sentences, set bail or even contribute to determinations of guilt or innocence (Kehl et al., 2017). CJAs are mostly proprietary and their workings largely hidden from public scrutiny. The processes and validity of CJAs is rarely examined, and if they are, the investigations are mostly carried out by the same party that developed the tools (Desmarais and Singh, 2013). This has led to strong controversy about the outcomes of CJAs and the ways in which they gather and process information (Scurich and Monahan, 2016). Such criticism is not propelled through public media but there are multiple challenges of the validity of CJA results in juridical proceedings. While the accountability issues related to CJAs raise reputational concerns regarding in-puts, outputs, as well as opacity, it is especially the opacity concerns that have the potential to undermine the public's faith in the criminal justice system (EPIC, 2017).

Principle 1: Access to Deliberation (Participation)

The intricate issues around algorithmic accountability need to be discussed in open fora, where every subject with the competence to speak and act is allowed to take part in debate. Specifically, all those who potentially suffer 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 and, if necessary, leads to broadly acceptable decisions. Stakeholders need to have institutionalized access to such deliberative settings so that they have a chance to voice their concerns, opinions, and arguments.

Possible deliberative fora for algorithm accountability could be connected to public code repositories that allow for version control or wiki-based documentation and commentary/debate. For instance, a number of news organizations, such as *BuzzFeed*, maintain repositories in which data and code used for data-driven articles are at least partially published (Diakopoulos and Koliska, 2017, p. 810). The limited functionality of published code in the context of machine learning received early criticism, which led to the development of machine learning repositories. Perhaps the most popular (and oldest) among them is the *UCI Machine Learning Repository*, which has benchmark datasets used to audit machine learning algorithms.[1] Today, media outlets such as *The New York Times* upload datasets they use to feed their machine learning algorithms to *GitHub*.[2] Both the case of *Buzzfeed* but also that of *The New York Times* serve as examples of algorithmic accountability management *as* reputation management. Both can be considered attempts to either reproduce (in the case of the *NYT*) or produce (in the case of *BF*) a reputation

[1] http://archive.ics.uci.edu/ml/index.php

[2] See, for example, their documentation of a machine learning model for their recipe database: https://open.nytimes.com/our-tagged-ingredients-data-is-now-on-github-f96e42abaa1c

of a news organization that applies highest journalistic standards not only to the humans working for them but also for their algorithms (and other non-human journalistic actors). In both cases, the actors go beyond mere trouble shooting and their activities on machine learning repositories represent a forward-looking form of reputation building as journalistic institutions.

Furthermore, as with other issues of public concern, the debate on algorithmic accountability can occur via public media platforms. While such discourses tend to be more inclusive than, e.g., conversations on forums in code repositories (which means better involvement of laypersons), the discursive quality relies strongly on the independence of the media system ('watchdog capacity') and its foundation in a responsive civil society (Habermas, 2006). Furthermore, possible algorithm issues must have a considerable magnitude to draw the attention of public media.

Finally, as regulatory bodies for algorithms have been suggested (Tutt, 2016), new fora for debate on algorithmic accountability can also emerge from political institutionalization: If regulators required organizations to disclose information on fundamentally opaque systems, these organizations would instead need to facilitate accessible fora where a debate about processes and ramifications of their algorithms can develop. If account-giving was simply rejected due to opacity, organizations would risk potentially arbitrary regulation.

Box 2: Participation in the Case of CJAs
The access to deliberation about CJAs is mainly happening on two dimensions. First, CJAs have become a public issue. In such times of strong public attention, the accountability discourse profited to some degree from a rather broad attention

of quality media (see, e.g., Smith 2016). Second, as CJAs are increasingly part of juridical procedure, accountability issues are tied directly to the formal and institutionalized process of legal reasoning and deliberation in applied law (Wojciechowski, 2010). As such, concerns with CJAs have commonly good access to legal deliberation. However, in many cases jurisdictions have taken decisions to adopt CJAs without prior attempt at debating their use and ramifications (EPIC, 2017). As a result, access to deliberation about the use and consequences of CJAs in many cases only started after critical cases occurred.

Principle 2: Access to Information (Comprehension)

All those that take part in the deliberative process need to have full information about the issues at stake, the various suggestions for their solution and the ramifications of these suggestions. This principle points directly to the 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)" (Burrell, 2016, p. 5).

Where the mere provision of the needed information does not allow for straightforward comprehension or is technically impossible—if it were possible, a discursive process would not be needed to jointly work out accountability issues—the emphasis is on the free exchange of arguments via inclusive platforms. However, there are some ways to increase comprehensible information on principally inconceivable algorithms, e.g., through "experiment databases" that enable comparisons between algorithms (van Otterlo 2013, p. 17) or methods of simplifying machine learning models by visually translating their actions to humans (Burrell, 2016, p. 9).

Another pathway to increased access to information is via *reverse engineering*, i.e., the approaches applied to produce transparency in a system without disclosing its inner workings. Through the observation of the inputs and outputs of a given system, a model is developed that explains how this system works (for an overview of the diverse field of scholarly literature on reverse engineering of algorithms, see Diakopoulos, 2015). For example, journalists apply several methods to understand how algorithms, such as those of the iPhone's autocorrection, work or to understand price discrimination in online commerce (cf. Diakopoulos, 2015). In the case of the iPhone, for example, the provided API was used in combination with model user data to reverse-engineer a conglomerate of algorithms responsible for the autocorrection, which turned out to have biases towards certain words. Methods of reverse engineering algorithms are already in journalistic practice, and they will become increasingly sophisticated in the future.

Further, information can be made accessible by *expert third parties* trusted by both organizations and the public if they are granted exclusive access to the algorithms in order to scrutinise them without disclosing their details (Pasquale, 2010). This approach has been deemed unlikely to work given the complexity and observer or user relativity (Sandvig et al., 2014b, p. 10). However, it remains unclear how this objection disqualifies the general advantages that the privileged access of third-party experts would provide.

Finally, as an alternative to reverse engineering a whole system, there are approaches available for generating information by focusing on actual use scenarios (cf. Sandvig et al. 2014a): *Algorithm audits* propose a sophisticated set of methods that simulate or follow actual algorithm users in order to determine how much algorithms discriminate in these realistic use cases. A

possible implementation scenario for these are subject-centric explanations that explain to a specific individual why a particular interaction with an algorithm yielded a particular result in his or her case. This involves explanations that employ sensitivity (what changes in an individual's input data will yield a different outcome), similarity (which other categories of individuals receive similar outcomes), and performance metrics (how sure the algorithm is about its classification).

A parallel issue to technical opacity in information access is the potential strategic motivations for not releasing information. For example, as proprietary entities, many algorithms are not subject to open government laws, which can make it very difficult for the public to access information. In these instances, it is often the algorithm developers that produce and publish their own validation studies with questionable informational value (e.g., EPIC, 2017). Furthermore, proprietors tend to focus these reports on output validity and do not publish the essential information on *how* calculations are made. To address this issue, some NGOs (such as EPIC or ProPublica) or state agencies (cf. Hunt and Dumville, 2016) have started to produce their own research in which they test and provide information about algorithms. As mentioned above, access is not only a question of how a given algorithm can be made visible in its functioning, but also of how clear code is structured and organized. Clarity of code thus becomes an increasingly important field for organizations to provide accountability.

Box 3: Comprehension in the Case of CJAs
As many other algorithms, CJAs pose a fundamental issue to the principle of information transparency and comprehension. As CJAs are proprietary they are not subject to open government law. This makes it very difficult for the public to

access information. While some algorithm developers produce and publish their own validation studies, the value of this information is questionable. Furthermore, proprietors focus these reports on output validity and do not publish the essential information on *how* calculations are made. In response to this issue with information access, some have submitted requests to state agencies to release source code through state freedom of information requests (EPIC, 2017).

Principle 3: Inclusion of All Arguments (Multivocality)

In addition to the inclusion of informed stakeholders (participation and comprehension), the inclusion of all arguments is a key principle to enable rational discourse and deliberation. Participants need to have the opportunity to see an issue from all relevant points of view. All those possibly affected should have a chance to voice their concerns.

This is a critical aspect especially in relation to people's often limited ability to comprehend the technicalities of algorithms and thus their limited means to formulate and present concerns. While public code repositories, for instance, may allow open access (participation) and provide extensive code (information), the actual discourses via such platforms are limited to experts that can make sense of the information that is made available and use it in argumentation. Further, some essential arguments may be excluded because stakeholders that could, in principle, participate and be informed, are simply unaware that and how they are affected by the respective processes and outcomes. As noted by Nanz and Steffek (2005), it is in principle challenging to account for essential arguments that are excluded from a debate.

Here, it is important to differentiate between different forms of ignorance regarding algorithmic action, or rather, different forms of unknowing. When observing research processes with digital methods software *Gephi*, Author and colleagues (2017 - blinded) discuss how researchers and the software itself make their unknowns accountable in a way that quarantines their unknown algorithms' agency from research results. More important than opening algorithmic black boxes (that remain opaque in practice anyway) is to turn *unknown unknowns* into *known unknowns*. To include all arguments, it is essential that unknowns be rendered as such. As a consequence, more fundamental than making algorithms transparent is creating awareness of their opacity. This is central not only from a normative point of view but also reputation-wise: If a public discourse becomes aware that proprietors of algorithms have made efforts to keep these unknowns unknown, this might yield remarkable reputational ramifications.

Box 4: Multivocality in the Case of CJAs
Most of the available arguments on CJA accountability issues are provided by the proprietors, defense advocates, and NGOs active in the fields of information technology and justice. A look into the recent debate in the press shows that most journalists also draw on these three groups as main sources. To a smaller degree, also academic voices are included. Thus, arguments seem fairly concentrated around a few specialized actors. Moreover, defense advocates tend to criticize that they are actually unable to challenge the validity of the results at sentencing hearings, which excludes their potential arguments from debate (EPIC, 2017).

Principle 4: Impact on Recommendations and Decisions (Responsiveness)

While participation (1) and access to information (2) are preconditions for a process of deliberation to take place, the inclusion of all arguments (3) is the main precondition of the rationality of that process. However, all three principles are meaningless if the different concerns and suggestions regarding algorithmic systems that are put forth by various stakeholders are not adequately taken up in the debate and cannot influence actual recommendations or decision as a result of the discourse (Nanz and Steffek, 2005). Thus, the process needs to be clearly *responsive* to these suggestions and concerns.

Here, in principle, developers and proprietors of algorithms may a) remain unresponsive, b) stall, c) allow for change according to discursive outcomes or d) install secondary measures. While practices of non-response and stalling are common and change is often achieved only through legal intervention—not discourse itself—some cases of secondary measures can be observed. One such case is the use of news bots in automated journalism (cf. Carlson, 2015, Dörr and Hollnbuchner, 2017, Montal and Reich, 2017). High pressure to remain accountable has prompted news agencies that employ content-producing systems to purposefully limit these systems. For instance, in relation to the example laid out above, journalists try to confine the agency of the algorithmic system to a level that they can control by using simple algorithms whose actions remain understandable for non-experts—i.e., the journalists take over accountability themselves rather than implementing opaque systems that may pose reputational concerns (Fanta, 2017, p. 10).

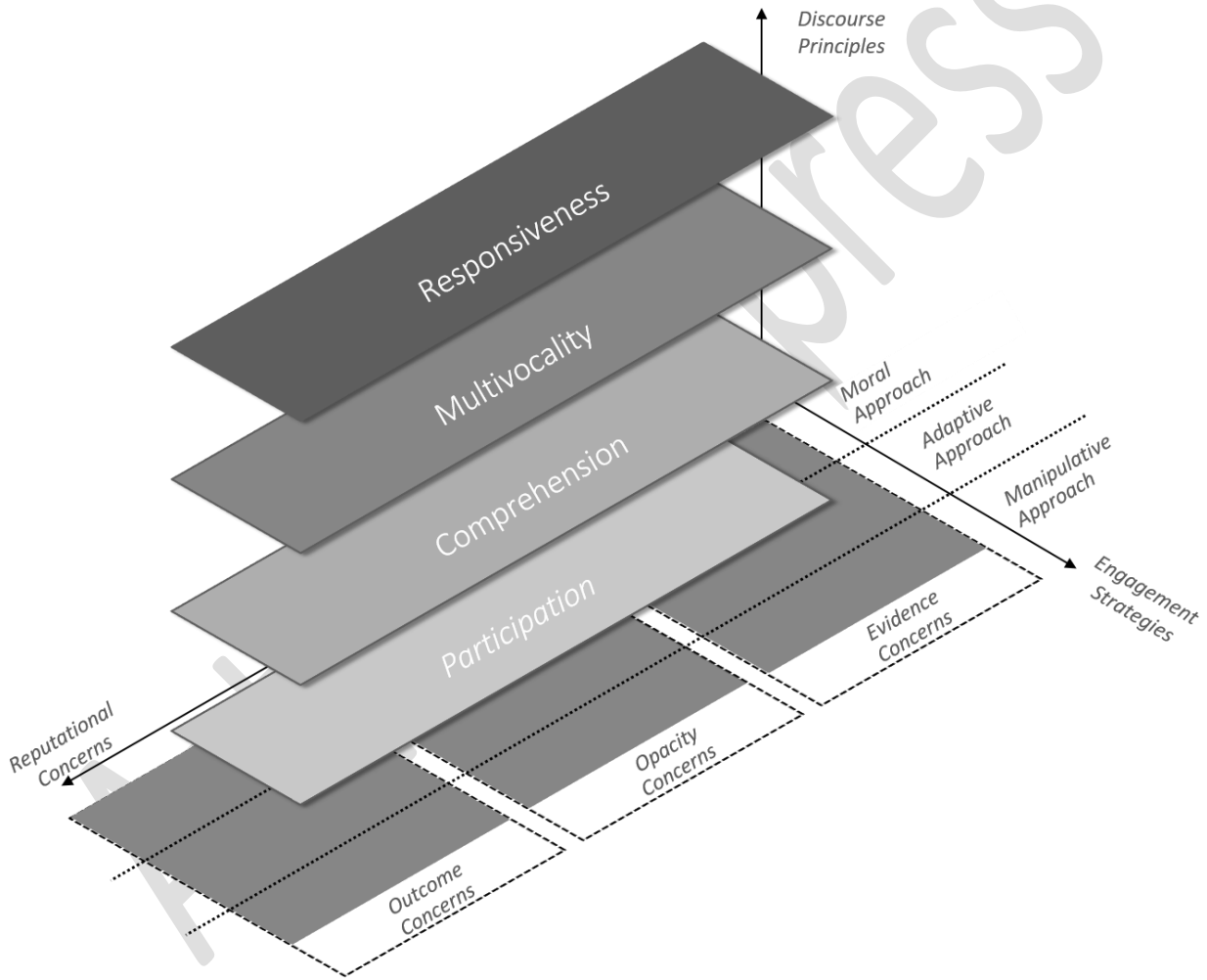
Box 5: Responsiveness in the Case of CJAs
Based on the current literature and debate, it is difficult to say much about responsiveness. However, there is a growing public record of a) contested cases and b) some cases that have been collected where the validity of the CJA output has successfully been rejected (EPIC, 2017). Further, to the degree that CJAs are linked to a formal process of deliberation (in the legal system), the fact that algorithms making risk assessments are proprietary and opaque leads to a *formal problem* in procedure: In such cases the fact that defendants are unable to challenge the validity of algorithm outputs becomes visible as a possible violation of the defendant's own right to due process.

5. Discussion

The above discussion suggests a framework for managing algorithmic accountability that encompasses three core dimensions: reputational concerns, engagement strategies, and discourse principles (cf. Figure 1). Specifically, the framework suggests a) that concerns about the epistemic setup, outcomes, and opacity of algorithms drive accountability processes, b) that the way in which organizations then practically engage with emergent expectations about algorithms may be manipulative, adaptive, or moral, and c) that to the extent to which accountability processes involve opaque and fluid algorithms, organizations and their environments need to engage in a deliberative process to jointly shape expectations about algorithmic accountability—a process that is put into effect by adherence to discourse-ethical principles of participation, comprehension, multivocality, and responsiveness.

When discourse principles are used to detect and describe the extent to which specific cases show potential for resolving accountability issues through discourse, several overarching issues emerge. First, because of the dynamic changes in complex algorithmic systems, the *fostering of access to deliberation* must be supplemented by platforms that allow for a sufficient continuity of debate (and not just for debate at selected time points). The fluidity of algorithms necessitates fluid observation and debate. Rigid certification processes, for instance, would not be able to do justice to the speed at which most complex algorithmic systems change. Recent suggestions for cooperative and procedural audits of algorithms (Mittelstadt 2016; Sandvig et al., 2014a) address this aspect of the continuity of deliberation platforms directly. The same aspect is also increasingly considered for public code repositories that use benchmark datasets to audit dynamic machine learning algorithms. This discussion indicates that deliberative forums for algorithmic accountability are likely to become an important area of contact and interaction between organizations and their environments, thus emerging as a new playing field of reputation management.

Figure 1
A Framework for Managing Algorithmic Accountability



Second, in the face of algorithmic opacity, *access to information* appears as a central obstacle to discursive accountability settings. From the perspective of reputation management, this dimension may pose an area of likely reputational threats for modern organizations that develop or apply algorithmic systems. This, in turn, emphasizes the relevance of prosocial positioning in relation to this issue as a means to create a "buffer" (McDonnell and King, 2013). Beyond such a defensive approach, several approaches are available to help both organizations and stakeholders actively mitigate this problem. While reverse engineering (Diakopoulos, 2015) focuses on algorithms as whole objects that one tries to read, algorithm audits (Sandvig et al., 2014b) take a subject-centred approach. The idea of neutral third parties (Pasquale, 2010) also approaches algorithms as whole objects, but whereas reverse engineering addresses them from an external perspective, third parties can provide internal observations. These different approaches, of course, are not mutually exclusive. Rather, they can be combined in a discursive triangulation. Ideally, reputation-aware corporations optimize their performance on all three dimensions. Furthermore, from a reputational angle, the inclusion of third parties then poses questions about the role of the reputation of neutral third parties and their ability to feed 'trusted information' into the accountability relationship (Swift, 2001).

Third, we should remain aware of challenges to *fostering inclusiveness of arguments.* We established that potentially deliberative formats, such as audits or repositories, may in principle be hosted on open access platforms but in practice still give access mostly to arguments of specialty audiences. This is a twofold problem: Algorithmic harms often arise from the way groups are classified or stigmatized. These groups are not only laypersons to algorithms; they are also often unaware that they are disadvantaged by them. This problem is currently, only for the most severe instances, balanced by a public deliberation supported through platforms of the quality press and watchdog journalism: where concerns about algorithms become the object of

broad public debate, the accountability discourse benefits to some degree from deliberation bolstered by quality media (see, e.g., Garber 2016, Naughton 2016, Smith 2016 for the case of criminal justice algorithms). Specifically, such a high involvement of journalism points at both the access to deliberation as well as the inclusion of diverse arguments.

This, quite generally, highlights that holding algorithms accountable necessitates a responsive civil society that can feed diverse arguments into the debate. As in other contexts involving complex and complicated issues (Nanz and Steffek 2005), empowering agents, such as NGOs, regulators, or civil society organizations, are essential for detecting and reviewing potential algorithmic failures and deliberating the intricate questions of accountability for multiple angles.

Fourth and finally, reputational concerns should in principle *foster responsiveness*. Algorithmic organizations enjoy a certain degree of freedom in providing accounts; however, public pressure might serve as a watchdog not only to engage in discourse but to also ameliorate eventual shortcomings discovered through this process. There is hope that this ongoing public scrutiny will keep algorithmic organizations accountable. Increasing emphasis is put on policy that aligns with the common problems many algorithms currently struggle with. The harms and benefits of algorithms tend to fall along fault lines societies struggle with in general in that markers such as race, class, gender identity and sexual orientation, (dis)ability, language, or geographic location can lead to discrimination. It is conceivable that organizations that have an interest in correcting these wrongs might have an incentive to ensure that their systems are audited and as a consequence potentially better explained or even changed, based on the deliberative norms outlined. If not, organizations scrutinizing civil society may bolster a reputational norm system that can make discursive accountability in their best interest. These

processes are inherently riddled with conflict, but we contend that while some technical challenges to transparency will always exist, they can always be made to be managed more responsibly through accountability processes that are based on principles of discourse and deliberation.

6. Conclusion

Algorithms are no longer a special-interest subject of internet activists, programmers or marketers. They are a major public issue, and their development and application relate to significant reputational concerns. Not all these concerns are equal. While the evidence they produce may or may not be inconclusive or misguided and their outcomes may or may not be unfair or discriminating, their processes will *at any rate* remain, at least in part, opaque and inconceivable, even to the data engineers that create them. This opacity constitutes an inherently *obtrusive* reputational concern for the developers, proprietors, and users of algorithms: when critical stakeholders demand information and transparency, the proprietors and users will inevitably struggle to give explanations.

We made the argument that such concerns about reputation may pressure organizations towards accountability and compel them to enter proactive debates. The practically unattainable algorithmic transparency, however, produces limitations on the ethical duties of organizations to deliver conventional accounts within such debates. 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. In this sense algorithmic accountability can also be seen as a "grand challenge" (c.f. Colquitt & George, 2011, p. 432) that directs the attention not just towards "robust action" (Ferraro, Etzion, &

Gehman, 2015) as a pragmatic solution but rather "communicative action" as an ethical one. Technical opacity cannot serve as a general excuse. This is to say that while the practical opacity of algorithms may free proprietors of some of the common duties to account-giving, it charges them with additional duties to facilitate discourses about them—and that also includes discourses in the future, when technologies for understanding algorithmic action are further developed. Similar to other domains in which the mere right to transparency does not necessarily create fairness (Edwards and Veale, 2017), the ethical solution lies nevertheless in the creation of interactive and discursive fora and processes. In the context of practical opacity, algorithmic accountability needs not reporting standards but standards for accountability discourses.

In practice, this will play out based on the socio-technological approaches that are available to increase the quality of discourses on opaque and fluid technologies, approaches that will in turn need to be evaluated. On the organizational level, such approaches and empirical measures can extend reputation management practices for securing actors' room for manoeuvre and ability to acquire resources. On the societal level they can inform benchmarks and forms of regulation with high reputational relevance that may act as future standards for algorithmic discourses. Scrutiny of such practices will allow to assess if and in how far actual mechanisms can set up reputation management as a practice of listening and learning or if they are structured in a way that ultimately leads to manipulation or mere image work on an impression of accountability.

Of course, actual accountability practices, inquiries, and discourses rub up against these ideal requirements. However, the empirical use of this model has important critical potential in that it allows for identifying possible contradictions between the data and the model, which indicate constraints that require serious attention (cf. Habermas, 2006). Furthermore, its relative blindness towards alleged specificities of algorithms allows for questioning the extent to which

these developments are actually as novel as many commentators claim and thus actually require distinctive discourse-ethical treatment. As demonstrated with the opacity discourse starting in the early 1960s, software has always been said to transcend human comprehension. As a result, current discourses about the exceptionality of algorithms also need to be aware that in the near future, this exceptionality maybe deemed only an incremental step in a long history of technology.

As such, the normative principles help not only a) detect and describe the extent to which specific cases lack or have the potential to resolve accountability issues through discourse but also b) clarify that the 'opacity challenge' found in the contemporary debate is also a common (and recurring) diagnosis about new technologies. This requires their integration in long-established normative frameworks of democratic societies rather than adjusting these frameworks to their alleged specificity—dispensing of them as objects of ethical concern.

References

ACM Association for Computing Machinery US Public Policy Council (2017). *Statement on Algorithmic Transparency and Accountability*. Available at: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf (accessed December 1st, 2017).

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93-117.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, (3)2, 1-17.

Author & Colleagues (2017). *Blinded for review purposes.*

Bachmann, P. (2017): *Medienunternehmen und der strategische Umgang mit Media Responsibility und Corporate Social Responsibility*. Wiesbaden: Springer.

Barnet, B.A. (2009). Idiomedia: The rise of personalized, aggregated content. *Continuum* 23(1), 93–99.

Bartlett, J. L., Pallas, J., & Frostenson, M. (2013). Reputation and legitimacy: accreditiation and rankings to assess organizations. In C. E. Carroll (Ed.), *The Handbook of Communication and Corporate Reputation* (pp. 530-544). Malden, MA: Wiley & Sons.

Beck, M. (2016). Can a Death-Predicting Algorithm Improve Care? *Wall Street Journal*, 2. December 2016.

Beer, D. (2009). Power through the Algorithm? Participatory Web Cultures and the Technological Unconscious. *New Media & Society*, 11(6), 985-1002.

Beller, M., Zaidman, A., Karpov, A. & Zwaan, R. (2017). The Last Line Effect Explained. Empirical Software Engineering, 22(3), 1508-1536, doi:10.1007/s10664-016-9489-6.

Bernaz, N. (2013). Enhancing Corporate Accountability for Human Rights Violations: Is Extraterritoriality the Magic Potion? *Journal of Business Ethics*, 117(3), 493-511. doi:10.1007/s10551-012-1531-z

Borgman, C. L. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge/MA: MIT Press.

Bovens, M. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. West European Politics, 33(5), 946-967.

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13(4), 447–68.

Burrell, J. (2016): How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1-17.

Busuioc, M. & Lodge, M. (2017). Reputation and Accountability Relationships: Managing Accountability Expectation through Reputation. *Public Administration Review*, 77(1), 99-100.

Busuioc, M. (2013). *European Agencies: Law and Practices of Accountability*. Oxford: Oxford University Press.

Carlson, M. (2015). The Robotic Reporter. *Digital Journalism*, 3(3), 416-431.

Carmona, S., Donoso, R., & Reckers, P. M. J. (2013). Timing in Accountability and Trust Relationships. *Journal of Business Ethics*, 112(3), 481-495. doi:10.1007/s10551-012-1273-y

Coombs, T. W. (2013). Situational theory of crisis: situational crisis communication theory and corporate reputation. In C. E. Carroll (Ed.), *The handbook of communication and corporate reputation* (pp. 262-278). Malden, MA: Wiley-Blackwell.

Crawford, K. (2016). Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology & Human Values*, 41(1), 77-92.

Danna, A., & Gandy, O. H. (2002). All that glitters is not gold: Digging beneath the surface of data mining. *Journal of Business Ethics*, *40*(4), 373-386.

De Cremer, D., & Barker, M. (2003). Accountability and cooperation in social dilemmas: The influence of others' reputational concerns. *Current Psychology*, 22(2), 155-163. doi:10.1007/s12144-003-1006-6

Deephouse, D. L., & Carter, S. M. (2005). An Examination of Differences Between Organizational Legitimacy and Organizational Reputation. *Journal of Management Studies*, 42(2), 329-360. doi:doi:10.1111/j.1467-6486.2005.00499.x

Desmarais, S. L., & Singh, J. P. (2013). Risk assessment instruments validated and implemented in correctional settings in the United States. Lexington: Council of State Governments.

DeZoort, F. T., & Harrison, P. D. (2016). Understanding Auditors' Sense of Responsibility for Detecting Fraud Within Organizations. Journal of Business Ethics. doi:10.1007/s10551-016-3064-3

Diakopoulos, N., & Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*, 5(7), 809-828.

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism*, 3(3), 398-415.

Dörr, K.N., & Hollnbuchner, K. (2017). Ethical Challenges of Algorithmic Journalism, *Digital Journalism*, 5(4), 404-419.

Doshi-Velez, F., & Kortz, M. (2017). Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper.

Dubnick, M. J., & Frederickson, H. G. (2010). Accountable Agents: Federal Performance Measurement and Third-Party Government. J*ournal of Public Administration Research and Theory*, 20(suppl_1), i143-i159. doi:10.1093/jopart/mup039

Edwards, L., & Veale, M. (2017). Slave to the algorithm?. *Why a 'Right to Explanation' is Probably Not the* Remedy You Are Looking For. 16 Duke Law & Technology Review 18 (2017). Available at SSRN: https://ssrn.com/abstract=2972855 or http://dx.doi.org/10.2139/ssrn.2972855

Eisenegger, M., & Imhof, K. (2008). The true, the good and the beautiful: Reputation management in the media society. In A. Zerfaß, B. v. Ruler, & K. Sriramesh (Eds.), *Public relations research: European and international perspectives and innovations* (pp. 125-146). Wiesbaden: VS Verlag.

EPIC, Electronic Privacy Information Center (2017). Algorithms in the Criminal Justice System. Available at: https://epic.org/algorithmic-transparency/crim-justice/ (accessed August 25th, 2018).

Fanta, A. (2017). Putting Europe's Robots on the Map: Automated journalism in news agencies. Available at: https://reutersinstitute.politics.ox.ac.uk/our-research/putting-europes-robots-map-automated-journalism-news-agencies (accessed December 19th, 2017).

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, 25(4), 435–437.

French, W., Zeiss, H., & Scherer, A. G. (2001). Intercultural Discourse Ethics: Testing Trompenaars' and Hampden-Turner's Conclusions about Americans and the French. *Journal of Business Ethics*, 34(3-4), 145–159.

Garber, M. (2016). When Algorithms Take the Stand, *The Atlantic*. June 30, 2016.

Gilad, S., Maor, M., & Bloom, P. B.-N. (2015). Organizational Reputation, the Content of Public Allegations, and Regulatory Communication. *Journal of Public Administration Research and Theory*, 25(2), 451-478. doi:10.1093/jopart/mut041

Gilbert, D. U., & Rasche, A. (2007). Discourse Ethics and Social Accountability: The Ethics of SA 8000. *Business Ethics Quarterly*, 17(2), 187-216.

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P.J. Boczkowski & K.A. Foot (Eds.), Media Technologies. Essays on Communication, Materiality, and Society, Cambrdige/MA: MIT Press, pp. 167-194.

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

Gray, R. (2002). The social accounting project and Accounting Organizations and Society Privileging engagement, imaginings, new accountings and pragmatism over critique? *Accounting, Organizations and Society*, 27(7), 687-708. doi:https://doi.org/10.1016/S0361-3682(00)00003-9

Greenwood, M. (2007). Stakeholder Engagement: Beyond the Myth of Corporate Responsibility. *Journal of Business Ethics*, 74(4), 315-327.

Greenwood, R., Raynard, M., Kodeih, F., Micelotta, E. R., & Lounsbury, M. (2011). Institutional Complexity and Organizational Responses. *Academy of Management Annals*, 5(1), 317-371. doi:10.5465/19416520.2011.590299

Habermas, J. (1999). *Moral Consciousness and Communicative Action* (trans. Christian Lenhardt, Shierry Weber Nicholsen). Cambridge, Mass.: 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), 411-426.

Hildebrandt, M. (2008). Profiling and the rule of law. *Identity in the Information Society*, *1*(1), 55-70.

Hoos, F., Pruijssers, J. L., & Lander, M. W. (2017). Who's Watching? Accountability in Different Audit Regimes and the Effects on Auditors' Professional Skepticism. *Journal of Business Ethics*. doi:10.1007/s10551-017-3603-6

Hunt, S. K., & Dumville, R. (2016). *Recidivism Among Federal Offenders: A Comprehensive Overview*. United States Sanctioning Commission. Available at: https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-publications/2016/recidivism_overview.pdf (accessed December 19th, 2017)

Karpoff, J. M. (2012). Does reputation work to discipline corporate conduct? In M. L. Barnett & T. G. Pollock (Eds.), *The Oxford handbook of corporate reputation* (pp. 361-382). Oxford: Oxford University Press.

Kehl, D., Guo, P., & Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Berkman Klein Center for Internet & Society, Harvard Law School. Available at: https://cyber.harvard.edu/publications/2017/07/Algorithms (accessed December 19th, 2017)

Kim, H., Giacomin, J., & Macredie, R. (2014). A qualitative study of stakeholders' perspectives on the social network service environment. *International Journal of Human-Computer Interaction*, *30*(12), 965-976.

Kim, M., Bergman, L., Lau, T., & Notkin, D. (2004, August). An ethnographic study of copy and paste programming practices in OOPL. In *Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on* (pp. 83-92). IEEE.

King, B. G., & Whetten, D. A. (2008). Rethinking the Relationship Between Reputation and Legitimacy: A Social Actor Conceptualization. *Corporate Reputation Review*, 11(3), 192-207. doi:10.1057/crr.2008.16

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), 164-178. doi:https://doi.org/10.1016/S0090-2616(01)00026-2

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*, San Francisco: W.H. Freeman & Company.

Martin, K. (2018). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, Online First, pp. 1-16, doi:10.1007/s10551-018-3921-3.

McDonnell, M.-H., & King, B. (2013). Keeping up Appearances:Reputational Threat and Impression Management after Social Movement Boycotts. *Administrative Science Quarterly*, 58(3), 387-419. doi:10.1177/0001839213500032

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

Minsky, M. (1967). Why programming is a good medium for expressing poorly understood and sloppily formulated ideas. *Design and planning II-computers in design and communication*, 120-125.

Mittelstadt, B. (2016). Auditing for Transparency in Content Personalization Systems. *International Journal of Communication*, *10*, 4991-5002.

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

Montal, T., & Reich, Z. (2017). I, Robot. You, Journalist. Who is the Author? Authorship, bylines and full disclosure in automated journalism. *Digital Journalism*, *5*(7), 829-849.

Nanz, P., & Steffek, J. (2005). Assessing the Democratic Quality of Deliberation in International Governance: Criteria and Research Strategies. *Acta Politica* 40, 368-383.

Naughton, J. (2016). Opinion, Even Algorithms Are Biased Against Black Men. *The Guardian*. June 26, 2016.

Niemi, J. I. (2008). The foundations of Jürgen Habermas's discourse ethics. *The Journal of Value Inquiry*, *42*(2), 255-268.

Nissenbaum, H. (2011). A Contextual Approach to Privacy Online. *Dædalus, the Journal of the American Academy of Arts & Sciences*, 140(4), 32–48.

Norris, P. (2014). Watchdog journalism. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford handbook of public accountability*. Oxford: Oxford University Press.

Owen, D. L., Swift, T. A., Humphrey, C., & Bowerman, M. (2000). The new social audits: accountability, managerial capture or the agenda of social champions? *European Accounting Review*, 9(1), 81-98. doi:10.1080/096381800407950

Palazzo, G., & Scherer, A. G. (2006). Corporate legitimacy as deliberation: A communicative framework. *Journal of Business Ethics*, 66(1), 71-88.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Pasquale, F. (2010). Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries. *Northwestern University Law Review*, 104, 105.

Passig, K. (2017): Fünfzig Jahre Black Box. *Merkur. Gegründet 1947 als Deutsche Zeitschrift für europäisches Denken,* 823(12), 16-30.

Pentland, B. T., & Feldman, M. S. (2005). Organizational routines as a unit of analysis. *Industrial and Corporate Change*, 14(5), 793-815.

Rasche, A., & Esser, D. (2006). From Stakeholder Management to Stakeholder Accountability Applying Habermasian Discourse Ethics to Accountability Research. *Journal of Business Ethics*, 65(3), 251–267.

Rindova, V. P., Pollock, T. G., & Hayward, M. L. A. (2006). Celebrity Firms: The Social Construction of Market Popularity. *The Academy of Management Review*, 31(1), 50-71. doi:10.2307/20159185

Romenti, S. (2010). Reputation and stakeholder engagement: an Italian case study. *Journal of Communication Management*, *14*(4), 306-318.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014a). An Algorithm Audit. In: S.P. Gangadharan (ed.), *Data and Discrimination: Collected Essays* (pp. 6-10). Washington, DC: New America Foundation.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014b). *Auditing algorithms: Research methods for detecting discrimination on internet platforms*. Paper presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA.

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), 259-284. doi:doi:10.1111/joms.12014

Scurich, N., & Monahan, J. (2016). Evidence-based sentencing: Public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law and Human Behavior*, 40(1), 36.

Seele, P. & Lock, I. (2015). Instrumental and/or Deliberative? A Typology of CSR Communication Tools. *Journal of Business Ethics*, 131(2), 401-414.

Smith, M. (2016). In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures. *NY Times*. June 22, 2016.

Stalder, F. (2016): *Kultur der Digitalität*. Suhrkamp: Berlin.

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., Bachtiger, A., Sporndli, M., & Steiner, J. (2003). Measuring Political Deliberation: A Discourse Quality Index. *Comparative European Politics*, 1(1), 21-48.

Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *Academy of Management Review*, 20(3), 571-610. doi:10.5465/amr.1995.9508080331

Suurmond, G., Swank, O. H., & Visser, B. (2004). On the bad reputation of reputational concerns. *Journal of Public Economics*, 88(12), 2817-2838. doi:https://doi.org/10.1016/j.jpubeco.2003.10.004

Swift, T. (2001). Trust, reputation and corporate accountability to stakeholders. *Business Ethics, a European Review*, 10(1), 16-26.

Tutt, A. (2016). *An FDA for algorithms. Social Science Research Network*. Available at http://papers.ssrn.com/abstract=2747994 (accessed December 14th, 2017).

Van Buren, H. J. (2001). If Fairness is the Problem, Is Consent the Solution? Integrating ISCT and Stakeholder Theory. Business Ethics Quarterly, 11(3), pp. 481-499.

Van de Walle, S., & Cornelissen, F. (2014). Performance reporting. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability*. Oxford: Oxford University Press.

Van Otterlo, M. (2013). A machine learning view on profiling. In M. Hildebrant & K. de Vries (eds.), *Privacy, due process and the computational turn: Philosophers of law meet philosophers of technology* (pp. 46–64). London, UK: Routledge.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, 131, 1355-1358.

Wojciechowski, B. (2010). Discourse Ethics as a Basis of the Application of Law. In: Jemielniak J., Miklaszewicz P. (eds), Interpretation of Law in the Global World: From Particularism to a Universal Approach (pp. 53-69). Springer: Berlin.

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 & Human Values*, 41(1), 118–132.