Rethinking Transparency as a Communicative Constellation

Florian Eyert[*]
WZB Berlin Social Science Center
and Weizenbaum Institute
Berlin, Germany
florian.eyert@wzb.eu

Paola Lopez[*]
University of Vienna
and Weizenbaum Institute
Vienna and Berlin, Austria and Germany
paola.lopez@univie.ac.at

ABSTRACT

In this paper we make the case for an expanded understanding of transparency. Within the now extensive FAccT literature, transparency has largely been understood in terms of explainability. While this approach has proven helpful in many contexts, it falls short of addressing some of the more fundamental issues in the development and application of machine learning, such as the epistemic limitations of predictions and the political nature of the selection of fairness criteria. In order to render machine learning systems more democratic, we argue, a broader understanding of transparency is needed. We therefore propose to view transparency as a communicative constellation that is a precondition for meaningful democratic deliberation. We discuss four perspective expansions implied by this approach and present a case study illustrating the interplay of heterogeneous actors involved in producing this constellation. Drawing from our conceptualization of transparency, we sketch implications for actor groups in different sectors of society.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; *Philosophical/theoretical foundations of artificial intelligence*; • **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

transparency, explainability, science communication, deliberation, prediction

ACM Reference Format:
Florian Eyert and Paola Lopez. 2023. Rethinking Transparency as a Communicative Constellation. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3593013.3594010

1 INTRODUCTION

With the spread of tools for the generation of text and images, machine learning continues to gain public attention. In this process, different narratives about "artificial intelligence" are competing and reinforcing each other, from groundbreaking innovation to existential threat. How these technologies are talked about and, thus, reflected on, is strongly shaped by influential actors who are able to back up their interests with resources. This leads to a situation where critical and sober outlooks are often sidelined. Technological hypes are sociologically complex phenomena and in order to systematically engage with them, a thorough reflection on societal knowledge practices is important. The FAccT community has so far only marginally investigated the ways in which societal communication about machine learning and artificial intelligence is structured and how a discourse based on careful evaluation and democratic deliberation can be fostered. In terms of facilitating knowledge about machine learning systems, the literature within FAccT has focused on technical methods of explainability and, partially, on the notions of disclosure and algorithmic literacy. We argue that these approaches do not exhaust the potential of "transparency" and that it is fruitful to also take into account the societal environment that surrounds machine learning applications: the conditions necessary to enable broad debates, the actors who can facilitate discourse, and the ways in which existing transparency efforts can be brought into conversation with the general public. As of yet, FAccT has not systematically engaged with these topics.

In this paper, we investigate this research gap and propose to address it by rethinking transparency as a communicative constellation – an approach that takes seriously the epistemic, systemic, multidirectional and contextual dimensions of the negotiation of machine learning systems and that links the conversation within the FAccT community to the literature on science communication. In the following, we proceed in five steps: In Section 2, we characterize the development and application of machine learning systems as a social process that inevitably entails choices that have no unambiguously optimal outcome. These choices, we argue in Section 3, call for a democratic response. Conventional approaches to transparency, however, fail to fully enable such a democratic response, as we show in Section 4. In Section 5, we then offer an expanded conception of transparency that we call *transparency as a communicative constellation* (TaCC). Finally, we lay out some of the implications of such a perspective for different societal groups in Section 6 and conclude with a brief discussion.

2 CHOICES IN MACHINE LEARNING

In the past years, the development of machine learning systems has been subjected to a myriad of critiques, especially in, but not limited to, the context of the FAccT conferences. It has been pointed out that when machine learning methods and large-scale models are deployed in sensitive areas and, consequently, make or guide decisions that affect the lives of human beings, various negative effects can occur. Among them are issues of bias and subsequent

[*] Both authors contributed equally to this research.

FAccT '23, June 12–15, 2023, Chicago, IL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594010

discrimination [84] and representational harm [76], the exacerbation of inequalities [10, 34], surveillance [93], misinformation [25] and environmental degradation [8].

As a response to these issues, the FAccT community has developed numerous valuable technical approaches to them, from fairness frameworks to explainability tools. While they provide strategies for more effectively realizing the values of those applying them, they do not resolve the problem of negotiating these values in the first place. Even when maximizing overall accuracy – an endeavor that, compared to others, is easy to quantify and assess – there is the issue of "model multiplicity" [14], i.e., multiple models yielding similar accuracy, which subsequently requires active choices. As scholars in science and technology studies have demonstrated for decades, technology is never neutral, but rather embeds and realizes particular courses of action, values and worldviews [1, 90], which has been confirmed in FAccT research [13]. Human choices are plentiful in all steps of the development and deployment of algorithmic systems. These choices, we stress, are not merely choices about technology, but choices about the way in which the social world is imagined and society is represented and ordered. Two particularly important examples illustrate this.

Firstly, using machine learning as a tool for prediction enacts a very particular relation to the future that is, curiously, scarcely discussed in the FAccT community. As a *terminus technicus* in the context of machine learning, prediction consists in fitting a function to existing data points in order to estimate some outcome for a previously unseen instance. The "pre" in prediction is, first of all, merely a question of definition and refers to inferring missing data from existing data, whether the former is unrecorded, falsely recorded, a missing word in a sentence, or refers to future events. In many models, time does not mathematically exist as a parameter or the future is simply assumed to be a continuation of the past. In non-technical contexts, however, "prediction" is often understood or presented in the sense of clairvoyance [33]. Contextualizations in economic, political and media discourse infuse the "pre" with this non-technical meaning. This leads to the expectation of prediction in the strong sense, clearly what Raji et al. call an overstated capability and impossible task [65]. When employed to actually make statements about the future, this can have substantial effects for those affected. Among the most striking are examples from the criminal justice area, such as the prediction of recidivism [4] or predictive policing [24, 30]. While critical analyses often focus on the injustice of bias through demographically differing accuracy rates in prediction, the injustice of prediction itself is less often discussed. There are a variety of ways of relating to the future [6, 61], and using data about the past and patterns inferred through machine learning to produce knowledge and, subsequently, actions about the future is a particular choice among them. It imagines society in a very specific way and prefigures ways of acting upon it. The assumption that the future is a continuation of the past is an even stronger assumption with respect to complex social systems and individual lives than it is with respect to other domains. Given the societal ramifications as well as the way predictions are framed in public discourses, the decision to use machine learning to relate to and, thereby, shape the future is a fundamental value choice.

Secondly, we want to draw attention to the choices surrounding fairness metrics. Proposed as a way to mathematically formalize normative demands regarding the minimization of bias, there are now famously numerous different metrics [37] that cannot be achieved simultaneously [26]. While the FAccT conversation increasingly challenges this technical definition of fairness and complements it with a more justice-oriented perspective [29, 73, 92], fairness metrics still play a significant role in the field. Selecting appropriate fairness criteria is not a purely technical task, but requires choices about how society is to be envisioned. Each fairness metric is related to particular assumptions about the social world and particular political philosophies [12] and must be evaluated with respect to social context [69]. Rather than providing a "view from nowhere", from which normative questions can be assessed in an objective manner, fairness metrics are embedded in contexts of justification [16] that rely on "situated knowledge" [44], i.e. knowledge that is based on lived experience and positions in society, which, for instance, feminist critique has aimed to make visible [63]. Here again, we emphasize, fundamental decisions about some ideal state of society are irreducibly part of the development and use of machine learning systems.

Both aspects – data-based prediction as a mode of relating to the future and fairness metrics as a mode of encapsulating justice claims – involve fundamentally normative decisions. They are examples of choices that have to be made without exclusively relying on technical necessities or optimal solutions. From this perspective the question of who takes part in making these decisions inevitably comes into view.

3 CHOOSING DEMOCRATICALLY

The relations we build between the past and the future, as well as the ideas we have about what is fair and just, are, in principle, not problems that can be "solved" once and for all. Rather, they have to be continually decided upon and reconfirmed. Since these issues are political and not merely technical, they are determined within temporary social arrangements that emerge from a multitude of negotiations, frictions, competing claims, power relations, and conflicts. From this perspective, the focus shifts from the appropriateness of a given technical solution to the *processes* through which answers are found to these questions.

In this paper, we are interested in the implications following from the position that these processes should be shaped democratically. We take democracy to refer to not just an electoral system, but a more general principle for structuring society that aims at the just distribution and critical evaluation of power. As machine learning systems are integrated into important societal institutions, such as the criminal justice system, news and social media platforms or welfare services, they should, from this perspective, be democratically shaped and assessed. While we do not endorse a particular tradition of democratic theory, this paper is inspired by the approaches of deliberative democracy [42] and radical democracy [52, 81]. From deliberative democracy, we take the focus on the sites in which different arguments and worldviews can be exchanged and translated into political agency, while at the same time acknowledging the effect of power relations that are at play in every social constellation. From radical democracy, we take the emphasis on the radical contingency – the possibility of things being different — of all collective arrangements and the fundamental necessity of the confrontation of disagreeing views and interests. Our perspective, thus, requires that there are mechanisms that make it possible to bring issues of societal concern to public dialogue, politicize them, and treat them as collectively shapeable rather than as inevitable facts. This is particularly important in the case of machine learning systems, since, as outlined above, they are intertwined with particular ways of representing and governing societies.

Following this approach, the appropriateness of a machine learning tool is to be measured by the quality of democratic deliberation that legitimizes its use. Precisely this, however, turns out to be a challenge when it comes to complex technologies [22]. If decisions are to be rooted in collective deliberation but it is not clear how the issues at stake can be understood or what the issues even are, a democratic vacuum emerges. Machine learning then appears as something inevitable and merely technical rather than the result of societal choices. The political problem of how choices around machine learning systems can be made democratically is therefore strongly tied to the epistemic problem of how these systems are known and can be known within a society. We refer to the extensive literature in science and technology studies that discusses this intricate interrelation between knowledge and politics [19, 49, 71].

In order to systematize the ways in which the FAccT literature has addressed the issue of knowledge so far, it is helpful to draw on Burrell's taxonomy of forms of opacity in machine learning [21]. She distinguishes between three obstacles that stand in the way of the endeavor of knowing or understanding machine learning systems: firstly, intentional organizational secrecy that renders it impossible for an outsider to get access to information; secondly, a lack of technical expert knowledge on the part of the general public that hinders non-experts from understanding machine learning systems; and thirdly, high technical complexity, e.g., a large number of features in a model, that makes it difficult even for experts to understand the reasons behind a decision. In the FAccT literature, these have, for the most part, been addressed by, respectively, calling for disclosure, stressing algorithmic literacy, and developing technical methods for fostering model transparency, with a very strong emphasis on explainability. In the remainder of this paper, we discuss this understanding of transparency and propose a critical rethinking that centers the objective of choosing democratically.

4 LIMITATIONS OF CURRENT APPROACHES TO TRANSPARENCY

Transparency as a topic in itself has received comparatively little conceptual consideration in FAccT scholarship. In their systematic study of the topics of "Four Years of FAccT", Laufer et al. show that while *fairness* has been in the center of FAccT attention, with 69% of papers addressing it, *transparency* was only talked about in 26% of papers [56]. Half of those papers, according to the authors, are concerned with explainability and interpretability, i.e. a very specific and largely technical approach to transparency. Analyzing the communities forming around different research topics, they identify an explainability community but no transparency community among the ten largest communities. When reviewing the programs of past FAccT conferences, it is striking that while there are numerous paper sessions about explainability and interpretability, none about transparency can be found. While there is a plethora of papers on explainability, reviewed in [74], transparency is rarely talked about for its own sake. It seems that to the extent that transparency is a part of the FAccT conversation, it is largely treated in terms of a very specific approach.

In this paper, we are concerned with the question of what might lie beyond this approach. While explainability plays an important role in making societal uses of machine learning more intelligible within democratic deliberation, we stress that it is only one possible way of thinking about transparency. In the FAccT literature, there is a growing reflexivity about this and recent research has started to point out limitations [17]. Partly drawing on these insights, we identify four structural limitations of the currently dominant approach to transparency with respect to the normative ideal of democratic deliberation sketched above.

Firstly, as pointed out, the current approach to transparency is largely *centered around technology.* Transparency is primarily seen as the endeavor to develop tools that provide additional information about a given model. This frames the problem in a particular way and enables particular courses of action. It has, for instance, been shown that explainability tools are often mainly used for debugging [11]. While there are suggestions to take into account human sensemaking [50] and social context and position [43] in the design and evaluation of explainability methods, the suggested instruments for reaching transparency are nonetheless largely technical. Besides explainability, other measures aiming at transparency, such as documentation frameworks like model cards [62] and data cards [64] or transparency information languages [41], have the advantage of focusing transparency on social rather than technical processes, but are themselves mostly articulated through technical means. None of these transparency approaches facilitate critical engagement with the underlying methodologies of prediction, nor do they assist in the process of defining what notion of fairness – if any – is to be implemented.

Secondly, current approaches to transparency mostly *focus on individuals*, for instance as recipients of explanation or as addressees of disclosure. This is rooted in a conceptualization of transparency as a problem of individual knowers, which in turn narrows how we think about transparency. Lima et al. point out that post-hoc explanations could even lead to a conflict with accountability if individuals subjected to algorithmic decisions are scapegoated [59]. Machine learning, however, operates under a fundamentally non-individualist paradigm: Its predictions are based on large-scale samples, generalize from groups to individuals and intertwine individual privacy and group privacy [80]. Matters of collective concern, like the temporalities of prediction and the negotiation of justice, cannot be addressed through approaches only aiming at individuals.

Thirdly, transparency is commonly approached in a *unidirectional manner*, in which a given technology is the starting point that is then made transparent through multiple successive layers, from engineers to, ideally, the public. This unidirectionality, however, has consequences for the social relations in which it operates. It has been pointed out that in adversarial situations post-hoc explanations have considerable limitations due to potential conflicts of interest between providers and receivers of explanations [17]. With regard to emerging documentation practices, it has been argued that their focus on tech firms as the center of agency might be counterproductive [38]. Furthermore, if the ways in which we relate to the future and the forms of justice we choose are fundamentally to be determined in a democratic exchange and struggle, a unidirectional approach can only go a limited part of the way. Another example of unidirectionality is the approach of disclosure. Legally forcing, or trying to force, tech companies to disclose technical information, underlying business models, or practices around data processing is crucial for enabling debates. But disclosing information to users or the broader public is an *ex post* practice: What is being disclosed is information about a system that has already been set up. Relevant communication only flows one way. This also applies to many systems in which user feedback is used for improvement. Since the system has already been set up, the "whether" question has been answered and it is the "how" question that is being optimized – often with premises that are not up for debate. This is very different from a democratic dialogue that takes a step back and reflects on the question of deployment itself.

Fourth, established conceptualizations of transparency problematize machine learning primarily with *existing models as a point of departure* and attempt to create explanations, interpretations or disclosures after the fact, instead of starting with societal context. A case in point is the issue of trust. The FAccT literature has suggested that information and explanations can play a role in strengthening overall trust in algorithmic systems: If a user knows more about what a machine does then her interaction with it will be infused by trust [58, 68, 82]. This is, however, an ambivalent endeavor, as algorithmic systems, in many constellations, are better not to be trusted [36, 65]. When taking existing machine learning models as a given, the more fundamental question about the suitability of the kinds of relations to the future it implies remains invisible. The issue of the selection of fairness metrics is another case in point: Fairness metrics are always articulated through and, thus, constrained by the quantitative logics and functionalities internal to machine learning models. They aim to solve the social problem of bias, but only within the constraints given by the technology.

All in all, current approaches to transparency do not address the full spectrum of issues involved in democratizing the use of machine learning – not because they fail at the task they are developed for but because they are developed for different tasks. They do not engage with media discourses that create harmful expectations, nor do they tackle underlying value choices or facilitate a discourse between those who build, those who employ, those who interact with, and those who are affected by a system. It is therefore important to explore ways of conceptualizing transparency in a more comprehensive way.

5 TRANSPARENCY AS A COMMUNICATIVE CONSTELLATION

In the following we introduce the notion of transparency as a communicative constellation (Section 5.1), describe four conceptual shifts associated with it (Section 5.2) and provide a case study that illustrates it (Section 5.3).

5.1 Towards transparency as a communicative constellation

We suggest expanding the concept of transparency towards *transparency as a communicative constellation* (TaCC). By doing so, we do not aim to dismiss or replace the other forms of transparency mentioned so far. We view critiquing the approaches to transparency discussed in the previous section as engaging constructively with existing research and as actively expressing care towards the collective project of fair, accountable and transparent machine learning. TaCC serves as a starting point for expansive thinking and, speaking with Patricia Hill Collins [28], viewing transparency through the lens of TaCC is a way of "organizing the thinking tools": those that are already present and those that we add.

Transparency as a communicative constellation refers to the society-level quality of the negotiations and debates around the development and application of a given technology. TaCC is a societal achievement that emerges as the result of the communicative effort of multiple actors and that can be thought of along several dimensions, which are described below. It includes translations across different fields, disciplines, domains and spaces. Understanding transparency as a communicative constellation acknowledges the ongoing-ness of the effort towards meaningful transparency and participation, and the push and pull of power relations that different actors striving for transparency are a part of. The project of meaningful transparency is never finished. Rather, it is an open-ended endeavor. As such, TaCC describes the extent to which there is an inclusive, critical and meaningful discourse about a given technology, as well as participation in the shaping of the technology.

Since the democratic community is at the center of our conceptualization, one component of TaCC consists in the general public having, or otherwise acquiring, some kind of technical knowledge. This is often discussed under variations of the umbrella term *literacy*: digital literacy, data literacy or algorithmic literacy are viewed as essential prerequisites for participating in discourses about technology. After all, one can only talk and, hence, negotiate, about what one knows at least a little bit about. Schoeffer et al., for instance, show that available information has an impact on the assessment of the fairness of automated decision systems [68]. Accounts of technical literacy often view those who are to gain literacy in a rather passive position – one in which they receive information and are merely educated about existing technologies that have been implemented without meaningful discourse. Recent proposals for critical data literacy [67] also include the capacity to critically assess digital technologies and the ability to reflect on the wider context in their definition of literacy. While literacy, especially critical literacy, is certainly a worthwhile endeavor, the literacy approach remains within a unidirectional paradigm of communication: Experts are to transfer technical knowledge to non-experts. Moreover, it often centers the individual as responsible for learning.

Going beyond this, we propose to draw from the field of *science communication* [20, 48, 66] as a valuable resource that has been curiously missing in the FAccT literature so far. We understand science communication in a broad sense, as all societal conversations – whether they are professional or not, public or private – about science and technology, rather than merely as the attempts of individual scientists and professional science communicators to unidirectionally communicate final results. Through enabling conversations about the facts and values involved in science and technology, science communication can be a vehicle for democratic deliberation rather than an afterthought in a grant proposal. These broad societal conversations can take many forms. Science communication scholar Brian Trench distinguishes between three models of science communication: the *deficit model*, the *dialogue model*, and the *participation model* [83]. The deficit model views the science communication situation as one in which the public is primarily characterized by a lack of knowledge about science and technology and in which it is the role of experts to disseminate their knowledge as best as possible. This model is closely related to the notion of literacy. Contrary to this unidirectional approach, both the dialogue model and the participation model pursue a more complex conception of science communication. In the dialogue model, a second direction of communication is added: "the problem is not that publics don't listen to scientists, but that scientists don't listen to publics" [47]. Dialogue, then, is achieved when scientists and technologists incorporate perspectives and preferences of those that are affected by their work through an ongoing exchange. The participatory approach shares the perspective of two-way communication and further intertwines different actors. It suggests that scientific research and technology development as activities themselves should include "lay people" and invite them to participate. Trench argues that rather than thinking of these models as in a succession or normative hierarchy, each model is appropriate in different situations. Efforts to achieve transparency in machine learning can then draw on a range of approaches depending on the circumstances.

A related and crucial component of TaCC is the role of language. For many people, the interaction with machine learning is mediated by the concepts that are dominant in public discourses around it. However, notions like "artificial intelligence", "self-learning algorithm" or "prediction" ultimately contribute to the obfuscation of the actual capabilities of machine learning systems through anthropomorphization or the invocation of other exaggerated expectations. Transparency as a communicative constellation, then, aims to facilitate the demystification of algorithmic systems by pointing to the gaps between the technical terms and the actual capabilities of and assumptions behind data-based models.

Viewing transparency through the lens of TaCC also provides an entrance point to unpacking the two inherently political issues discussed in Section 2. Firstly, while technical transparency only ever allows the discussion of particular applications and manifestations of machine learning, it provides no way of facilitating a debate about whether machine learning is an appropriate tool for a given purpose *in principle*. It remains unaddressed whether the answers a machine learning system can provide match the questions we ask of it. TaCC aims at bringing the societal debate closer to the limitations of machine learning systems and at overcoming the grave "communication failures" currently associated with machine learning [65]. As a language intervention in the service of TaCC, denoting *pre*dictions as *post*dictions might be a way to indicate the fundamental discrepancy between data-driven analyses based on data about the past on the one hand and prophetic capabilities on the other. Referring to predictions as *postdictions* might open up the debate through more sober and descriptive terminology.

Secondly, viewing transparency as a communicative constellation aims at addressing the questions sketched above with respect to technical fairness metrics. While they are in many cases an important tool for the mitigation of harms that machine learning systems would otherwise cause, the selection of a specific fairness metric can not be done on a purely technical basis but needs to be connected to societal ideas of justice. Transparency as a communicative constellation aims to make the relations between technical definitions of fairness and cultural ideas of justice accessible to the debates and negotiations of those who are concerned. Enabling conversations about both facts and values, TaCC then functions as a bridge between the development of technical tools and social justice efforts.

As these examples suggest, such a societal conversation has instruments at its disposal to circumvent the issues that technical transparency typically targets. While technical transparency aims at specific and concrete implementations of machine learning algorithms, TaCC acknowledges that machine learning systems can be evaluated and made sense of in other ways. If, for instance, public debate could rest on a solid understanding of and focus on the ways in which postdictions deviate from the ideal of predictive machines, whole classes of systems and use cases could be assessed on a societal level. If general properties and risks of machine learning technologies, such as the inherent restriction to being informed by the structures of the past, are more broadly visible, the sometimes less than obvious option of not using AI at all might become more plausible.

Transparency as a communicative constellation seeks to counteract the hype that surrounds artificial intelligence and machine learning technologies. It invites the FAccT community to collectively establish and systematize ways in which different methods, such as science communication, can effectively pierce through the hype and, thus, hopefully counter some negative effects. These efforts have a long intellectual tradition that dates back to early critics such as Joseph Weizenbaum. As a computer scientist working in the emerging field of AI during the 1960s, he developed a rule-based language model called ELIZA that was supposed to mimic a psychotherapist [86]. Enabling an interaction in natural language, it operated as a chat bot. After being shocked by the extent to which users overestimated ELIZA's capabilities and even trusted it with their innermost secrets, Weizenbaum became a critic of the hype discourses on artificial intelligence. He insisted that a more sober language and clarity about what a system is, in fact, able to do, as well as a focus on its limitations, was necessary. Situating itself in this tradition, the perspective of TaCC already has its contemporary examples, many of which are active in the field of FAccT. A thorough theoretical reflection of this, however, will allow a more systematic investigation of the conditions and strategies concerning this collective endeavor and will help to bring a broad range of heterogeneous efforts into a common horizon of analysis.

5.2 Perspective expansions

In the following, we specify our conceptualization of TaCC by way of describing four directions of expansion. With each shift of perspective, we address one aspect of the endeavor towards a broader understanding of transparency. We note that in each dimension the second category does not exclude the first one but contains it. We therefore do not suggest to discard previous accomplishments but embed them in a broader and more multifaceted approach.

Direction of expansion 1: From technical to epistemic transparency. Here we ask what is to be achieved when aiming for transparency. As discussed in Section 4, transparency in machine learning is currently primarily viewed as a technical task. Efforts towards technical transparency in the form of explainability strive to increase what individuals – such as developers or users – can know about a given machine learning system: e.g., which features have been most relevant for a conducted classification. Expanding this idea and abstracting from a specific system, one can ask about what it means – in general – for a machine learning system to be *known*. This opens up the question of epistemology: What kind of knowledge is produced within a machine learning system? Does this – data-driven – knowledge serve our purposes in all areas of application? *Epistemic transparency* includes the possibility of abstracting from a specific machine learning system to classes of systems that operate under the same general logic, and, subsequently, making this logic of decision-making known. The technical approach to transparency, then, is decentered as only one way among many in which transparency can be achieved. New ways come into view, including collective forms of knowledge like "algorithmic imaginaries" that can shape how technical systems are politicized [9]. Epistemic transparency contains a moment of construction in which the systems are discussed in a somewhat simplified way that centers the concerns of those involved. From this perspective notions like prediction become visible as what Bachelard has referred to as "epistemic obstacles" [7]: ideas and ways of thinking that have meaning within a certain historical constellation but constitute a barrier to advances in understanding and reflexivity. Going from a technical to an epistemic understanding of transparency then implies addressing opacity as well as obstacles in the attempt to affect how machine learning can be known.

Direction of expansion 2: From individuals to a societal constellation. Here we ask who is to be involved in transparency and on which societal level it is located. In many cases, the technical approach to transparency operates under an individualistic premise. Transparency as explainability is to be achieved with respect to a single individual that applies some strategy to query a given machine learning system. While of course this strategy is supposed to be used by many people, this constitutes a merely additive scaling-up without any interaction between the individuals. Through the conceptualization of TaCC, we view transparency as a collective effort happening in different areas, fields, and contexts and ultimately located on the systemic and structural level. Precisely because democracy, too, is a systemic and not an individual or group notion, the communicative constellation we have in mind is systemic in scope. But what kind of discourse is possible within a social and political context? It is specific configurations of power relations between different kinds of actors that facilitate or hinder debates. Thus, viewing transparency as a communicative constellation means being alert to the ways in which different actors – such as tech companies or political decision-makers – pursue their interests. Transparency emerges as the result of push-and-pull dynamics shaped by structural inequalities, power relations and resistance to them. The question of *who* communicates to *whom* about *what*, as well as the question of who will listen to whom is, and has always been, shaped by power relations but in turn also shapes them.

Direction of expansion 3: From unidirectional to multidirectional communication. Here we ask about the structure of the communication involved in constituting transparency. Current approaches to transparency are often constructed around a unidirectional logic, in which a technical product is deciphered by experts, who then, ideally, provide the means for lay people to retrace their insights. Our broader understanding, in contrast, assumes a more multidirectional pattern. Inspired by the dialogue and participation models of science communication, we aim to center the ways in which communication about machine learning systems can occur in multiple directions. This includes those who are building systems or doing research as possible receivers of communication from other contexts: from informing themselves about societal concerns to inviting inclusive debate from the beginning of projects or being a part of institutionalized public forums. This perspective considers the extent to which different groups of actors take positions as speakers, and it investigates the institutional requirements for achieving multidirectionality. It also implies an expansion from disclosure to discourse. While disclosures are an important means for achieving TaCC, other forms of transparency that involve dialogue even before a system is constructed come into view. Both deliberative and radical theories of democracy stress the importance of the possibility of such dialogue and possible confrontation, the former in order to translate concerns into politics, the latter in order to put into practice the contingent and political nature of existing structures.

Direction of expansion 4: From internal to contextual problematization. Here we ask about the content of communication and the kinds of problems that are discussed. Current notions of transparency commonly address problems that relate to a given model that is to be rendered transparent, such as providing reasons for a particular classification. We call this internal problematization because it takes its point of departure from the internal properties of the model. In contextual problematization, in contrast, the problems to be solved are first of all defined from the point of view of the model's external context. The construction and evaluation of the model then happens as a second step and in response to the societal demands that have been articulated. A somewhat parallel expansion has been proposed by Green and Viljoen, who envision algorithmic realism as "a new mode of algorithmic thinking that is attentive to the internal limits of algorithms and to the social concerns that fall beyond the bounds of algorithmic formalism" [40]. Contextual problematization relates to a multidirectional approach in that relevant issues can often be identified through engagement with a variety of external actors.

"Organizing our thinking tools" through these four expansions opens up a larger territory of transparency, in which particular approaches can be located and analyzed more systematically. The approach of literacy, for instance, can be understood as realizing epistemic transparency in that its strategies for furthering transparency do not primarily aim at technology but at education. In most cases, however, it remains tied to an individualist position that locates responsibility on the level of individuals or, at best, partly in the education system [21], which will lead to unequal outcomes depending on unequal resources. Participatory design approaches [53, 57] fulfill the principle of multidirectionality in that stakeholders are thought of as active participants. Yet, the focus on stakeholders might not fully realize the systematic and collective nature of democratic deliberation and, thus, not treat the issue as a "res publica" that concerns everybody. Including only those that are directly affected by the deployment of an algorithmic system fails to take into account the fact that the development and deployment of machine learning systems might have broader implications for the way a democracy functions. Furthermore, participatory design usually begins from the assumption that the deployment of a system has been decided upon – it is merely the way the system is designed that is up for discussion.

A crucial characteristic of transparency as a communicative constellation is the fact that it emerges from the interaction between, and through the synergy of, various parts in a larger ecosystem: civil society actors, government agencies, science journalists, NGOs, critical scientists and others. The perspective of TaCC acknowledges that transparency is not a one-time individual accomplishment, but rather a gradual and sometimes collaborative, sometimes confrontational endeavor. We illustrate this in the following section's case study.

5.3 The Austrian AMS algorithm as a case study

Several years ago, the Austrian Public Employment Service (*Arbeitsmarktservice* in German and *AMS* for short) planned to introduce a predictive classification system to segregate job-seekers into different categories with differing eligibility for support. The group placement was to be made according to the job-seeker's *chances* on the labor market, as predicted by the system using different data about job-seekers, including personal data on sensitive attributes such as, e.g., gender, age and health status. The classification system was not supposed to decide automatically, but to be used as a decision support tool for case workers. Currently, the case of the *AMS algorithm*, as it came to be known in the media, is pending before the Austrian Supreme Administrative Court due to the question of its accordance with the EU's General Data Protection Regulation. In the following, we sketch the communicative constellation at work and the efforts by multiple actors in furthering meaningful transparency and public discourse around the AMS algorithm.

In October 2018, Austrian media reported about the implementation of the algorithmic system [77], and published interviews with a responsible board member of the AMS explaining the advantages of the data-driven decision support tool [78, 87]. A few days later, some technical details were published in the form of a 15-page document by the research institute that built the algorithmic system [46]. The document includes some technical descriptions on the features that the algorithm used for decision-making, such as gender, age, health status, nationality, as well as some information about the ways in which the classification was supposed to function. Scientists later criticized the lack of technical transparency, as only little information was disclosed in the document [23] – information that was even, to some degree, misleading [3].

Still, some fundamental aspects of the classification and the ensuing rationale of decision-making could be inferred from the published document. However, these fundamental aspects had to be rendered understandable and, thus, transparent to the broader public. Several groups of actors were crucial in sparking a larger debate: science and tech journalists, activists, and critical scientists, among others. Science and tech journalists played a vital role in communicating the algorithmic system and its potential negative effects, such as the automation of discriminatory practices, to the broader public. They reported on the algorithm and used different techniques to make its decision rationale more understandable. One article, for example, created an interactive application so that readers could click on and, thus, try out different combinations of data entries to calculate their chances on the labor market according to the AMS algorithm [51]. This reconstruction of the AMS algorithm is based on incomplete information and mimics only one of the proclaimed 96 models for classification, as the technical details were not made fully transparent. Nevertheless, this technical tool served the purpose of science communication by acquainting the readers with the decision rationale without assuming technical knowledge. In this constellation, they enabled readers to view the system critically – an ideal example of epistemic transparency as a facilitator of public critique. This interactive and easy-to-use application also had more than double the comments than an article that only explained the published technical details in a text [79]. This demonstrates the interest by the public – interest that was elevated by the availability of tools for critique in discourse.

Scientists, too, criticized the algorithmic system. From the document that was published, they concluded that the implementation might entail negative effects, such as the built-in potential for discrimination. Scientists served as translators in two ways: via interviews with media outlets when the application of the AMS algorithm was announced [88, 89], and, after the application of the algorithmic system was decided on, in blogs of media outlets [23, 75]. Scientific publications were written that engaged with, and critiqued, the AMS algorithm (e.g., [2, 60, 70]). The responsible board member of the AMS published blog articles responding to the public critique and defending the system [54, 55] (see also [18]).

Twitter users and activists criticized the tool as well (e.g., [27, 45]). The Austrian network of non-profit, labor market oriented social integration enterprises, *arbeit plus*, published a position paper that included explanations about the AMS algorithm's decision rationale [5]. The NGO *epicenter.works* submitted inquiries to the AMS under a legal title that required the AMS to answer and, thus, to disclose more details [31]. The NGO also started a campaign including press conferences and a video that explained, in an easy-to-understand fashion, the potential risks of the AMS algorithm, thereby translating and making available the effects and the critique of the algorithm to the broader public [32]. The case of the AMS algorithm gained a lot of attention – also internationally – and, hopefully, sparked general and far-reaching debates about the caution with which one has to proceed when developing and implementing an algorithmic system, especially in a sensitive area.

This case study shows, on the one hand, the multi-actor and multi-stage nature, and, to a degree, the messiness of the quest for transparency as a communicative constellation. It is the constellation of several actors that enabled a broader discourse. This case also shows the potential for politicization, even in the absence of comprehensive technical transparency: politicization occurred even though only some details about the algorithmic classification were made public. Transparency in the broad sense served as a vehicle for resistance against the implementation of the algorithmic system because it didn't have to rely on complete technical details. The conclusions about the decision rationale, together with their translation, were enough to initiate public controversy and debate.

One can imagine what kind of discourse can be made possible with more transparency – not only ex post, but ex ante: a broad discourse about the implementation of algorithmic systems themselves before they are decided upon. Taking transparency seriously, the respective political decision-makers and the state agency actors as well as the practitioners who developed the tool would have contributed to the transparency of the algorithmic system *before* its decided implementation. Further, it would have been their obligation to translate the technical details to the broad public and to thereby facilitate a wider debate. This debate, in turn, could have had an effect on the system's development (see, e.g., [70]) or subsequent policy decisions, as discourse, of course, is not an end in itself.

6 IMPLICATIONS FOR SOCIETAL GROUPS

As argued above, transparency as a communicative constellation unfolds its potential in the interaction of various actors within a society-wide ecosystem. In order to make our suggestion more tangible, we describe its implications for a number of sets of actors. We stress that synergies between actors will not be straightforward and must be established and maintained across diverging interests. In the following, we lay out several potential elements of transparency as a communicative constellation.

Machine learning practitioners and researchers. While work on explainability methods and related strategies continue to be crucial, it will be worthwhile to investigate possible entry points for science communication approaches. Considering the political agency of machine learning professionals [39] and their role in shaping societal self-understanding, transparency can be furthered by more active engagement both with affected groups and non-technical disciplines. We acknowledge the constraints and various kinds of incentive misalignments that employment relations put on this, but point to emerging forms of collective organizing [15] as opportunities for transforming professional norms. Recognizing the need for continued learning about the complex societal implications of machine learning technologies can play a role in an attitude more strongly oriented towards transparency. In this, recent proposals for frameworks through which practitioners can engage with the limitations of their work [72] can be helpful.

Technology companies. Technology companies face evident conflicts of interest regarding a broad concept of transparency, as a discourse centering limitations and sober debate poses a challenge to current hypes. Similar to "ethics washing" [85] there is then a risk of transparency washing, which is further complicated by the entrenchment of the FAccT research landscape with business interests [91]. Working towards the realization of a broader form of transparency, however, can help in highlighting the role that these interests play in the design of technology within democratic debates about machine learning systems. Transparency in the sense of a communicative constellation cannot alter economic power relations by itself, but it can shape how societies relate to them.

Political decision makers in education, science and innovation. From the perspective of policy, our perspective suggests that it will be highly beneficial to reserve a certain amount of the large quantities of funding and investment going towards machine learning for strengthening transparency. This can happen through science communication and transparency budgets allocated to research projects or even making extensive forms of critical and independent science communication and public dialogue efforts mandatory. Funding science journalism projects and transparency hubs could be another approach. Finally, including critical reflection on machine learning in school and university education for a variety of disciplines can strengthen the communicative constellation.

Politicians and civil servants. Politicians and civil servants can also take part in societal communication about machine learning in a variety of ways. Most importantly, they can proactively assume the role of communicators and facilitators of transparency as well as listeners, for instance through the establishment of public forums, citizen dialogue formats and contact with NGOs. Actively establishing transparency institutions such as the AI registers already introduced in Helsinki and Amsterdam or even improving existing websites through clearer communication and improved language accessibility can be first steps. Lastly, similar to the way in which government agencies often draw on external technical expertise for providing services, they would further transparency by increasingly incorporating inter- and transdisciplinary expertise from critical research into their work.

Science and technology journalists. Even though journalists are subject to the logics of the attention economy in their field, they can play a crucial role in transparency as a communicative constellation by systematically covering risks and limitations associated with machine learning technologies and reporting about cases in which harm is caused in a way that explains the underlying principles. Through interviews, public discussions and research collaboration with experts on science and technology they can give voices to critical assessment, gain further insights and contribute to fostering a web of transparency. Lastly, a critical examination of the language used around machine learning, such as "self-learning", "prediction" or "artificial intelligence" will also be effective towards more transparency.

NGOs. NGOs and similar civil society actors in the space of technology justice, civic tech and other areas play a crucial role in the ecosystem of transparency as a communicative constellation as well. Close cooperation with other actors in this ecosystem, such as journalists and critical researchers, can help make strategies more effective. Focusing campaigns on clear communication of what machine learning systems are actually doing and how they are falling short of expectations can form a strong component of their work. Among many examples, the above-mentioned organization *epicenter.works* did just that with regard to the AMS algorithm through campaigns involving public events, an animated explanatory video and other material. Organizations like the German *AlgorithmWatch* combine research and advocacy to contribute to transparency in debates around machine learning and publish accessible reports and analyses. These examples show that while there are challenges to overcome with respect to the audience that is actually reached, seeing science communication as a more explicit part of their work can be an important methodology for civil society actors.

Activists. For activists, too, our perspective implies to treat the political and the epistemic as strongly connected and incorporate efforts of science communication into their approaches. Providing counter-narratives about machine learning and facilitating empowerment through epistemic transparency constitutes an important form of political action. One can expect an increasingly important role of what might be called "dissident technologists": individuals that have strong expertise in a technical field and use it to act as "public intellectuals" by challenging widespread misconceptions and bringing technically complex issues of common concern within the reach of public discourse.

Scientists and scholars. While the research space of machine learning is often strongly interdependent with industry, there is also potential for other scientists and scholars to be voices for transparency. Especially interdisciplinary fields like FAccT have the potential for developing perspectives that cover both facts and values and can be leveraged to frame and tackle important political issues democratically. While FAccT is in various ways already engaged in science communication and outreach, the perspective we propose entails that it would be well worth it to pay more attention to the 'T' in FAccT and strengthen its efforts in science communication in the future. This might require learning new skills and welcoming groups with different forms of expertise into the field. More generally, scientists from disciplines that are until now not typically engaged in science communication could begin to incorporate such efforts into their work more strongly. For instance, mathematical science communication has a potential to assume an increasingly public role by explaining the core mathematical principles underlying machine learning and critically contextualizing them with respect to their societal use [35]. Moreover, cooperation with science journalism can help scientists and scholars to reach wider audiences and initiate public debates.

While these are suggestions and conjectures about plausible ways forward, only empirical experimentation can show how dynamics will unfold. As argued in the previous section, the interaction between different groups of actors is crucial. Since the application of machine learning is fundamentally about diverging interests and views of how to refer to ourselves as a society, there will always be dissent and negotiation around it. Finding viable arrangements for cooperation is therefore vital.

7 DISCUSSION

In this paper, we have highlighted that the development and application of machine learning systems involves a number of implicit or explicit value judgements and that their political nature requires that they be democratically negotiated. We have argued that current notions of transparency fall short of providing the conditions for enabling such negotiation. As a remedy we have suggested an expanded understanding of transparency: transparency as a communicative constellation. This interpretation of transparency conceptualizes it as epistemic, systemic, multidirectional and contextual and brings into view the various societal actors involved in creating the conditions for an inclusive and democratic conversation about the uses of machine learning. It stresses the importance of technical approaches like explainability and fairness metrics – as well as the fact that in many contexts technical forms of communication are necessary for various reasons – and aims to situate them in a broader context.

Viewing transparency as a communicative constellation is a proposal for conceptually shifting our perspective on what transparency is to achieve and decentering the technical interpretation of transparency that has been dominant in the FAccT field. With this suggestion we aim at opening up new research perspectives and relations between literatures. We understand it as an invitation to the FAccT community to think further along the paths outlined here and to more systematically work on transparency as a research topic. Building on this proposal, more work is needed to consolidate the implications of adopting a broad, deliberative and reflexive understanding of transparency. This includes (1.) systematically reviewing, developing and evaluating concrete strategies for creating transparency, (2.) further extending our conceptual and analytical framework for describing such a broad account of transparency and (3.) strengthening the bridges to the research field of science communication and related areas. We hope that our suggestion will itself prove to be a valuable component of future emerging communicative constellations.

ACKNOWLEDGMENTS

We would like to thank the research group Technology, Power and Domination at the Weizenbaum Institute for fruitful discussions and the anonymous reviewers for their valuable remarks.

REFERENCES

[1] Madeleine Akrich. 1992. The De-Scription of Technical Objects. In *Shaping Technology / Building Society*, Wiebe E. Bijker and John Law (Eds.). MIT Press, Cambridge, MA, 205–224.

[2] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data* 3 (2020), 1–17. https://doi.org/10.3389/fdata.2020.00005

[3] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. Der AMS Algorithmus - Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). https://doi.org/10.1553/ITA-pb-2020-02

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] arbeit plus. 2019. Algorithmen und das AMS Arbeitsmarkt-Chancen-Modell. https://arbeitplus.at/wordpress/wp-content/uploads/2019/09/2019-09_Position-Algorithmus-und-Segmentierung.pdf

[6] Stefan Aykut, David Demortain, and Bilel Benboudiz. 2019. The Politics of Anticipatory Expertise: Plurality and Contestation of Futures Knowledge in Governance — Introduction to the Special Issue. *Science & Technology Studies* 32,

4 (2019), 2–12. https://doi.org/10.23987/sts.87369

[7] Gaston Bachelard. 2002 [1938]. *The Formation of the Scientific Mind. A Contribution to a Psychoanalysis of Objective Knowledge.* Clinamen, Manchester.

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Virtual Event, Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[9] Garfield Benjamin. 2022. #FuckTheAlgorithm: algorithmic imaginaries and political resistance. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 46–57. https://doi.org/10.1145/3531146.3533072

[10] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity, Medford, MA.

[11] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, Barcelona, Spain, 648–657. https://doi.org/10.1145/3351095.3375624

[12] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research* 81 (2018), 1–11.

[13] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 173–184. https://doi.org/10.1145/3531146.3533083

[14] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul Republic of Korea, 850–863. https://doi.org/10.1145/3531146.3533149

[15] William Boag, Harini Suresh, Bianca Lepe, and Catherine D'Ignazio. 2022. Tech Worker Organizing for Power and Accountability. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 452–463. https://doi.org/10.1145/3531146.3533111

[16] Luc Boltanski and Laurent Thévenot. 2006. *On Justification: Economies of Worth.* Princeton University Press, Princeton. https://www.degruyter.com/document/doi/10.1515/9781400827145/html

[17] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. 2022. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 891–905. https://doi.org/10.1145/3531146.3533153

[18] Katharina Braunsmann, Korbinian Gall, and Falk Justus Rahn. 2022. Discourse Strategies of Implementing Algorithmic Decision Support Systems: The Case of the Austrian Employment Service. *Historical Social Research* 47, 3 (2022). https://doi.org/10.12759/HSR.47.2022.30

[19] Mark B. Brown. 2015. Politicizing science: Conceptions of politics in science and technology studies. *Social Studies of Science* 45, 1 (2015), 3–30. https://doi.org/10.1177/0306312714556694

[20] Massimiano Bucchi and Brian Trench (Eds.). 2022. *Routledge Handbook of Public Communication of Science and Technology.* Routledge, London.

[21] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 1–12. https://doi.org/10.1177/2053951715622512

[22] Michel Callon, Pierre Lascoumes, and Yannick Barthe. 2009. *Acting in an Uncertain World: An Essay on Technical Democracy.* MIT Press, Cambridge, MA.

[23] Florian Cech, Fabian Fischer, Soheil Human, Paola Lopez, and Ben Wagner. 2019. Dem AMS-Algorithmus fehlt der Beipackzettel. *Futurezone* (March 2019). https://futurezone.at/meinung/dem-ams-algorithmus-fehlt-der-beipackzettel/400636022

[24] Adriane Chapman, Philip Grylls, Pamela Ugwudike, David Gammack, and Jacqui Ayling. 2022. A Data-driven analysis of the interplay between Criminological theory and predictive policing algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 36–45. https://doi.org/10.1145/3531146.3533071

[25] Bobby Chesney and Danielle Citron. 2019. Deep Fakes: A Looming Challenge for Privacy. *California Law Review* 107 (2019), 1753–1819. https://doi.org/10.15779/Z38RV0D15J

[26] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[27] Wolfie Christl. 2020. Tweet by @WolfieChris on November 17th 2020, 13:09 CET. https://twitter.com/WolfieChristl/status/1328671651521835008

[28] Patricia Hill Collins. 2019. *Intersectionality as Critical Social Theory.* Duke University Press, New York.

[29] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press, Cambridge, MA.

[30] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research* 81 (2018), 1–12. https://proceedings.mlr.press/v81/ensign18a.html

[31] epicenter.works. 2019. AMS Algorithmus - Auskunft gem. §§ 2, 3 Auskunftspflicht G. https://en.epicenter.works/document/2104

[32] epicenter.works. 2022. Stoppt den AMS-Algorithmus. https://amsalgorithmus.at/de/

[33] Elena Esposito. 2013. Digital prophecies and web intelligence. In *Privacy, Due Process and the Computational Turn*, Mireille Hildebrandt and Katja De Vries (Eds.). Routledge, London, 117–138.

[34] Virginia Eubanks. 2017. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, New York.

[35] Florian Eyert. 2023. Mathematical Science Communication as a Strategy for Democratizing Algorithmic Governance. In *Handbook of Mathematical Science Communication*, Anna Maria Hartkopf and Erin Henning (Eds.). World Scientific, Hackensack, NJ, 295–321. https://doi.org/10.1142/9789811253072_0017

[36] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 1457–1466. https://doi.org/10.1145/3531146.3533202

[37] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, Atlanta, GA, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[38] Ben Gansky and Sean McDonald. 2022. CounterFAccTual: How FAccT Undermines Its Organizing Principles. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 1982–1992. https://doi.org/10.1145/3531146.3533241

[39] Ben Green. 2018. Data Science as Political Action: Grounding Data Science in a Politics of Justice. (2018). https://arxiv.org/abs/1811.03435

[40] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, Barcelona, Spain, 19–31. https://doi.org/10.1145/3351095.3372840

[41] Elias Grünewald and Frank Pallas. 2021. TILT: A GDPR-Aligned Transparency Information Language and Toolkit for Practical Privacy Engineering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Virtual Event, Canada, 636–646. https://doi.org/10.1145/3442188.3445925

[42] Jürgen Habermas. 1996. *Between facts and norms: Contributions to a discourse theory of law and democracy.* MIT Press, Cambridge, MA.

[43] Leif Hancox-Li and I. Elizabeth Kumar. 2021. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Virtual Event, Canada, 817–826. https://doi.org/10.1145/3442188.3445943

[44] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599.

[45] Frank Herrmann. 2018. Tweet by @herrfrankherrmann on October 13th 2018, 11:00 CET. https://twitter.com/herrfrankmann/status/1051035016627712000

[46] Jürgen Holl, Günter Kernbeiß, and Michael Wagner-Pinter. 2018. Das AMS-Arbeitsmarktchancen-Modell. Dokumentation zur Methode. http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

[47] Maja Horst. 2008. In Search of Dialogue: Staging Science Communication in Consensus Conferences. In *Communicating Science in Social Contexts*, Donghong Cheng, Michel Claessens, Toss Gascoigne, Jenni Metcalfe, Bernard Schiele, and Shunke Shi (Eds.). Springer Netherlands, Dordrecht, 259–274. http://link.springer.com/10.1007/978-1-4020-8598-7_15

[48] Maja Horst, Sarah R. Davies, and Alan Irwin. 2016. Reframing Science Communication. In *The Handbook of Science and Technology Studies*, Ulrike Felt, Rayvon Fouché, Clark A. Miller, and Laurel Smith-Doerr (Eds.). MIT Press, Cambridge, MA, 881–907.

[49] Sheila Jasanoff (Ed.). 2004. *States of Knowledge.* Routledge, London.

[50] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 702–714. https://doi.org/10.1145/3531146.3533135

[51] Sebastian Kienzl and András Szigetvari. 2018. Jobchancen-Berechnung: Testen Sie einen der 96 AMS-Algorithmen. *Der Standard* (Oct. 2018). https://www.derstandard.at/story/2000089925698/berechnen-sie-ihre-jobchancen-so-wie-es-das-ams-tun

[52] Alexandros Kioupkiolis. 2011. Keeping it open: Ontology, ethics, knowledge and radical democracy. *Philosophy & Social Criticism* 37, 6 (2011), 691–708. https://doi.org/10.1177/0191453711402941

[53] Goda Klumbytė, Claude Draude, and Alex S. Taylor. 2022. Critical Tools for Machine Learning: Working with Intersectional Critical Concepts in Machine Learning Systems Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul, Republic of Korea, 1528–1541. https://doi.org/10.1145/3531146.3533207

[54] Johannes Kopf. 2018. Wie Ansicht zur Einsicht werden könnte. https://www.johanneskopf.at/2018/11/14/wie-ansicht-zur-einsicht-werden-koennte/

[55] Johannes Kopf. 2019. Offener Brief an Fr. Prof. Sarah Spiekermann zum Thema Einsatz von KI im AMS. https://www.johanneskopf.at/2019/09/24/offener-brief-fr-prof/

[56] Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 401–426. https://doi.org/10.1145/3531146.3533107

[57] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35. https://doi.org/10.1145/3359283

[58] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 1257–1268. https://doi.org/10.1145/3531146.3533182

[59] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 2103–2113. https://doi.org/10.1145/3531146.3534628

[60] Paola Lopez. 2019. Reinforcing intersectional inequality via the AMS Algorithm in Austria. In *Conference Proceedings o f the STS Graz Conference 2019. Critical Issues in Science, Technology, and Society Studies*. 289–309. https://doi.org/10.3217/978-3-85125-668-0-16

[61] Giuliana Mandich. 2020. Modes of engagement with the future in everyday life. *Time & Society* 29, 3 (2020), 681–703. https://doi.org/10.1177/0961463X19883749

[62] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta, GA, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[63] Bianca Prietl. 2019. Big Data: Inequality by Design?. In *Proceedings of the Weizenbaum Conference 2019*. https://doi.org/10.34669/wi.cp/2.11

[64] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 1776–1826. https://doi.org/10.1145/3531146.3533231

[65] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 959–972. https://doi.org/10.1145/3531146.3533158

[66] Simone Rödder, Martina Franzen, and Peter Weingart (Eds.). 2012. *The Sciences' Media Connection – Public Communication and its Repercussions*. Springer Netherlands, Dordrecht.

[67] Ina Sander. 2020. What is critical big data literacy and how can it be implemented? *Internet Policy Review* 9, 2 (2020). https://doi.org/10.14763/2020.2.1479

[68] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 1616–1628. https://doi.org/10.1145/3531146.3533218

[69] Pola Schwöbel and Peter Remmers. 2022. The Long Arc of Fairness: Formalisations and Ethical Discourse. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 2179–2188. https://doi.org/10.1145/3531146.3534635

[70] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 2138–2148. https://doi.org/10.1145/3531146.3534631

[71] Steven Shapin and Simon Schaffer. 1985. *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton University Press, Princeton.

[72] Jessie J. Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 587–597. https://doi.org/10.1145/3531146.3533122

[73] Wonyoung So, Pranay Lohia, Rakesh Pimplikar, A.E. Hosoi, and Catherine D'Ignazio. 2022. Beyond Fairness: Reparative Algorithms to Address Historical Injustices of Housing Discrimination in the US. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 988–1004. https://doi.org/10.1145/3531146.3533160

[74] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 2239–2250. https://doi.org/10.1145/3531146.3534639

[75] Sarah Spiekermann. 2019. Warum das AMS keine KI auf österreichische Bürger loslassen sollte. https://www.derstandard.at/story/2000108890110/warum-das-ams-keine-ki-auf-oesterreichische-buerger-loslassen-sollte

[76] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, USA, 1–9. https://doi.org/10.1145/3465416.3483305

[77] András Szigetvari. 2018. AMS bewertet Arbeitslose künftig per Algorithmus. *Der Standard* (Oct. 2018). https://www.derstandard.at/story/2000089095393/ams-bewertet-arbeitslose-kuenftig-per-algorithmus

[78] András Szigetvari. 2018. AMS-Vorstand Kopf: "Was die EDV gar nicht abbilden kann, ist die Motivation". *Der Standard* (Oct. 2018). https://www.derstandard.at/story/2000089096795/ams-vorstand-kopf-menschliche-komponente-wird-entscheidend-bleiben?ref=article

[79] András Szigetvari. 2018. Leseanleitung zum AMS-Algorithmus. *Der Standard* (Oct. 2018). https://www.derstandard.at/story/2000089720308/leseanleitung-zum-ams-algorithmus

[80] Linnet Taylor, Luciano Floridi, and Bart van der Sloot (Eds.). 2017. *Group Privacy: New Challenges of Data Technologies*. Springer International Publishing, Cham.

[81] Lars Tønder and Lasse Thomassen (Eds.). 2014. *Radical Democracy: Politics between Abundance and Lack*. Manchester University Press, Manchester.

[82] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona, Spain, 272–283. https://doi.org/10.1145/3351095.3372834

[83] Brian Trench. 2008. Towards an Analytical Framework of Science Communication Models. In *Communicating Science in Social Contexts*, Donghong Cheng, Michel Claessens, Toss Gascoigne, Jenni Metcalfe, Bernard Schiele, and Shunke Shi (Eds.). Springer Netherlands, Dordrecht, 119–135. https://doi.org/10.1007/978-1-4020-8598-7_7

[84] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review* 123, 3 (2021), 735–790. https://www.ssrn.com/abstract=3792772

[85] Ben Wagner. 2019. Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In *BEING PROFILED*, Emre Bayamlioglu, Irina Baraliuc, Liisa Albertha Wilhelmina Janssens, and Mireille Hildebrandt (Eds.). Amsterdam University Press, Amsterdam, 84–89. https://www.degruyter.com/document/doi/10.1515/9789048550180-016/html

[86] Joseph Weizenbaum. 1966. ELIZA — A computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45. https://doi.org/10.1145/365153.365168

[87] Barbara Wimmer. 2018. AMS-Chef: "Mitarbeiter schätzen Jobchancen pessimistischer ein als der Algorithmus". *Futurezone* (Dec. 2018). https://futurezone.at/netzpolitik/ams-chef-mitarbeiter-schaetzen-jobchancen-pessimistischer-ein-als-der-algorithmus/400143839

[88] Barbara Wimmer. 2018. "AMS-Sachbearbeiter erkennen nicht, wann ein Programm falsch liegt". *Futurezone* (Oct. 2018). https://futurezone.at/netzpolitik/ams-sachbearbeiter-erkennen-nicht-wann-ein-programm-falsch-liegt/400147472

[89] Barbara Wimmer. 2018. Der AMS-Algorithmus ist ein „Paradebeispiel für Diskriminierung". *Futurezone* (Oct. 2018). https://futurezone.at/netzpolitik/der-ams-algorithmus-ist-ein-paradebeispiel-fuer-diskriminierung/400147421

[90] Langdon Winner. 1980. Do artifacts have politics? *Daedalus* 109, 1 (1980), 121–136.

[91] Meg Young, Michael Katell, and P.M. Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 1375–1386. https://doi.org/10.1145/3531146.3533194

[92] Marilyn Zhang. 2022. Affirmative Algorithms: Relational Equality as Algorithmic Fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul, Republic of Korea, 495–507. https://doi.org/10.1145/3531146.3533115

[93] Shoshana Zuboff. 2018. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs, New York.