Proposing a Postcritical AI Literacy: Why We Should Worry Less about Algorithmic Transparency and More about Citizen Empowerment

Eugenia Stamboliev

Special Issue: Critique, Postcritique and the Present Conjuncture

Proposing a *Postcritical AI Literacy*: Why We Should Worry Less about Algorithmic Transparency and More about Citizen Empowerment

Media Theory
Vol. 7 | No. 1 | 201-232

http://mediatheoryjournal.org/

EUGENIA STAMBOLIEV

University of Vienna, AUSTRIA

Abstract

So-called artificial intelligence (AI) is infiltrating our public and communication structures. The Dutch childcare benefit scandal, revealed in 2019, demonstrates how disadvantageous the opacity of AI can be for already vulnerable groups. In its aftermath, many scholars urged for the need for more explainable AI so that decision-makers can intervene in discriminatory systems. Fostering the explainability of AI (XAI) is a good start to address the issue, but not enough to empower vulnerable groups to fully deal with its repercussions. As a canon in data and computer sciences, XAI aims to illustrate and explain complex AI via simpler models making it more accessible and ethical. The issue being that, in doing so, XAI *depoliticises* transparency into a remedy for algorithmic opacity, treating transparency as artificially stripped of its ideological meanings. Transparency is presented as an antidote to ideology, though I will show how this is an ideological move with consequences. For instance, it makes us focus too much on algorithmic opacity, rather than explaining the wider power of AI. Second, it hinders us from having debates on who holds the power around AI's explanations, application or critique. The problem is that those affected by or discriminated against by AI, as in the Dutch case, have little tools to deal with the opacity of AI as a system, while those who focus on data opacity are shaping the literacy discussion. To address these concerns, I suggest moving beyond the focus on algorithmic transparency and towards a *post-critical AI literacy* to strengthen debates on access, empowerment, and resistance, while not dismissing XAI as a field, nor algorithmic transparency as an intention. What I challenge here is the hegemony of treating transparency as a depoliticised and algorithmic issue and viewing the explainability of AI as the sufficient path to citizen empowerment.

Keywords

Artificial Intelligence, Transparency, Explainability of AI (XAI), AI literacy, Post-critique

Introduction: How to turn post-critical on transparency

A growing number of semi-automated but discriminating applications, known collectively as "artificial intelligence" (AI)[1], are permeating private, public, and communication structures. How well do we as citizens comprehend and contest their functions and authority? This question is prompted by a need to respond urgently. Between 2013 and 2019, the Dutch tax authorities utilised a "self-learning" algorithm[2] to develop risk profiles for detecting fraud involving childcare benefits. The results were highly discriminatory.[3] Thousands of caregivers have been forced into poverty, divorce, and even suicide as a result of mistakenly claimed tax authority obligations, leading to many caregivers losing custody of their children. As a result of this scandal, the Dutch government was forced to resign, some of the falsely accused were compensated, and the case moved up to European legal bodies investigating what went wrong.[4]

Exactly *what* went wrong? According to some, the opaque algorithm was to blame. The government employed a highly secretive AI system that was neither visible to those who used it nor accessible to those who were the targets of discrimination. The system was highly biased in its risk assessment, favouring Dutch nationals and penalising non-Dutch or dual citizens. For many scholars analysing this case in its aftermath, it was crucial to ensure that this AI system was explained and made transparent to those using it, before deploying it (e.g., Kuźniacki, 2023).

On the one hand, in order to ensure that algorithms are ethical, comprehensible, and trustworthy (HLEG, 2019; Shin, 2021), AI must be made transparent. Explainability of artificial intelligence, or XAI, is a research field, but also a data-oriented critique of AI's opacity that aims to make algorithms more accessible and understandable (Hoffman et al., 2018; Samek et al., 2017). It aims to increase the transparency and accessibility of AI and algorithms through reorganised and simplified models, texts, and visualisations. By advocating for greater explainability, computer scientists, data researchers, and AI developers have positioned themselves as authorities in the fight against the rise of AI opacity (Liefgreen et al., 2022). Transparency is mostly presented as a non-ideological goal (Samek et al., 2017; Shin, 2021; Chazette et al., 2022) and as an antithesis to opacity (Rosenfeld & Richardson, 2019). Data and computer sciences

might define themselves as apolitical due to the engineering and programming focus, hence this focus is understandable. However, the problem is that what these fields enable (and now explain), is not at all apolitical, nor is the role of the data scientist (Green, 2021).

Given that algorithms are more than technical instructions and integrated into wider opaque systems like corporate agendas (Pasquale, 2015), we need to question the sole focus on explaining the algorithmic realm and treating its opacity or transparency as apolitical. We face many ethical issues on responsibility and trustworthiness around this topic as AI is a mysterious technology while problematically pervasive (Coeckelbergh, 2021), but we lack discussion of the role of transparency as a hegemonic and limited path towards AI empowerment. In the coming years, it will be extremely challenging to comprehend and explain AI in its entirety. It could also be an impossible task. Educators (Klein, 2023) as well as journalists (Diakopoulos & Koliska, 2017) already face this issue.

Algorithms are opaque, but their opacity does not end in their computational features. AI is all about ideology and power (Cave et al., 2020; Bartoletti, 2020), so is the intention to explain it and conventions about how this is to be done. Not all XAI debates are apolitical or uncritical (Knowles, 2020; Singh et al., 2021). Scholars such as Bran Knowles (2020) point out that we have no ways to measure the success of explainable AI and most people have little opportunities to realistically stop AI (3). I agree. In addition, I also see value in talking about transparency; it shapes our awareness around AI being opaque, but it just cannot provide the transparency we would need as a society. Is transparency the wrong goal or a limited goal? Both. Focussing *only* on algorithmic opacity seems wrong, focussing *only* on transparency seems limited.

In what follows, I will examine how transparency might be *depoliticised* in the process and canon of explaining AI, known as XAI, and show why this is a problem for AI literacy approaches. By focussing on algorithmic truths and explanations given from a technocratic authority, I fear that we will overlook the ideological assumptions embedded in algorithms (and AI) and also displace discussions of how to empower citizens in dealing with these new power structures infiltrating their lives. A growing depoliticisation is not simply apolitical, but ideological because it wrongly limits the

concern of a wider opacity in techno-political decision-making to one of opaque algorithms (even if XAI does not ignore the social effects of algorithms). I use the term depoliticising also in relation to "the political" as an antagonistic force. According to Chantal Mouffe's work (2000), "the political" is a dynamic and antagonistic force, crucial for a democracy but often stifled due to a reactionary understanding of pluralistic discourses. As a force, it draws from public engagement and opportunities for counter-hegemonies and dissensus. However, I mainly apply it to emphasise the artificial removal of transparency from its ideology and from enabling a societal and antagonistic potential to shape counter-hegemonies to XAI as top-down discourse and to algorithmic truths as computational ambitions.

Transparency might be depoliticised in XAI, but this in itself is an ideological move (Birchall, 2014; Valdovinos, 2022). How I use ideology should not be overloaded with a political meaning here, but mainly pointing to an unquestioned, yet implied, data authority system that dominates the XAI debates. Ideological also means that transparency is indirectly used to divert from the ideological and hegemonic manifestations and dispositions, and to reduce transparency to its functional-instrumentalist orientation (Owetschkin et al., 2021: 5). This is a strategy that is presented as having eliminated ideological barriers and neutralised the agenda by revealing their *pure* intentions (Birchall, 2014), and can only be understood by looking at these aspects as ideological but not yet contested (Valdovinos, 2022). The hype around AI and its complexity plays into an algorithmic hegemony as it gives more room to the algorithmic issues (while reducing them to engineering ones). This distorts the view on why we need transparency and how far our focus on opaque algorithms can take us. For instance, OpenAI's chatbots (using neural networks, which can be argued as AI) seem very accessible and democratised, but their algorithmic underbelly is deceptive in terms of sources, algorithmic instructions, reinforcement strategies, and ideology; hence, chatbots are opaque on various levels not making their algorithmic opacity less of a problem, but not the *only* problem.

Depoliticising transparency will not make AI less opaque, but less accessible. It keeps us from having more conversations about how to equip individuals with more accessible, critical, and inclusive AI literacies. The Dutch case study I refer to here shows how the use of AI in governmental platforms automatises welfare management

and tax matters, not only cutting out most of the human oversight, leaving it to a few to comprehend and control the societal effects, but also showing that AI systems are opaque for decision-makers and those implementing AI, creating terrible consequences for those affected.

Below, I will follow a two-fold structure. In the first section, I highlight and critique current objectives and issues in XAI; in the second, I suggest ways to reconstruct a broader notion of literacy via questioning the authority and framework of XAI, revisiting the history and canon of media literacy, emphasising citizen empowerment. In neither case am I dismissive of the intention to aim for more transparency; rather, I challenge the dispositions, frameworks, and authorities to enable a better path to make AI more transparent for *all*, not only for a few. Hence, I will approach AI literacy as a *post*-critical response to XAI, not its dismissal.

The *post-critical* is not an anti-critique, nor a move beyond critique. I try to resolve a tension between moving beyond the limits of transparency as a solution while not dismissing it as an objective to face opacity. Adopting Rita Felski's work on post-critique (2015), I will show how a *post*-critical AI literacy allows me to unpack additional sensibilities that fall short in XAI. I mainly use the *post*-critical to address what goes "wrong" with XAI as a critique, not as a method to dismiss XAI. Similar to Castiglia's remark on the limits of critique (2017), I equally assume that "something is wrong" with how transparency is used as a remedy to fight AI's opacity. I am not against the desire to fight opacity, but against unquestioned dispositions about what deserves transparency as some "dispositional change is necessary for critique to get a second wind" (Anker & Felski, 2017: 226). I challenge the view that explaining or educating about AI has to be done in the present framework of data or computer sciences, the framework we use to get there, who delivers this information, and what they promise. The main concern I have is that transparency is used in XAI to artificially erase power relations, while limiting the spaces for antagonistic debates (Mouffe, 2000: 18), feeding an illusion that we democratise AI by explaining it (*we are making it more transparent, aren't we?).* My post-critical turn in discussing AI literacy does not aim to develop a *new* or *meta* critique of transparency or media literacy, but rather to challenge the disposition that algorithms are the only level that matters, or that it is a level that can be explained in a straightforward, non-ideological way. It is an attempt to stir up the ground we

stand on when it comes to AI; one that minimises the gap between people's everyday experience with AI and the intellectual and academic discourse around it; something a post-critical view tries to minimise (Habed, 2021: 500). The difficulty I see lies in explaining or educating about the opacity of AI (and similar) and to do justice to all parties involved.

Experts must understand AI, but so do laypeople and those who will be affected by it. By challenging the technocratic hegemony, I also challenge the common attitude traditional critiques share on the "political" being disempowering (Felski, 2015: 212). Aiming for an AI literacy that is access-focused can operate in parallel with XAI while building on the learning curve and past errors of media literacy. Despite being ambitious in pointing out that transparency is used as a limited concept, the shift towards *post-critical AI literacy is not* a *solution or denial of the critique of opacity or the value of XAI as critique*, but *a post-critical AI literacy aims to empower those affected by AI by debating access to AI as power, not AI as algorithm*. It is a first step in expanding on an urgent debate on who is allowed at the table when we develop, implement and *explain* AI. Still, the main element in calling my AI literacy shift *post-critical* is not to dismiss the value of critique but to "reconstruct" the critical engagement (Felski, 2015: 212) around what deserves explaining, and by whom.

Section One: On the good and the bad about transparency

What is explainable AI? Explainability is a not well-defined, but vastly growing, canon dealing with the computational and data-oriented side of AI. According to Larissa Chazette and colleagues (2021) and Wojciech Samek and colleagues (2017), the use of transparency is crucial to most debates. When it comes to neural networks like Large Language Models (LLMs), algorithms are opaque and become increasingly difficult to comprehend as they become more complicated. According to David Gunning and David Aha (2019), the goal of XAI is "to create a suite of new or modified machine learning [ML] techniques that produce explainable models that, when combined with efficient explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems" (45). The European Union's ethics guidelines for trustworthy AI mention "explicability as an ethical

principle of trustworthy AI and transparency as a way to achieve it" (Amann et al., 2022: 12), despite the fact that the terms "explicability" and "transparency" are not clearly defined.

Explainability and transparency are related in three ways, according to Chazette et al. (2021). First, to lessen software opacity, which allows for users to comprehend the decisions made by algorithms better. Second, in order to minimise a possible frustration, and third, to influence the connection of trust and reliance on the system (197). The difficulty in engineering common ground qualities and in using catalogues and models is the key level that is addressed here. The latter in particular seems to provide a chance "to ask stakeholders about their interest with respect to the general system quality" (197). Because it is difficult to "elicit, negotiate, and validate (...) due to the subjective, interactive, and relative nature" (197), they note that explainability can have both positive and negative effects. Chazette et al. (2021) present a valid discussion, but my concern is that their focus is too practical, while not ideologically critical enough. The problem, I say, reaches beyond their paper; while mainly aware of the complexity behind algorithmic values and embedded interests, most XAI discussions do not engage more deeply with how they use transparency as a solution or as a promise.

Transparency acts as a remedy deprived of any ideology, while only aimed at the computational opacity of AI. With regard to one problem – the opaque algorithm – transparency becomes a target, a solution, and a practical one. For example, Avi Rosenfeld and Ariella Richardson (2019) make the following point: "Within the Machine Learning/Agent community, transparency has been unofficially defined to be the opposite of opacity or "blackbox-ness"" (677), but neither the black box nor the notion that transparency might not be the opposite of opacity are contextualised. Even when discussing the epistemology of transparency, Juan Durán and Karin Jongsma (2021) focus on reading the sequences and variables to understand algorithmic decision making. There is no mention of any non-engineering issues; transparency is not contextualised as a legitimate path. But this could go against what Durán and Jongsma aim for, to provide evidence that algorithmic decisions are reasonable and not only reliable (330). Even if XAI might use transparency as a valid response to the

growing obscurity in the field, since algorithms like neutral networks are inaccessible universes, this is no neutral path even if it presented as a pragmatic one.

While transparency is used as if deprived of ideological meaning, it does not yet exist in a computational vacuum in these debates; rather, it is frequently employed to increase societal trust in algorithms (Shin, 2021; HLEG, 2019; Robinson, 2020). For example, Dhongee Shin (2021) asserts that trust is a combination of explainability and causability (not causality), which implies that explanations are equally crucial to user demands and wants. In his critique, Shin urges AI developers to consider user requirements and emphasises the significance of developing and implementing metrics that guarantee we can "best use AI to help users and give better insights" (7). Shin broadens the transparency debate by addressing how AI and algorithms effect society and the need to understand "how to avoid unfair and discriminatory situations" by asking "how to balance the need for technological innovation and the public interest with accountability and transparency for users" (7). His perspective on trust continues to be based solely on algorithms, however, not unpacking any ideological or political meanings. Cory Robinson (2020), who mentions the cultural values ingrained in algorithms, brings the issue of transparency closer to the forefront of consideration. According to him, it is "pretty clear" that technology trust is a prerequisite for digital trust (2). Whether or not this is evident, I understand his line of reasoning, which is that people do not trust algorithms because they are opaque. I agree that algorithms and machine learning are "technologies that are hidden from citizens' view" and produce "a 'black box' effect" (Robinson, 2020: 2), but I am not sure how we will obtain the values out of the algorithm he refers to as a "black box".

Even if transparency is insufficiently dealt with in most XAI debates and approaches, XAI has faced certain critiques of its methodology. For instance, there is critique of how explanations can be faulted when not providing enough justifications (Knowles, 2020; Singh et al., 2021), referring to lacking alignments with trust-building or to random measures of success. Not only do the gaps for situating the scope for explanations grow, but there is little awareness of the fact that trust and transparency do not always correlate with one another. Caroline Reinhardt (2023) has provided substantial research on this matter with an overview of how AI research works with trust and trustworthiness as parallel, but often conflated, concepts leading to different

debating canons. For instance, she considers the risks of "false trust, misused trust, the perils of trust, and (productive) distrust" as potential problems. (5) Additionally, Weller (2019) intervenes, arguing that having too much information on factors like ethnicity could very likely encourage discrimination, not as a mistake, but supported by the algorithm (12). What we take from this overview is that aiming for transparency makes sense, but the focus on algorithmic transparency as a promise to shine a light onto AI as a system is limited, creating unresolved tensions. On the one hand, we need to work with transparency somehow – aiming for less opaque algorithms is a valid goal. On the other hand, aiming for less opaque algorithms cannot be the only goal in terms of transparency, as by keeping only this mind we support a techno-determinist version of transparency. This alignment of transparency to algorithms seems removed from the wider ideology behind transparency while, in fact, this very reduction is an ideological move. I will contextualise my concerns with transparency and its ideology in the next section.

The transparency delusion: On why making algorithms transparent is an ideological choice on what to conceal and what to reveal

Transparency appears to be a common-sense idea. Who would *want* opaque systems for governing technology? This seems appealing but it hides its ideological core through its wider agreeability. Assuming this statement as true and unproblematic suggests we can provide unbiased information and bring a new objectivity in the complex setting of AI or algorithmic power. When used as a "keyword" (Valdovinos, 2022), transparency acts as a superior type of disclosure. Transparency presents itself as an objective strategy for addressing opacity, however it is founded on a cultural desire to purge information of the ethical, legal, and epistemological problems associated with conflicting agendas (Birchall, 2014). I walk through the criticism already addressed in transparency research without fully summarising it, to address the flaws of a pragmatized stance on transparency (Valdovinos, 2022; Birchall, 2014). According to my interpretation, 'transparency' is not about trying to understand the origins of algorithms (although scholars do make an effort to do so), but rather the signifier acts to hide some of the more challenging opacities, like power or impotence, which XAI might not necessarily address but should.

By *wrenching* transparency through its existing interdisciplinary critique, I identify three problems with depoliticising transparency in the present XAI discourse. They are issues directly affecting how we inform and educate about AI and whom we empower.

First, we cannot treat transparency as a remedy to opaque algorithms. This depoliticises its meaning by suggesting that once we reveal the inner workings of algorithms, we can democratise AI or make algorithms more neutral and accessible. The illusion of this being possible is bound to the common-sense use of transparency; it comes in handy.

> Although it is obvious that the term transparency has a variety of applications, its tight semantic relationship to other words (such as openness, clarity, intelligibility, insight, obviousness, accountability, access, participation, etc.) raises the possibility that we may be dealing with more than just a straightforward "word" here (Valdovinos, 2022: 44).

What is highlighted here is that transparency acts as a chain of associations and as a promise to deliver *more*. More insight, more clarity and more information implying this as a straightforward process that removes all uncertainty and opacity. This narrative is problematic as it maintains the "functional-instrumentalist orientation" of transparency (Owetschkin et al., 2021: 5) while bearing the risk of detracting from a more encompassing transparency we might need on AI as an algorithmic power. In the complex setting of AI or algorithmic power, transparency is reduced to only supplying unbiased information and bringing a fresh perspective, acting as a superior type of disclosure. But transparency is more, it is a strategy for addressing opacity but also founded on a cultural desire to purge information of the ethical, legal, and epistemological problems associated with more disparaging versions of transparency (Birchall, 2014). Transparency is contextual and a trade-off between interests on what should remain opaque. According to Valdovinos (2022), the ambiguity and fluidity surrounding words like "transparency" permit many readings and perceptions of its purpose. Therefore, suggesting more transparency as a governmental remedy is never just an idea; it is an ideological and often unfulfilled promise.

And, still, disclosure is a good thing and possible. For instance, in the Dutch case study, simplifying the system and explaining the crucial decision points might have helped

those deciding to decide *better*. It might have stopped the system earlier. I can only speculate. The issue is when algorithmic transparency begins to outsource ideology as if this would be possible, or as if we have an issue with algorithms being *broken*, which invokes the promise that we can *fix the algorithm*. In every aftermath of flawed or discriminatory algorithmic systems, we urge for more transparency of "biased" algorithms, but in doing so, we only legitimise the impression that flawed algorithms are devoid of ideology, and only *broken* (Valdovinos, 2022). Hence, as popular and useful as this viewpoint might be, it wrongly assumes that we can fix or reveal algorithms (while still not justifying their use and intentionality enough). This pragmatic idea builds on the hegemony of fixing algorithms without fixing their wider context of application, as if they are distinct from it. Once viewed through an ideological lens, it is not about fixing, but about concealing the motifs and effects of how different actors will employ transparency differently (Valdovinos, 2022). This concern shows in another ideological aspect besides the fixing; the *democratising promise*. Transparency suggests that we can also make information more broadly and democratically accessible. AI is not democratic in that sense, as it is an expert technology, but explaining it would make it more accessible, ideally. But practically, this only works if democratic access is about power and not technical insights. There is nothing wrong with providing technical insights. I do not want to sound technophobic, but only providing them is an ideological decision linked to the pseudo-democratisation of power that transparency commonly implies (Berger et al., 2021).

Second, the attempt to make algorithms *fully* transparent is a way to depoliticise transparency. This shows in terms of revealing some imaginary core of the algorithm, or trying to materially/logically access the algorithmic decision process (while potentially not looking at the agenda behind it). The depoliticising aspect I rather want to emphasise is that of exclusivity and focus. One main aspect is mirrored in the question of Marilyn Strathern (2000): "what does visibility conceal?" (310). And this question is key to understanding why us focusing on algorithmic truths is not simply one sided, but by focusing on algorithms, we might even distract from there being another side that remains opaque, and this might not be coincidental. Further, mistaking transparency for visibility is misleading. Transparency is not simply about *more* visibility, considering that algorithms are not visible but inaccessible to human vision, yet this does not mean that algorithms are strategically concealed from our sight

or knowledge. We face two forms of non-transparency; an aesthetic inaccessibility, on the one hand, and someone's decisions on what not to explain or share about algorithmic power, on the other hand. It might not be possible to visually or logically access "algorithmic thought" or the deep layers of neural networks (Fazi, 2020), but this is not an excuse to overlook their ideological input/output. I say that while the first should not worry us that much, the second should worry us more. Neither in the algorithmic realm are all the ideological perspectives ever fully accessible or autonomous, despite indications to the contrary.

Looking at algorithmic transparency from a media-material critique (Parisi, 2013), transparency might act as a promise towards making algorithms more visible but it struggles with the impossibility of *seeing* algorithms as computational realms. This way of using transparency, common for XAI but overcome in most media-material studies, builds on the outdated, but popular, assumption that seeing is the same as knowing. This pre-algorithmic trend is deeply embedded in the century-old ideology and hegemony of the visual that has been "broken in the algorithmic culture that runs our social and political lives," according to Zylinska (2020: 94), but XAI has not yet faced this critique. Here again, transparency promises a democratisation of an otherwise *clouded* knowledge, it promises to reveal the invisible, while fostering the concealment of other information which remains unknown and unseen. In addition, it is likely that XAI draws from another outdated assumption of data sources being truly objective and raw, which is not only contradicted by scholars such as Lisa Gitelman (2013), but brings up the issues with how transparency acts as an ideological purification of algorithms from a human contamination.

Besides there being material limits in accessing algorithmic thought per se, due to the different systems of logic (Fazi, 2020), transparency has a revealing and a cleansing role. I draw on Birchall (2014) to illustrate the underlying mechanism, relating it to how transparency is applied in XAI. According to my interpretation, what XAI assumes it is doing is to "cleanse" data of both technological riddles and human imperfections and errors, one of which is the human inaccessibility of algorithmic aesthetics. What would a clean-up enable? Making algorithms more transparent should – according to this hegemonic discourse on algorithmic truths – result in less political prejudice and more fairness (Birchall, 2014: 77). This has to do with the distinction

between what Birchall refers to as "hard" or "neutral" facts and what the desires and morals of individuals "contaminate" (78). However, neither the algorithm nor the preferences of people are neutral or unbiased. There is no getting to the core, as algorithms and AI are rather networks of relations than a line to trace. Having a transparent algorithm is therefore impossible because it makes the false promise of returning to an imaginary core or ground that does not exist (even if this should not be seen as excuse to avoid justifications). The ultimate purpose of XAI is to demonstrate AI's operation so that we can understand what it is doing. Even if AI were to become more democratic (Bartoletti, 2020; Sudmann, 2019), society needs to be able to access its fuller scope; from intentions to design to critique.

Third, we encounter still another repercussion of this depoliticised narrative: Algorithmic knowledge is aimed as its own experts, excluding everyone else or requiring for others to become an algorithm expert. We can already see this reflected in the literature on AI literacy (Kong and Abelson, 2022). Explainability automatically becomes significant solely to individuals who comprehend algorithms once we introduce transparency regarding algorithms. "The target of XAI is an end user who depends on decisions or recommendations made by an AI system, or actions taken by it, and therefore needs to understand the system's justification", claim Gunning and Aha (2019:45). For XAI, this refers to professionals, but to me, it also must refer to other groups, such as people who are impacted by AI. Even though Ramya Srinivasan and Ajay Chander (2020) indicate that far more stakeholders must understand how AI functions, the language is still overly complex and authoritative from the top down. They disagree on who deserves explanations, but not on whether they are valid. To find the correct explanation (4812) and form the right instruction (4815), it is crucial to consider how algorithms and AI influence society. However, we have not sufficiently discussed who, aside from developers or political deciders (like tax authorities), deserve to know this.

What does it mean to know more about AI? When we take the ideological side of transparency into account, we might be able to see how discussing AI as an expert debate (even if not intended), excludes those who will be affected by AI from understanding and from taking action.

Limiting transparency to an algorithmic issue, therefore, wrongly limits the concern of a wider opacity in techno-political decisions to one on opaque algorithms (even if the latter can include debates on the effects and impacts of algorithms, not only their operationality). I showed how the discourse around explainability can depoliticise transparency and why it simply masks the ideological elements in XAI like their focus and authority. Next, I will move to linking these insights back to the Dutch case study and addressing new means to shape AI literacy. I want to move *on* but also move *with* (a wider angle on) transparency towards a framework of AI literacy shaped by post-critical views on access, inclusivity, and empowerment. My attempt to shift to an AI literacy is more a post-critical move that pays attention to an integration of this critique with issues of empowerment, and less to create an intellectual or technocratic model that can suit experts.

Section Two: Moving from *more* transparency to an *empowering* AI literacy

Revisiting the Dutch algorithm scandal: On explaining away a power disbalance

Returning to the Dutch tax authority algorithm, known as the "toeslagenaffaire"[5], opacity and inaccessibility in AI can have severe consequences. What went wrong was that authorities not only developed a flawed Risk Classification Model (RCM)[6], an algorithmic decision-making system designed to detect childcare allowances tax fraud, but that it was extremely discriminatory and that no one interfered for several years. How did this system (mal)function? The RCM was a self-learning algorithm that created risk profiles for potential applicants while identifying those who are more likely to provide false information and maybe engage in fraud. However, the system began incorrectly labelling caregivers based on their citizenship and nationality, implying inappropriate government action, such as strict interpretations of the law and ruthless policies. The primary issues were the accusations made and the subsequent demands for repayment, which could not be contested by many.

In this escalating situation, explainability models or approaches would have been essential for the developers and decision-makers to step in at an early stage. I have no

reason to dispute that, and no information on why these were missing. Tax authority employees using and deploying this system were unaware of the algorithm's decisions and were unable to verify them. This was further complicated by the time it took to notice structural errors due to the different intuitions involved. Hence, I agree that it is important to know why AI chooses the decision path it does, as Błażej Kuźniacki (2023) puts it, further saying that we cannot reason officially that the AI told us to impose tax on someone. That might result in unfair treatment. Tax authorities cannot properly defend their conclusions unless they can adequately explain them. Algorithms, for example, cannot totally or even primarily replace human trust.

Importantly, there must be a person with decision-making authority and knowledge of AI's internal logic. The incident has shown us what happens when a procedure is too hidden and computerised. The use of non-legally significant information in decision-making by AI is supposedly possible, including sex, religion, race, and address. Furthermore, we must keep in mind that privacy is in competition with transparency; that which we make public is no longer private. In this instance, the Dutch tax law secrecy, for example, made this algorithm even more opaque by prohibiting the disclosure of personally identifiable information. The fact that legal restrictions competed with algorithmic explainability resulted in its obscuration, but this was not an error; rather, it was a conflict of interests between institutions. However, according to Weller (2019), disclosing personal information can still encourage even greater bias and discrimination through algorithms. There is no simple equation here; having more transparency is never a good motif in and of itself, especially in balancing it with privacy.

In the aftermath of the Dutch welfare fraud scandal, the absence of human oversight came under fire. Why was no one noticing this issue? And how come those using the algorithms were not better informed? Retrospectively from a scholarly perspective, it is difficult to identify what went wrong and who unintentionally pressed the *wrong* button. But what we know in the field of AI ethics is that this dilemma is a common responsibility and accountability issue when discussing the consequences of a faulty AI design. Mark Coeckelbergh (2020) speaks of the "many hands" and "many things" problem, which says that once an algorithm or AI is implemented and malfunctions, it is difficult to assign responsibility retrospectively to a single point or

person. This can be misappropriated to get out of responsibility as well. In the Dutch case, this point about responsibility is important but was resolved in one way, by having the most responsible authority, the Dutch government, step down. While a valid consequence, it lacks much meaning in the big picture; to empower and foster literacy for those affected by these systems. However, authorities investigated, claiming that there was too little human oversight when applying this algorithmic system. How little human monitoring this system had is unclear, though. The European AI Act, which sought to ensure human oversight in complex and risky AI systems, would maintain that having none is not acceptable (EC, 2022). Further, it came to light that the government had promised rewards for uncovering fraud. This might have influenced the current oversight, making it significantly biased and less inclined to act. While discrimination is abstract and a question of knowing the code for people who program AI and make decisions, those who are harmed are powerless to stop, alter, or intervene in this code. The childcare benefit controversy underlines the need for transparency, but it also demonstrates how powerless people are in the face of these structures.

Nevertheless, as Virginia Eubanks (2018) notes, the problem with automating discrimination is not only that it is opaque, but also that there is no mechanism to intervene and stop it. On many levels, it keeps individuals in a state of powerlessness. Therefore, for individuals impacted by this algorithm, XAI would not have altered much (but I am speculating here). What I do not speculate about is that the authority in explaining AI is with the data sciences, which are still refusing to take on full responsibility as a political actor and force (Green, 2021). As a result of AI's opacity, discrimination is occurring more covertly than people are aware of, and while this is addressed increasingly, it is not discussed as a problem of access to AI or application but as an unfortunate consequence. Denying access is the first sign of impotence, hiding it qua depoliticised transparency debates only fosters this shortcoming. As Eubanks puts it, "Although we all live in this new digital data regime, not everyone has the same experience with it. The access to information, free time, and self-determination that middle-class professionals frequently take for granted made my family's experience bearable" (2018: 4). Although her book laid the foundation for the discussion of inequality and AI, it has not yet influenced discussions on XAI to consider this access issue rather than just discussing the technology.

How can we strengthen our veto power or gain a better knowledge of these algorithms or systems? One approach is to investigate the evolution of media literacy and education, and focus less on making things transparent, but more on how we can act upon it.

A postcritical AI literacy: From explaining the algorithm to empowering citizens (ideally)

Can AI literacy help individuals interact with opaque AI systems? Let me discuss the *ideal* literacy constellation while recognising that my comprehension of access is limited to people being able to use their knowledge to judge the consequences and implementation of AI. First, there is no perfect literacy model to use or design, but we may learn what to avoid and what to do better from media literacy literature. Second, while literacy could be regarded wrongly as a depoliticised response free of ideology, just like transparency, it is not. I would like to rephrase the subject, but I do so with care because I am aware of all the potential problems, like diminishing or simplifying. When it comes to pitfalls, my focus will be on learning from media literacy.

Me shifting to an AI literacy is a *post*-critical attempt, not an anti-critique. It allows me to unpack additional sensibilities that fall short in XAI. I mainly use the *post*-critical to address what goes "wrong" with XAI as a critique (Castiglia, 2017), by contesting the XAI framework and authority. It is an attempt to challenge the present framework we operate in when discussions around AI literacy are mostly based in the data sciences, which fosters valid expert discussions, but which might widen the gap between people's everyday experience with AI and the intellectual and academic discourse around it (Habed, 2021: 500).

The main concern I have is that transparency is used in XAI to artificially erase power relations while limiting the spaces for antagonistic debates (Mouffe, 2000: 18), feeding an illusion that we democratise AI by explaining it (*we are making it more transparent, aren't we?).* Hence, revisiting media literacy allows us to think of literacy as more political and more inclusive. Media literacy and education got caught up in a similarly pragmatic battle: that of teaching media as a pragmatic skill, not a techno-political power system. However, it had much more time to reflect and yet, not all would agree that this

learning curve is successful. In this context, we see the emergence of transparency equally as operational and not complex enough to empower individuals; linked to skills training and rationalising AI to a manual of neutral choices we can simply learn. Teaching skills seems to be the pragmatic solution to a complex problem (impotence to deal with AI), and it is advertised as such. EU governments have emphasised the importance of digital, informational, and media skills as "basic skills" for individuals (De Abreu, 2022: 35). However, media literacy may overpromise in producing critical media citizens who know the difference between truth and falsehood while coding their way to happiness. The definition of media (and AI) literacy matters, with definitions ranging from tautological education to an idealistic version of cultural ideals. Livingstone (2004) emphasises the relationship between textuality, competence, and power that sets those who see literacy as democratising and empowering ordinary people against those who see it as elitist, divisive, and a source of inequality.

Media literacy education has been critiqued for mainly offering technological training and tautological training by suggesting, for instance, school curricula frequently emphasising technological skills (Buckingham, 2018: 4); something XAI is about to continue and advance (Kong & Abelson, 2022). This is not wrong per se, but it has downsides. Despite recent advances, Alfonso Gutiérrez-Martín and Alba Torrego-González (2022) warn that media education can be reduced to developing technological and instrumental digital competence. If media literacy focuses solely on technical knowledge, operating procedures, and device or software operation, it may neglect the impact of ideologies, attitudes, and values on critical thinking. Yet we see new trends. David Buckingham (2019) highlights the importance of media literacy education, but he warns in his wider research of what I fear happens because of the depoliticisation of transparency in XAI; we overlook the rhetorical construction of literacy as a common-sense "good thing" (Buckingham, 2018: 29), and do not question the authority of those teaching it and to whom (2018). Critical attempts around media literacy, aligned to the work of Buckingham (2018; 2019) and de Abreu (2022), try to accommodate a more democratic school curriculum and a less authoritarian top-down mentality. In times of disinformation, politics, inequality, and authority crises, the context of new technologies and media (including AI) became critical, as does media education. Media literacy understands that opacity is about a political entanglement of

technology and discourses, but it has also realised that the political is hard to teach or address in education.

What can we learn from media literacy? Mainly, that we have frameworks which tapped into the depoliticising trap around transparency but also around how political technology is taught or not. We can learn to pay more attention to empowering citizens to engage with AI and its implementation. In addition, as Buckingham (2019) mentions, it will be key to enable citizens, and not only experts, to question the authorities of those explaining to them as this is an ideologically charged process, not a given setting. Shifting to a post-critical AI literacy is an attempt to question the disposition around AI expertise and audience, one that minimises the gap between people's everyday experience with AI and the intellectual and academic discourse around it (Habed, 2021: 500). Hence, it aligns well to being a "framework through which digital civic discourse can take place" (Baldi & Seraydarian, 2022: 211). To clarify, I do not position political or technical abilities against one another; I think that XAI can be integrated into this literacy, hence, the post-critical is a "reconstruction" of critique not its dismissal (Castiglia, 2017: 212).

How should I go about *reconstructing* a new literacy model? My intention is not to provide a full check box, but to pitch in first steps. One first step is to demand that citizens get a seat at the table. The second is to have a chance to organise more inclusive "hubs" to foster more antagonistic discourses and/or oppositions. Both are not only post-critical responses to XAI but ways to transform the depoliticised application of transparency into a politicised literacy model. Let me unpack these two points in more detail.

First, people should be able to access and evaluate AI decision-making and critique AI as a power structure by having a basic understanding of its design, purpose and effects. This applies to public service systems in particular as it might be harder to demand for corporate AI. Now, is this suggestion too simple or too complicated? Perhaps both, but we have such a strong hegemony in focusing on algorithmic issues surrounding AI that we overlook the citizen perspective. In the scandal involving the Dutch tax algorithm, an early public debate, including something like a public vote, or even a test run evaluated by citizens, could have changed a great deal for the better, and could have stopped this system in its tracks earlier than it did. But the debate over

transparency, which involves experts, demonstrates that citizens are rarely the focus of technocratic approaches. Furthermore, I see a direct link to the depoliticisation of transparency, which pushes not only ideology but also citizens' problems out of sight. Why? If, for example, AI systems were up for a vote or a public discussion, citizens would make the decisions in collaboration with government officials. As an illustration, efforts to involve citizens more in local climate regulations are made and are successful. As an example to show that this is possible, the Berliner Klima Bürger:Innen Rat (BKB) is a citizen-led committee that discusses and implements climate laws and regulations which they decide and discuss collaboratively after being selected in a lot system.[7] This is one strategy to consider when implementing artificial intelligence into wider democratic systems. Still, I am not sure I can advocate for a total shift in voting systems or a transformation to participatory democracy at this time, but I do advocate for more participatory mechanisms in deciding over AI and more points of intervention and discussion. This would be one way to politicise this discourse while risking an antagonistic discourse (Mouffe, 2000), which is key for having a democratic plurality, but might not be easy to manage if it turns out that citizens do not support AI innovation or implementation goals. But a potential disagreement and opposition cannot be a reason to keep people impotent on these matters.

Now, having said we should question the hegemony of data sciences or of algorithmic truths, who would govern this kind of literacy? Would this require independent panels? Do we not have enough of those? I cannot answer this question, but I can say that, ideally, diverse groups should come together to avoid a technocratic bubble. Maybe this can work through facilitating more 'info-diverse' groups like academics, citizens, designers, and politicians, maybe by making sure that algorithms are not presented as the doom of humanity? And maybe by making sure that company agendas do not sneak into governmental structures? For sure, not by overlooking that society's interest in this matter differs. The social fabric conceals complex algorithmic applications, but society is not a single entity; it is divided into groups with competing interests. This will only become acceptable as long as consideration is given to this heterogeneity. The problem is that there is no discourse on whether or not AI should be implemented in government structures; there is regulation, but it is frequently focused on platforms and profit-oriented AI (EC, 2022).

Second, AI literacy should facilitate stronger resistance alliances between journalists, grassroots activists, academics, and the general public. This suggestion may seem 'out of the blue' after discussing only the academic discourse, but I am convinced that not only literacy models, but also XAI should consider this outreach (and might even do). Not only does my own research benefit tremendously from wider alliances of a few journalists, which enables a much quicker overview, or better insights on, for example, problematic labour and business agendas of OpenAI's agendas (Perrigo, 2023), but this synergy appears crucial given that access to AI is neither equalised nor complete. Without journalists or activists sifting through information more quickly and thoroughly on some issues, such as exploitation or unethical business practices, the work of AI scholars seems to be increasingly impossible (Vincent, 2023). Combined, critical groups may be able to influence more discussions. Since the use of transparency appears to be a hegemonic tactic to keep the focus on algorithms rather than power, I suppose this brings us back to creating counter-hegemonies (Mouffe, 2000), which I did not discuss to avoid a theoretical overload. Nevertheless, I view this as necessary given that the hegemonic use of transparency is depoliticising the AI debate.

Let me give one example of why and how these synergies matter. Just as journalists and media outlets first addressed bias and inconsistencies in the Dutch algorithm, so was the "AMS Algorithmus"[8] in Austria stopped from implementation because of wider protest from media, NGOs such as epicenter.works, and institutes like the ÖWA (Austrian Academy of Sciences, 2020). This complex algorithmic system was aimed to streamline and ease the recommendation of trainings and job offers, and although the authorities did not intend for those algorithms to discriminate or exclude, the early test phase showed that the system did (especially against women and people over 50). However, this system was never put into use due to a vocal opposition from the public; instead, it served as a case study for researchers looking at XAI bias and its limitations, including mine. Citizens might become 'illiterate' in using their democratic rights when there isn't a coordinated and widespread politics on the issue. It must be essential to promoting public discourse or participation. Yes, for these, we need XAI to open the "black box" and explain the inner workings, but we also need academics, journalists, activists, civic organisations, and affected groups. Who is to make this possible? How? I must admit, I do not have the solution nor answers at this point. Two things I would like to add, however. One thing is that once we teach the complexity of AI in its

breadth, we need to include media studies, social and political sciences and involve them in shaping AI literacy, which means to look beyond data and computer sciences hegemonizing this discourse. The other is that the timing matters in this context. *When* are we debating AI? Why only when things go wrong? It would be much more difficult for people to complain to or get involved in a dynamic techno-bureaucratic machine once they have been impacted by an algorithmic system, though. If we start *prior* to the AI application, we might change or dispute more collectively, despite the different agendas that might emerge. Still, we might have more power to halt or intervene if various societal groups discuss or reject specific applications in their test phase. Although I wish not to sound naive, this opportunity requires political and governmental will, but I do not think that this will is necessarily missing.

Concluding this section, I only spoke of the *post-critical* as an attempt to reconstruct the opacity critique by expanding on it. I have not found *the* solution to AI opacity, but I emphasise the danger of both depoliticising and relying too much on transparency. Skipping the ideological talk is only masking it, not resolving any issues around power and impotence when it comes to AI. Still, we cannot abandon talking about transparency either; but we need to do it differently. Hence, I stirred up the ground we stand on when it comes to transparency as wrong and limited by talking about AI literacy as empowerment not as explanations. I equally tried to minimise the gap between people's everyday experience with AI and the intellectual and academic discourse around it as a post-critical attempt to bring back critique into people's lives (Habed, 2021: 500).

Crisis of authority and the slippery truth concept: Two consequences for other debates

We gain more things from a post-critical AI literacy. We can situate this discussion into two parallel developments in society that were left undiscussed here, but are crucial for the wider transparency narratives; first, the authority crisis and second, post-trust discourses.

First, the technocratic authority and hegemony of XAI might be contested by publics, even if is not yet. My concerns about XAI pragmatizing transparency is because their

research agenda is to understand *algorithms*, not society's needs. We need to keep in mind that imposing AI in public governance is often an authoritative process – seen only as the pragmatic extension and tool of a democratic process and assigned government, even though it is implementing a new form of power and decision making, *undemocratically*. What if citizens protested to be governed by AI (even if this is still about human decisions)? What if AI only masks the existing imbalance but is sold as its reason? Citizens and publics will have an opinion, at least, even if it might be a reactionary one. So far, we see public debates on the relevance of science conflicting with each other. These are not new, nor are the conflicts. Somehow, a depoliticised transparency debate seems like a safety net for those who would like to appear as if they are ignoring the political implications, as a way to escape their regulations (like Big Tech companies). But public services cannot escape the public and legal scrutiny nor responsibility.

For instance, concepts like trust could be contested even more, because of conflicting transparency models. What could this look like? Will Davies (2018) points to the late 19th century, when nationalist movements mobilised the masses through extensive information campaigns, resulting in a democratic authority crisis. Trust might be flipped in this AI context. We will be increasingly confused and in the dark about AI, but not because AI is opaque as an algorithmic box (which it is!), but because we might not trust authorities to give us accurate information about what is in this box. Davies (2018) states, "instead of trusting experts on the assumption that they are impartial and objective, we have come to rely on services that are quick but whose public standing is questionable" (14). AI-explaining authorities lack legitimacy (ideally through public debate and not just technocracy), which increases distrust. Participatory systems must precede pure technocracies because citizens are more likely to legitimise them (Sætra, 2020: 6). Authorities are unstable, and no model of transparency, technology, or government sanctions can fix this (Marecos & de Abreu Duarte, 2022: 211).

Further, the discussion around trust and post-truth or misinformation is heavily fuelled by AI and its opacity, but it is not *due* to AI and its opacity. A post-critical literacy of AI may overcome the pragmatisation of transparency and dispel the myth that literacy is neutral, even if it empowers. But no literacy can fully empower nor reveal any algorithmic truths (hence, not claiming algorithmic transparency as the ultimate or

even achievable goal). We must debate AI's role in misinformation, post-truth campaigns, and stereotypes without relying on the algorithm to give us answers that are not algorithmic, but political. But we also need to discuss that algorithms are not total game changers in these ideological frissons caused by informational or political transparency. Algorithmic misinformation campaigns might have challenged our lives and infoscapes, as much as contesting our knowledge systems (de Abreu, 2022), but they are embedded in ideological and economic backgrounds that shape our political comprehension (Abreu & Oner, 2020). The public holds a legitimate distrust of media as powerful and opaque institutions (Fenton, 2019), especially in their neoliberal instantiation (Freedman, 2019). Every form of organised literacy will remain entangled in a wider truth and authority crisis. It is surprising that XAI has not yet addressed this issue. Unquestioned hegemonies validate and manifest truths, which rages as scientific authority declines (Davies, 2018; Mouffe, 2000), and this will affect how the data sciences position themselves in this AI discourse. Since information is never impartial nor apolitical, a depoliticised take on transparency overpromises to democratise and provide complete access to information, while potentially leading to more impotence in time. Information might be faster and more accessible, but it is not more transparent. Again, we see the pragmatics of the technology catching up on the concept of transparency, giving it a technical aura; one we should not overrate.

Conclusion

We cannot empower citizens in dealing with AI's opacity without transparency, but having algorithmic transparency is not going to empower on its own. Proposing an AI literacy was a post-critical effort to challenge the dispositions that explainability of AI (XAI) research builds on. Assuming that algorithms are the only level that matters, and that they can be explained in a straightforward, non-ideological way, XAI depoliticises transparency in a way that strengthens its own power and voice, but limits that of those affected by AI (often negatively).

Challenging the authority of those who explain AI and how it works also transforms the accessibility of academic critique, pointing to the value of a post-critical turn in this debate. As a way to minimise the gap between people's everyday experience with AI

and the intellectual and academic discourse, the post-critical impulse attempts to bring back critique into people's lives (Habed, 2021: 500). Thus, creating a post-critical AI literacy is an attempt to reconsider access to AI as a question of power rather than a question of algorithmic toolboxes. We require a much broader understanding and accessibility of artificial intelligence with less emphasis on the dichotomy of transparency versus opacity. Furthermore, while transparency is promoted as a pragmatic solution to opacity, it is ineffective in dismantling the opaque (one-sided) hegemony of AI held by those who develop it. Those subjected to discrimination are frequently unaware of it and are therefore powerless to stop it, which is exacerbated by AI's obscurity. Understanding how algorithms work is important, but I have argued that having the power and resources to challenge their use is at least equally important. Algorithms, particularly neural networks (also known as AI), are notoriously difficult to understand. However, addressing their algorithmic opacity does not remove its ideology; if we take the depoliticised route, we simply avoid discussing the ideology around what is revealed and to whom. Machines and algorithms are a part of our social fabric, which is both frightening and difficult to comprehend. However, algorithms are not neutral, or free of human values or bias; there is no way of cleansing them from human *contamination*.

In the case of the Dutch childcare benefit algorithm that accused caregivers wrongly of tax fraud, explainability would have been critical for the deciders, but it might have still automated inequality by making it impossible to intervene from the bottom up. Perhaps XAI would have addressed the discriminatory structure in the risk assessment on time and halted implementation. On this subject, I can only speculate. We must be aware, however, that our inability to challenge this system, as well as our lack of understanding of it, contributes to its power.

A post-critical discussion is not an anti-critique or meta-critique of XAI, but about challenging the dispositions and hegemony of algorithmic insights, and the negligence over the latter being as ideological as a seemingly pragmatic demand. Avoiding a *meta* critique only means that I do not position my voice on top of XAI by offering a critical engagement with their algorithmic focus only, but step aside by thinking of access and empowerment too, while still dealing with their blind spots. Although my ideal literacy model does not currently exist, the opportunities for knowledge and decision-making

it would provide would be empowering. This is difficult to do in practice. Who should make the call? Based on what? Although this is debatable, the current decision makers appear to have been decided too quickly. XAI manifests the hegemony of the data sciences as the technocratic power and authority in how we approach education and literacy. But knowledge and access to AI are inextricably linked, and neither is solely based on algorithms. On the issues for which we seek access and explanations, we must confront the politics behind it. We are at the beginning of a longer process that will change and shape many discourses, eventually affecting those who are already marginalised and undervalued. Hence, fostering literacy around algorithms cannot become a neutralised, technical *skills lab*. Rather than a core of technology that we can explain, we are surrounded by a web of embedded relationships, ideological potencies, and our inability to control them.

We may never be able to fully resolve the greater powerlessness we feel as citizens in the face of AI – even if we should not proclaim any doomsday scenarios that promote a sophistication these systems do not have yet – but we can foster a transdisciplinary dialogue worthy of its post-critical status. Shaping a post-critical AI literacy could be the first step in questioning not only how AI is explained, but also how its *mysterious* power is applied. Literacy can only work to eliminate powerlessness as an anti-transparency practice by incorporating a broader AI spectrum. It was beneficial and productive to revisit established social and political critiques and draw from their insights rather than dismissing them. Transparency, seen through a postcritical lens, may have revealed a simple but profound-*ish* truth: AI wears new clothes, but it covers old issues. It is critical to get to know both better, and, for want of a better term, to make both more *transparent*.

Acknowledgments

This research has been funded by the Vienna Science and Technology Fund (WWTF)[10.47379/ICT20058]

References

Abreu, M. and Öner, Ö. (2020) 'Disentangling the Brexit vote: The role of economic, social and cultural contexts in explaining the UK's EU referendum

vote', *Environment and Planning A: Economy and Space*, 52(7): 1434–1456. doi: https://doi.org/10.1177/0308518x20910752.

Amann, J., Vetter, D., Blomberg, S.N., Christensen, H.C., Coffee, M., Gerke, S., Gilbert, T.K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R.V. & Madai, V.I. (2022) 'To Explain or not to Explain?—Artificial Intelligence Explainability in Clinical Decision Support Systems', *PLOS Digital Health* 1(2): 16. doi: https://doi.org/10.1371/journal.pdig.0000016.

Anker, E. & Felski, R. (2017) *Critique and Postcritique*. Durham: Duke University Press.

Austrian Academy of Sciences (ÖWA) (2020) *Der AMS Algorithms* [online] oeaw.ac.at. Available at: https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus/.

Baldi, M. & Seraydarian, P. (2022) 'Media Literacy as Civic Discourse: A Framework for Inquisitive 'Listening' and Authentic 'Speaking' in a Digital Space', in: B. De Abreu, ed., *Media Literacy, Equity, and Justice*. London: Routledge.

Bartoletti, I. (2020) *Artificial Revolution: On Power, Politics and AI*. London: The Indigo Press.

Berger, S., Fengler, S., Owetschkin, D. & Sittmann, J. (2021) *Cultures of Transparency: Between Promise and Peril*. London: Taylor & Francis.

Birchall, C. (2014) 'Radical Transparency?', *Cultural Studies ↔ Critical Methodologies*, 14(1): 77–88. doi: https://doi.org/10.1177/1532708613517442.566t

Buckingham, D. (2018) 'Going Critical: On the Problems and the Necessity of Media Criticism', in H. Niesyto and H. Moser, eds., *Medienkritik im digitalen Zeitalter*. München: Kopead, pp. 45–59.

Buckingham, D. (2019) *The Media Education Manifesto*. Cambridge: Polity.

Castiglia, C. (2017) 'Hope for Critique?', in E. S. Anker & R. Felski, eds., *Critique and Postcritique*. Durham: Duke University Press, pp. 211-229.

Cave, S., Dihal, K. & Dillon, S. (2020) *AI Narratives: A History of Imaginative Thinking About Intelligent Machines*. Oxford: Oxford University Press.

Chazette, L., Brunotte, W. and Speith, T. (2021) 'Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue', *IEEE 29th International Requirements Engineering Conference (RE)*. doi: https://doi.org/DOI%2010.1109/RE51729.2021.00025.

Coeckelbergh, M. (2020) *AI Ethics*. Cambridge, Massachusetts: The MIT Press.

Davies, W. (2018) *Nervous States: How Feeling Took Over the World*. London: Jonathan Cape.

De Abreu, B. (2022) *Media Literacy, Equity, and Justice*. New York: Routledge.

Diakopoulos, N. & Koliska, M. (2017) 'Algorithmic Transparency in the News Media', *Digital Journalism* 5(7): 809–828. doi: https://doi.org/10.1080/21670811.2016.1208053.

Durán, J.M. & Jongsma, K.R. (2021) 'Who is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI', *Journal of Medical Ethics*, p.medethics-2020-106820. doi: https://doi.org/10.1136/medethics-2020-106820.

Eubanks, V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

European Commission (EC) (2022) 'Artificial intelligence Act (AI Act)' [online]. *European Parliament*. Available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792 [Accessed 1 Feb. 2023].

Fazi, B. (2020) 'Beyond Human: Deep Learning, Explainability and Representation', *Theory, Culture & Society*, 38(7):1–23. doi: https://doi.org/10.1177/0263276420966386.

Felski, R. (2015) *The Limits of Critique*. Chicago: University Of Chicago Press.

Fenton, N. (2019) '(Dis)Trust', *Journalism* 20(1): 36–39.

Freedman, D. (2019) 'Media and the Neoliberal Swindle: From 'Fake News' to 'Public Service'', in: S. Dawes and M. Lenormand, eds., *Neoliberalism in Context: Governance, Subjectivity and Knowledge*. Cham: Palgrave, pp. 215–231.

Gitelman, L. (2013) *'Raw data' is an Oxymoron*. Cambridge: MIT Press.

Green, B. (2021). Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing* 2(3): 249–265. doi: https://doi.org/10.23919/jsc.2021.0029.

Gunning, D. & Aha, D. (2019) 'DARPA's Explainable Artificial Intelligence (XAI) Program', *AI Magazine*, [online] 40(2): 44–58. doi: https://doi.org/10.1609/aimag.v40i2.2850.

Gutiérrez-Martín, A. and Torrego Gonzales, A. (2022) 'ICT and Media Education Curriculum for Teachers in the Post-Truth Era', *Routledge eBooks*, pp.33–44. doi: https://doi.org/10.4324/9781003175599-6.

Gutiérrez-Martín, A. & Tyner, K. (2012) 'Educación para los medios, alfabetización mediática y competencia digital', *Comunicar* 19(38): 31–39. doi: https://doi.org/10.3916/C38-2012-02-03.

Habed, A.J. (2021). The Author, the Text, and the (Post)Critic: Notes on the Encounter Between Postcritique and Postcolonial Criticism. *Postcolonial Studies*, 24(4): 498–513. doi: https://doi.org/10.1080/13688790.2021.1985248.

HLEG (High Level Expert Group) (2019) 'Ethics Guidelines for Trustworthy AI' [online] *futurium - European Commission*. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html. Accessed 10.07.2023.

Hoff, K.A. & Bashir, M. (2015) 'Trust in Automation', *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57(3): 407–434. doi: https://doi.org/10.1177/0018720814547570.

Hoffman, R.R., Klein, G. & Mueller, S.T. (2018) ,'Explaining Explanation For 'Explainable AI'', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62(1): 197–201. doi: https://doi.org/10.1177/1541931218621047.

Hollanek, T. (2020) AI Transparency: A Matter of Reconciling Design With Critique. *AI & Society*, no vol, no issue number, pp.1-9. doi: https://doi.org/10.1007/s00146-020-01110-y.

Klein, A. (2023) 'AI Literacy, Explained', *Education Week* [online] 10 May. Available at: https://www.edweek.org/technology/ai-literacy-explained/2023/05. Accessed 30.06.2023

Knowles, B. (2020) 'Explainable AI: Another Successful Failure?', in: *Chi 2020*. Chi Conference Proceedings.

Kong, S.C. & Abelson, H. (2022) *Computational Thinking Education in K-12: Artificial Intelligence Literacy and Physical Computing*. Cambridge: The MIT Press.

Kuźniacki, B. (2023) 'The Dutch Childcare Benefit Scandal Shows That We Need Explainable AI Rules' [online] *uva.nl*. Available at: https://www.uva.nl/en/shared-content/faculteiten/en/faculteit-der-rechtsgeleerdheid/news/2023/02/childcare-benefit-scandal-transparency.html?cb. Accessed 11.07.2023.

Liefgreen, A., Weinstein, N., Wachter, S. & Mittelstadt, B. (2023) 'Beyond Ideals: Why the (Medical) AI Industry Needs to Motivate Behavioural Change in Line with Fairness and Transparency Values, and How It Can Do It', *AI & Society,* no vol, no issue, pp. 1-17. doi: https://doi.org/10.1007/s00146-023-01684-3

Livingstone, S. (2004) 'Media Literacy and the Challenge of New Information and Communication Technologies', *The Communication Review* 7(1): 3–14. doi: https://doi.org/10.1080/10714420490280152.

Marecos, J. & de Abreu Duarte, F. (2022) 'A Right to Lie in the Age of Disinformation: Protecting Free Speech beyond the First Amendment 206 J', in B. de Abreu, ed.: *Media Literacy, Equity, and Justice*. London: Routledge, pp.206–213.

Mouffe, C. (2000) *The Democratic Paradox*. London: Verso.

Niesyto, H. & Moser, H. (2018) *Medienkritik im digitalen Zeitalter*. München: Koepad.

Owetschkin, D., Sittmann, J. & Berger, S. (2021) 'Cultures of Transparency in a changing world: An introduction'. In: *Cultures of Transparency*. London: Routledge.

Parisi, L. (2013) *Contagious Architecture*. Cambridge: MIT Press.

Pasquale, F. (2015) *The Black Box Society*. Cambridge: Harvard University Press.

Perrigo, B. (2023) 'AI By the People, For the People' [online] *Time*. Available at: https://time.com/6297403/the-workers-behind-ai-rarely-see-its-rewards-this-indian-startup-wants-to-fix-that/. Accessed 30.07.2023.

Reinhardt, K. (2023) 'Trust and trustworthiness in AI ethics', *AI Ethics* 3: 735–744. https://doi.org/10.1007/s43681-022-00200-5

Robinson, S.C. (2020) 'Trust, Transparency, and Openness: How Inclusion of Cultural Values Shapes Nordic National Public Policy Strategies for Artificial Intelligence (AI)', *Technology in Society* 63(101421): 1–15. doi: https://doi.org/10.1016/j.techsoc.2020.101421.

Rosenfeld, A. & Richardson, A. (2019) 'Explainability in Human–Agent Systems', *Autonomous Agents and Multi-Agent Systems* 33: 673–705. doi: https://doi.org/10.1007/s10458-019-09408-y.

Sætra, H.S. (2020) 'A Shallow Defence of a Technocracy of Artificial Intelligence', *SSRN Electronic Journal* 62(101283): 1–10 doi: https://doi.org/10.2139/ssrn.3494309.

Samek, W., Wiegand, T. & Müller, K.-R. (2017) 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models' [online] Available at: https://arxiv.org/abs/1708.08296 [Accessed 15 Aug. 2023]. – no journal

Shin, D. (2021) 'The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI', *International Journal of Human-Computer Studies* 146: 102551. doi: https://doi.org/10.1016/j.ijhcs.2020.102551.

Singh, D., Slupczynski, M. and Pillai, A.G. (2022). *Grounding Explainability Within the Context of Global South in XAI*. [online] doi: https://doi.org/10.48550/arXiv.2205.06919 [Accessed 15 Aug. 2023] – no journal.

Srinivasan, R. & Chander, A. (2020) 'Explanation Perspectives from the Cognitive Sciences---A Survey', *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*. doi: https://doi.org/10.24963/ijcai.2020/670.

Strathern, M. (2000) 'The Tyranny of Transparency', *British Educational Research Journal* 26(3): 309–321. doi: https://doi.org/10.1080/713651562.

Sudmann, A. (2019) *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*. Bielefeld: Transcript.

Valdovinos, J.I. (2022) *Transparency and Critical Theory*. Cham: Springer Nature.

Vincent, J. (2023). 'AI is Killing the old web, and the new web Struggles to be Born', [online] *The Verge*. Available at: https://www.theverge.com/2023/6/26/23773914/ai-large-language-models-data-scraping-generation-remaking-web. [Accessed 18 Jul. 2023].

Weller, A. (2019) 'Transparency: Motivations and Challenges', In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, KR. (eds.) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science*, vol 11700, pp.23–40. Cham: Springer. https://doi.org/10.1007/978-3-030-28954-6_2.

Zylinska, J. (2020) *AI Art: Machine Visions and Warped Dreams*. London: Open Humanities Press.

Notes

[1] The contested term "artificial intelligence" is used here for illustrative purposes only; it refers to dominant narrative around algorithmic systems situated in a tradition of human intelligence (Cave et al., 2020).

[2] This system was driven by a Risk Classification Model (RCM), an algorithmic decision-making system calculating risk profiles to detect potential tax fraud. https://www.dailysabah.com/politics/news-analysis/dutch-child-care-subsidies-scandal-exposes-countrys-systematic-xenophobia-turkophobia. Accessed 14 July 2013.

[3] https://www.theguardian.com/world/2021/jan/14. Accessed 13 July 2023.

[4] https://eulawenforcement.com/?p=7941. Accessed 10 July 2023.

[5] https://www.rijksoverheid.nl/documenten/kamerstukken/2022/05/17/antwoorden-op-vragen-van-het-lid-van-haga-over-toeslagenaffaire. Accessed 19 July 2023.

[6] https://www.dailysabah.com/politics/news-analysis/dutch-child-care-subsidies-scandal-exposes-countrys-systematic-xenophobia-turkophobia. Accessed 14 July 2023.
[7] https://green20s.de/Berliner-Klima-Burger-Innenrat. Accessed 08 July 2023.
[8] https://amsalgorithmus.at/en/. Accessed 06 July 2023.

Eugenia Stamboliev is a media philosopher and critical technology scholar at the University of Vienna. As a fellow in the WWTF project 'Interpretability and Explainability as Drivers to Democracy', she investigates the political power of complex algorithmic models and how to make these accessible to broader publics. Her research focuses on AI literacy as a techno-political undertaking, the relationship between climate crisis and AI innovation, and the (un)trustworthiness of platforms.

Email: eugenia.stamboliev@univie.ac.at