LUND UNIVERSITY

Four Facets of AI Transparency

Larsson, Stefan; Haresamudram, Kashyap; Högberg, Charlotte; Lao, Yucong; Nyström, Axel; Söderlund, Kasia; Heintz, Fredrik

Published in:
Handbook of Critical Studies of Artificial Intelligence

DOI:
10.4337/9781803928562.00047

2023

Document Version:
Peer reviewed version (aka post-print)

Link to publication

Citation for published version (APA):
Larsson, S., Haresamudram, K., Högberg, C., Lao, Y., Nyström, A., Söderlund, K., & Heintz, F. (2023). Four Facets of AI Transparency. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* (pp. 445-455). Edward Elgar Publishing Ltd.. https://doi.org/10.4337/9781803928562.00047

Total number of authors:
7

Creative Commons License:
Unspecified

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Download date: 18. Oct. 2026

Preprint version (2022). To be included in Lindgren, S. (ed.) *Handbook of Critical Studies in Artificial Intelligence*, Edward Elgar Publishing.

Four Facets of AI Transparency

Stefan Larsson,[1] Kashyap Haresamudram,[1] Charlotte Högberg,[1] Yucong Lao,[2] Axel Nyström,[3] Kasia Söderlund,[1] Fredrik Heintz[4]

[1] Department of Technology and Society, Faculty of Engineering, Lund University, Sweden
[2] Department of Information Studies, Faculty of Humanities, University of Oulu, Finland
[3] Department of Laboratory Medicine, Faculty of Medicine, Lund University, Sweden
[4] Department of Computer and Information Science, Linköping University, Sweden

Abstract
Transparency in artificial intelligence (AI) can mean many things, but at the same time, it is currently a central focus for both scientific and regulatory attention. We seek to critically unpack this conceptual vagueness. This is particularly called for given recent focus on transparency in much of AI policy. To this end, we construct our analysis of AI Transparency into four facets. Firstly, (1) *explainability* (XAI) has become an expanding field in AI, which we argue needs to be complemented by more explicit focus on the (2) *mediation* of AI-systems functionality, as a communicated artefact. Furthermore, in the policy discourse on AI, the importance of (3) *literacy* is underscored. We draw from the rich literacy literature in order to show both promising and troubling consequences of this. Lastly, we unpack transparency as a form of governance, within a (4) legal framework encompassing a structure of trade-offs. By these four facets we aim to bring more clarity to the multifaceted concept of transparency in AI.

Keywords: AI transparency; explainable AI (xAI); mediation; AI literacy; law as tradeoff

Introducing a Multifaceted Concept

'Transparency' is one of those contemporary concepts that, linked to AI, spans technical, legal, and ethical – and more – perspectives. While transparency is part of a wider trend in international governance (Koivisto, 2022), it is also one of the most common concepts in the recent surge of ethics guidelines on AI that has been developed by a wide variety of entities from governments, NGOs, and large companies to multi-stakeholder groups (Jobin et al., 2019). Often, it is framed as a mechanism for promoting accountability (Diakopolous, 2020). In recent EU policy on AI, there is a focus on risk assessments and auditing (Felländer et al., 2022; Mökander et al., 2021), with an emphasis on "human-centricity" (Larsson, 2020; Larsson et al., 2020), implicating how European countries strategize about AI (Robinson, 2020), their national mandates, and initiatives for various sectors, not the least the public sector (de Bruijn et al., 2022).

For some of the origin of transparency as a governance tool, firstly, one can point to the policy-debates on anti-corruption pushing for corporate and governmental transparency in

the late 1990s and early 2000s (Forssbaeck & Oxelheim, 2014; Koivisto, 2022), but some of its recent support in EU policy could arguably also be explained by its positive connotations as metaphorically linked to openness (Koivisto, 2022; Larsson & Heintz, 2020). As a reaction, it has also spurred the more aesthetically and politically framed emerging field of *critical transparency studies* (cf Alloa, ed., 2022; Koivisto, 2022), which we draw from in order to outline some of the implications of "AI Transparency" in contemporary policy debates. Recently, and secondly, transparency – particularly in terms of algorithmically focused "explainability" (cf Haresamudram et al., 2022) – has been put forward as a key element of ensuring AI to perform well and fulfil its promises as well as strengthen *public trust* in AI (cf Jacovi et al., 2021). In this chapter, we describe why common approaches to explainability constitutes a narrow concept and propose how it can be complemented for a richer understanding of its consequences for policy.

By drawing from critical examinations of AI transparency, such as Jenna Burrell's three forms of opacity (2016) and Ida Koivisto's account on the transparency paradox (2022), this chapter develops four facets of AI transparency. Firstly, we critically examine the growing body of literature on *explainable AI*, which stems from a call to make machine-learning processes more understandable. Secondly, and inspired by recent critique (Miller, 2019) that this field draws too little from how humans actually understand explanations, we see a need to break out the explicit *mediation* of machine learning processes that this leads to. Similarly, Burrell discusses these two facets in terms of a "mismatch between mathematical procedures of machine learning algorithms and human styles of semantic interpretation" (Burrell, 2016, p. 3). In addition, while Burrell has a narrower focus on technical illiteracy, mainly pointing to coding abilities, we – thirdly – expand by drawing from the rich field of *literacy* studies for our third facet. Lastly, and fourthly, Fourthly, and lastly, we point to transparency as a form of governance in itself, and place this analysis within a legal framework by pointing to how law often is tasked with balancing different interests. Law is thereby making *tradeoffs* between for example public needs to supervise and assess the use of corporate and governmental AI-systems, on the one hand, and security needs demanding secrecy or legally supported notions of secrecy to ensure competition, on the other. By Burrell identified as *intentional* corporate or state opacity.

Four Facets

In this section, we develop the analysis of the four facets of particular relevance to AI Transparency.

1. Transparency as explanation

The notion that complex AI systems entail "black box" issues that demand better *explainability methodologies* has been established for some time, and constitutes a central aspect of AI transparency (Larsson, 2019). While definitions vary, explainable AI (XAI) can generally be considered to produce "details or reasons to make its functioning clear or easy to understand" (Barredo Arrieta et al., 2020, p. 85). The challenge of interpreting and explaining AI systems has attracted growing attention as the methods, such as deep learning, have increased in complexity (cf Qi et al., 2021), but also as AI has been applied for

more diverse audiences and groups of users (cf Ribera & Lapedriza, 2019). Explanations by XAI models can take different forms, such as texts by generated captions, visualisations by generated images, local explanations by gradient maps/heat maps, or by generated nearest neighbours and counterfactuals (cf Lipton, 2018; de Vries, 2021). The rapid growth in XAI research is, at least in part, motivated by a need to maintain trust between the human user and the AI (Jacovi et al., 2021), a notion also echoed in European policy (cf Larsson et al., 2020).

Several attempts have been made to develop taxonomies of explainability techniques and their desiderata, with common axes of explanations including *global* versus *local* (explaining the model versus explaining a specific prediction), *model specific* versus *model agnostic* techniques (in regards to which set of AI models the explainability technique applies to), and *model complexity*. Another common distinction is between "inherently explainable" models (often called "transparent" or "interpretable"), and so-called post-hoc techniques that attempt to explain the behaviour of an otherwise black-box model (Barredo Arrieta et al., 2020). In line with this, Rudin (2019) argues that one should always strive to use inherently interpretable models, rather than resorting to post-hoc explanations of black box models, at least if the stakes are high.

Even if the development of XAI often is motivated by a need to contribute to trusted and fair AI applications, synthesised into a need to better understand aspects of causality and transferability, there seems to be a lack of a unified terminology. For example, there is a distinct lack of metrics, some argue, by which such objectives can be easily quantified and compared (Lipton 2018, Barredo Arrieta 2020). Although there are many XAI techniques available for black box models, the lack of explainability metrics makes it difficult to validate their utility. When such techniques have been tested empirically, counter-intuitive results are not uncommon. For example, Kaur et al. (2020) found that even data scientists tend to misuse and overtrust visual explanations, not noticing when the models misbehave, which we return to below in the section on mediation.

A key problem with XAI methods, stressed by Miller et al. (2017), is its lack of grounding in the social and behavioural sciences. Similarly, Mittelstadt et al. (2019) argue that there is a fundamental distinction between explanations provided by AI and everyday explanations intended for humans. The latter is, in short, not the same as the "interpretability" and explainability found in the XAI domain. This has led researchers to conduct meta-studies, drawing from social psychology and philosophy, on the critical properties of human explanations (cf Miller, 2019). From this perspective, Miller (2019) argues, explanations are often

1. *contrastive*, that is, people ask not necessarily why an event happened, but rather why an event happened instead of another event,
2. an outcome of the fact that we tend to *make a biased selection* of one or two causes from a sometimes infinite number of causes to be *the* explanation,
3. *not strictly depending on probabilities*, as much as referred causes, that is, the most likely explanation is not always the best explanation for a person, which leads to the last category stating that explanations are
4. *social*. That is, part of a conversation or interaction, which implicates the explanation.

The above-mentioned aspects, including risks of misused visual explanations and the lack of attention to how humans understand explanations, lead us to explicitly focus on the complementing mediation as a facet of AI transparency in its own right.

2. Transparency as mediation

Following the proposition that a distinction should be made between XAI and human explanations established in the social sciences (Miller, 2019; Russell & Wachter, 2019), this section conceptualises explanations as components of *mediation*, and highlights its role in achieving transparency. Here, we draw from what Koivisto (2022) refers to as transparency "as a medium", generative and non-neutral, albeit here more distinctly focused on the contemporary AI discourse. We elaborate on the modes of mediation, whose implications for AI transparency we argue is understudied and in need of further scrutiny. This should also be seen in light of contemporary calls for more transparency in the application of AI-systems, in terms of that we seek to underscore the meaning of mediation, as an important aspect of whatever goals transparent AI is set to reach. Transparency, as a metaphor linked to seeing (cf Koivisto, 2022; Larsson & Heintz 2020), is not neutral, but a "ocularcentric" notion, in the words of Koivisto (2022) that also seem to downplay other notions of mediation than the visual.

To facilitate more effective human-AI communication, Miller (2019) proposes that AI explanations should be designed to incorporate characteristics of human explanations (counterfactual, selective, contextual). A majority of the existing XAI techniques produce statistical probability explanations either textually or through graphical representations. XAI is often argued to serve specific use-cases meant to be handled by domain experts, and thus there is no need for general explainability intended for non-experts. Within this context, so-called *interpretability tools* have been designed to help data scientists and machine learning practitioners better understand how AI systems work (cf Kaur et al., 2020). These tools favour visualisation as a medium of communication. However, recent research indicates that data scientists risk overtrusting and misusing interpretability tools, and shows that visual output can be misleading (Kaur et al., 2020).

Evidently, mediation of explanations provided by AI is not, and should not be seen as, limited to visual representations; in a mundane everyday context, they can arguably also take the shape of text or symbols in user interfaces and user agreements (Larsson & Heintz, 2020), where online ads has been pointed to as a particularly problematic and opaque area (Andrejevic et al., 2022). When considering mundane and everyday practices, transparency as a medium of information poses a great challenge to non-expert users of AI, such as consumers, citizens or patients. The choice of words or symbols can metaphorically highlight certain features and downplay others, which structures and guides how certain phenomena are understood, potentially leading to normative implications for law's attempt to regulate new technologies (Larsson, 2017) as well as affecting users' understanding of technological interfaces (Stanfill, 2015). These choices need to be studied and understood also for the sake of improving AI governance in everyday life. This is especially important when translating explanations across languages, where different metaphor-relations may be at play. This level of nuanced mediation is largely lacking in XAI it seems, particularly in relation

to policy and for example consumer interests, which could be concerning in relation to overconfidence in automated tools (for a critical examination of the feasibility of AI policy focusing on human control over automation, see Koulu, 2020). With AI systems being integrated into commonplace products and services, they have to meet legal requirements aimed at protecting the user's privacy, amongst other things. This means that users may need to be informed and consent to data collection involved in the automation. To this end, for example cookie consent banners were implemented in the EU to provide more transparency regarding online data collection, as well as a rich plethora of consent agreements and privacy policy statements used to communicate how personal data is collected and processed – which often is a prerequisite for consumer-facing AI-applications like recommender systems. Several studies, however, indicate that most consumers do not understand such communications on how their data is collected and what it is used for. Similarly, research suggests that users find the quantity of information overwhelming, causing information overload, and leading them to disregard the information altogether (Cranor et al., 2015; Larsson et al., 2021).

Mediation between humans and AI is a field in need of more scrutiny and development. Research in this space is spread across a myriad of disciplines, such as psychology, cognitive science, communication and information studies, and interaction design; bridging knowledge from all these fields is a pressing challenge. Mediation implies an addressee and an audience, which we analyse in terms of *literacy* in the subsequent section (cf Burrell, 2016, on opacity as technical illiteracy).

3. Transparency as Literacy

The lack of literacy is often cited as a reason for why AI-applications are considered opaque. Technical illiteracy is for example identified by Burrell (2016) as one of the main reasons for the "state of not knowing". It is, however, also tightly interwoven with the particularly complex characteristics of machine learning algorithms and their scale of operation (Burrell 2016). This is what forms the basis for claims that we need "new forms of interpretability and literacy" (van Nuenen, et al., 2020, p. 43). In this section, we consider the nuances of literacy under the umbrella of AI transparency, and implications and limitations of literacy as a solution to AI opacity.

Originally, literacy referred to "the ability to express ourselves and communicate using written language", but it has since come to be defined as "skill sets in a variety of disciplines that have the same potential to enable expression, communication and access to knowledge" (Long & Magerko, 2020, p. 2). In the AI discourse, the call for literacy has become a common normative standpoint in addressing issues of governance (Larsson et al., 2020, Jobin et al. 2019). For instance, there is a clear emphasis on transparency in the *Ethics Guidelines for Trustworthy AI* prepared by the High-Level Expert Group on AI, (cf Larsson 2020), including calls for both data and algorithmic literacy in European policy more broadly. Similarly, Strauß argues that opaque AI systems can reinforce so-called wicked problems, involving "ill-formulated risks of undetected failure, self-fulfilling prophecies and an incremental normalisation of AI biases in society" (Strauß, 2021, p. 45). Therefore, he argues that having a basic understanding of AI and raising problem-awareness among decision-makers and persons interacting with AI systems, is essential to face these problems.

In the context of AI, different *types of literacies* are discussed. Long & Magerko (2020, p. 2) defined AI literacy as "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace". However, it is frequently pointed out that digital literacy is a precondition for AI literacy, while computational literacy might not be essential, though it depends on who the audience is (e.g. Barredo Arrieta et al., 2020). Scientific literacy can, similarly, inform AI literacy, whereas data literacy is argued to overlap with AI literacy (Long & Magerko 2020, p. 2).

Embedded in the strive for AI literacy is the question of *who* needs to be literate, i.e. to whom should algorithms be transparent? Is it the general public, auditors, legislative actors, policy makers, AI developers themselves, or any other group? Various actors are involved with AI systems, requiring different sets of knowledge and skills, and, therefore, different literacies. Furthermore, different groups, individuals and professions vary in their competence to review information about AI systems. For example, a report conducted for the European Parliament on algorithmic transparency and accountability, differs between if the intended transparency is for "everyone", regulatory staff, third-party forensics, or researchers (Koene et al., 2019). Moreover, the importance of *a critical audience* for transparency of AI is stressed (Kemper and Kolkman, 2019).

Yet, AI literacy cannot, due to system architectures and input data, be regarded without (big) data literacy and information literacy in general (Jandrić, 2019, p. 33). The call for increased AI literacy can be viewed in light of broader calls for literacy of media, information, data and digital technologies, and the imaginaries they hold. Firstly, they build on the notion that increased literacy by necessity leads to a surge of knowledge and empowerment, as well as general social progress. This is despite hardships of identifying what these literacies need to consist of; operation of skills or deeply critical and reflexive reasoning? (cf Jandrić, 2019; Ng et al., 2021; Lloyd, 2019). Secondly, often, these literacies are conceptualised as responsibilities of the individual, including being able to deal with ambiguous claims (Haider & Sundin, 2022, p. 30). Placing the onus for gaining and using them on the individual, can as a policy approach have serious implications for questions of how to distribute accountability. That is, while literacy indeed can foster empowerment, the request for AI literacy as a strategic policy for regulating the relationship between AI-using companies or authorities and their human consumers or citizens, may promote an individualistic approach that at worst displaces the required scrutiny down to the individual users. The user or subject is expected to educate themselves, stay informed about, understand the consequences of different AI use cases, and take action. This can be compared to critical and empirical research on data-intensive consumer research that points to a "corporate cultivation" of resignation amongst consumers (Draper & Turow, 2019), as a sort of obfuscation (Zalnieriute, 2021), which in some practices lead to "uninformed" consent (Utz et al., 2019).

Nevertheless, increased AI literacy is put forward by several governance initiatives as an important piece of the puzzle towards trustworthy AI (2019, p. 23). While education and awareness are framed as important for trust in AI, in ethics guidelines (Jobin et al., 2019) critical awareness might be shadowed by the governance landscape in which AI systems

reside, through legally sanctioned limits to insight and transparency. In international governance, transparency has come to take a central position as belonging to a "group of good concepts" (Koivisto, 2022, p. 162) – along with for example human rights and democracy. To contrast this discursive and normative connotation – that more transparency is better – we will now turn to ideas on when transparency needs balancing in a tradeoff with other interests. We use a legally informed framework for this demonstration.

4. Transparency in a Legal Tradeoff

Under the umbrella of AI transparency, we have so far related the algorithmically focused explainability concept to how AI-systems are mediated, as well as turned the gaze to the addressees in terms of their literacy. Lastly, we seek to place the transparency concept into a governance context in which law as an equilibrium tool becomes central. While recent emphasis on transparency in soft or "ethics-based" governance of AI-systems may fit well with the overarching goals of making these systems more explainable, better mediated and understood, the wider perspective of governance considers several values and interests to be balanced. Transparency, as a form of governance (cf. Flyverbom, 2015), will within law therefore have to be seen as something that takes part in a trade-off. As stressed in the literature (cf. Koivisto, 2022; de Laat, 2018), there are several legitimate reasons for keeping certain things secret, and regardless of what recent ethics guidelines put forward, this trade-off is inevitably to be played out in practice in most jurisdictions, not the least European.

To begin with, it should be noted that there is a firmly established legal notion that there are legitimate interests in limiting transparency. Access to AI systems, including their source code, associated parameters, training data, training processes, and resulting models, is according to this legal notion not always warranted. For instance, AI proprietors generally prefer not to reveal the inner workings of their systems in order to keep their competitive advantage on the market. Other motives for not revealing detailed information about AI systems include the need to prevent users from gaming the algorithms. Too much transparency in such cases, is argued to risk leading to abuse of the systems, e.g. cyberattacks, and defeat the purpose of said systems (cf. de Laat, 2018). Likewise, uncontrolled access might jeopardise personal data used for training the models (de Laat, 2018). Thus, organisations developing AI technologies often resort to various legal vehicles provided by IP law, especially trade secret protection, as well as data ownership restrictions, non-disclosure agreements, and other contractual provisions (Foss-Solbrekk, 2021; Pasquale, 2015; Tschider, 2021).

However, opacity enabled by such legal mechanisms may also serve as a convenient means for both corporations and governments to conceal both legal and illegal practices. The latter can mean such conduct as abuse of dominant position, discrimination or violation of other fundamental rights. Burrell (2016) refers to this opacity as *intentional corporate or state secrecy,* with a number of critics pointing to algorithmic secrecy as a big challenge for accountability and fairness in applied AI. This can for example relate to data-driven markets (Pasquale, 2015) or smart city transparency (Brauneis & Goodman, 2018), often in contexts where corporate systems perform public functions. This challenge has received increased attention, not only with regards to end-users, such as citizens, patients or consumers, but also as an issue of distorted competition driven by large-scale digital platforms (cf Larsson,

2021). Such problematic consequences of opacity prompt the need to curb its scope. Some limitations to the trade secret protection, for example, are provided within law, such as by the Trade Secret Directive, whereby trade secret holders may be obliged to disclose relevant information due to the public interest. This is an attempt of balancing interests. Other restrictions stem from competition law, which prevents the use of trade secrets as a means to abuse market dominance. Courts may invoke human rights protections, such as respect for private and family life (see for example the Dutch *SyRI* case concerning digital welfare fraud detection). The judicial approach is, however, dependent on drawn-out, cumbersome and sometimes expensive legal and administrative processes.

Another method to scrutinise AI systems is to appoint certain entities, such as competent authorities or auditing bodies, to examine the systems under a confidentiality regime. In this context, Pasquale (2015) proposes a *qualified transparency* approach for data-driven markets to counter some of the intentional opacity in the shape of proprietary claims. In a somewhat similar argument, but aimed at how to think of and handle aspects of gaming for machine-learning systems, de Laat (2018) argues that full public transparency may render "perverse effects", and particularly advocates for full transparency for oversight bodies as the only feasible option. Recent developments in EU policy point to this line of reasoning, especially the newly-proposed AI Act. It seeks to regulate high-risk AI systems by making them subject to a special compliance regime, giving competent authorities the right to access the systems' source code. This governance approach has been interpreted by Mökander et al. (2021) as a Europe-wide ecosystem for conducting AI auditing. However, in accordance with the draft AI Act, the resources foreseen for enforcing the regulation only provide between 1-25 extra full-time staff per Member State, which Veale et al. consider to be "dangerously optimistic" (Veale, 2021).

Although legal opacity of AI systems may be justified in certain cases, efforts to provide a more effective system to "limit the limitations" on transparency by legal or technical means, are intense and ongoing. The proposed methods to scrutinise AI technologies depend on either slow-paced judicial and administrative decisions, or the review of systems by competent authorities or auditing bodies under the confidentiality rule, with arguably deficient resources at their disposal. It therefore remains to be seen whether the measures to provide more AI transparency in the EU will be sufficient in order to address the negative aspects of legally warranted opacity.

Discussion: Observations

In this chapter, we have scrutinised and aimed to place the often explainability-focused notion of transparency in contemporary AI governance discourse, into an interdisciplinary understanding of the concept. Firstly, we have pointed to some of the critique in the XAI domain in order to contrast this to mediation. While part of this critique stresses the importance of taking different "audiences" into account, we have tried to deepen the abilities of these audiences in terms of the rich literature on literacy found in information studies. Lastly, which is important not the least in light of the central role the concept of transparency plays in the contemporary governance discourse on AI, we have placed all of these three facets into the framework of trade-offs provided in law. Here we acknowledge transparency as a form of governance in itself, that has many interests to take into account.

It is not new and unique for the development of AI, that processes of transparency are something that can both reveal and conceal, and sustain (or exacerbate) as well as disrupt power structures (c.f., Strathern, 2000, Fenster, 2015, Hansen and Flyverbom, 2015). Yet, the need for transparency is something that is moving to the foreground as AI implementations increase and expand in society, and the consequences of them and automation of decision-making become increasingly apparent and profound. In some cases, even Kafkaesque in its opaqueness and difficulty to object when interwoven in bureaucratic and technical complexities (Vredenburgh 2022). The four facets we discuss all play their part in the process of making AI transparent, but they are also interdependent of (possibly) conflicting interests of the plethora of actors involved, and with the intricate data ecologies and infrastructures in which AI-systems and technologies come to be, and come to use.

To counterbalance the "ocular-centric" notion of transparency (Koivisto, 2022), we need to focus more on how what we cannot see actually gets mediated and brought to our attention. The procedural and interface-related aspects of transparency we address in terms of mediation above not only points to how AI-systems often are attempted to be made more scrutable and explainable, but also that a posed ideal of explainability is heavily depending on mediation as such. Literacy is beneficial for certain types of empowerment. As a policy-instrument, it can however also lead to problematic effects by a strategic approach for larger players in a digitised society to tilt accountability towards overwhelmed end-users or "data subjects". How to enact transparency is by no means a neutral process, but value-laden and political. Critical analyses discuss "transparency washing" as a strategy whereby a focus on transparency can act as obfuscation from more substantive and fundamental questions about the concentration of power (Zalnieriute, 2021, p. 139). From a legal point of view, several scholars address a more complex issue of transparency, in the sense of being able to manage both legally justified claims for opacity and undesirable results of too much transparency (cf de Laat, 2018). However, the private-public complexity, in terms of proprietary claims of secrecy, has been seen as a problem for oversight and civic participation, and has led to arguments for strong oversight bodies.

What do these four facets of AI transparency add up to? They are entangled as concepts and realities, and build upon the sociotechnical assemblages of humans and non-humans that are forming AI development and use. Even though it is commonly conceptualised as a state, AI transparency can also be conceptualised as performative, processual and under negotiation (Cellard 2022). It is a process that is limited by matters such as what can actually be known and explained of how a system operates (XAI), the choices made in what and how information about it is conveyed (mediation), the expertise of oversight bodies and individual capabilities (literacy) and the interpretations and constraints regarding what information is required to be accessible (legality).

Conclusions

The aim of this chapter was to critically unpack the conceptual vagueness of AI transparency. This is particularly motivated by recent focus on transparency in AI policy. To this end, we construct our analysis of AI Transparency into four facets. Firstly, (1) as

explainability (XAI) is an expanding field in AI, we argue for a need for it to be complemented by more explicit focus on the (2) *mediation* of AI-systems functionality, as a communicated artefact. Furthermore, in the policy discourse on AI, the importance of (3) literacy is underscored. Subsequently, we draw from the rich literacy literature in order to show both promising and troubling consequences of this. Lastly, therefore, we argue for transparency being a form of governance, albeit laden with positive connotations – that more transparency is better – which we critically break up within a (4) legal framework set to balance between a multitude of interests. By these four facets, we examine a particularly complex concept in dire need of clarification, due to its central position in the governance of increasingly automated and AI-dependent corporate and governmental activities.

References

Alloa, E. (Ed.). (2022). *This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor*. Leuven University Press.

Andrejevic, M., Fordyce, R., Luzhou, L., Trott, V., Angus, D., & Ying, T. X. (2022). Ad Accountability Online: A methodological approach. In Pink, S., Berg, M., Lupton, D. and Ruckenstein, M., eds. *Everyday Automation* (pp. 213-225). Routledge.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale JL & Tech.*, *20*, 103.

Burrell, J. (2016). How the Machine 'Thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1-12.

Cellard, L. (2022) Algorithmic Transparency: On the Rise of a New Normative Ideal and its Silenced Performative Implications. In: Alloa E (ed) *This obscure thing called transparency : politics and aesthetics of a contemporary metaphor.* Leuven Leuven University Press, pp.119-144.

Cranor, L. F., Hoke, C., Leon, P. G., & Au, A. (2015). Are they worth reading-an in-depth analysis of online trackers' privacy policies. *ISJLP*, *11*, 325.

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, *39*(2), 101666.

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philosophy & technology*, *31*(4), 525-541.

de Vries, K. (2021). Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM. *Critical Analysis of Law*, *8*(1), 121-138.

Diakopoulos, N. (2020). Transparency. In Dubber, M.D., Pasquale, F., and Das, S. (eds.) *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Draper, N. A., & Turow, J. (2019). The corporate cultivation of digital resignation. *New media & society*, *21*(8), 1824-1839.

Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2022). Achieving a Data-driven Risk Assessment Methodology for Ethical AI. *Digital Society* 1(2): 1-27.

Fenster, M. (2015) Transparency in search of a theory. *European Journal of Social Theory* 18(2): 150-167.

Forssbaeck, J., & Oxelheim, L. (Eds.). (2014). *The Oxford Handbook of Economic and Institutional Transparency*. Oxford Handbooks.

Foss-Solbrekk, K. (2021). Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly. *Journal of Intellectual Property Law & Practice*, *16*(3), 247-258.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11).

Haider, J., & Sundin, O. (2022). *Paradoxes of Media and Information Literacy: The Crisis of Information* (1st ed.). Routledge.

Hansen, H. K., & Flyverbom, M. (2015). The politics of transparency and the calibration of knowledge in the digital age. *Organization, 22*(6), 872-889.

Haresamudram, K., Larsson, S. & Heintz, F. (2022) Three Levels of AI Transparency, *Computer*, special issue on Trustworthy AI.
Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).
Jandrić, P. (2019). The Postdigital Challenge of Critical Media Literacy. *The International Journal of Critical Media Literacy*, 1(1), 26-37.
Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.
Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14.
Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society,* 22(14), 2081-2096. doi:10.1080/1369118X.2018.1477967
Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency* (Study No. PE 624.262) Panel for the Future of Science and Technology, Scientific Foresight Unit (STOA), European Parliamentary Research Service.
Koivisto, I. (2022) *The Transparency Paradox: Questioning an Ideal.* Oxford: Oxford University Press.
Koulu, R. (2020). Human control over automation: EU policy and AI ethics. *Eur. J. Legal Stud.*, *12*, 9.
Larsson, S. (2017). *Conceptions in the code: How metaphors explain legal challenges in digital times*. Oxford University Press.
Larsson, S. (2019). The socio-legal relevance of artificial intelligence. *Droit et Société, 103*(3), 573-593.
Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, *7*(3), 437-451.
Larsson, S. (2021). Putting trust into antitrust? Competition policy and data-driven platforms. *European Journal of Communication*, *36*(4), 391-403.
Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2).
Larsson, S., Ingram Bogusz, C., & Andersson Schwarz, J. Eds. (2020) *Human-Centred AI in the EU. Trustworthiness as a strategic priority in the European Member States*. Brussels: European Liberal Forum.
Larsson, S. Jensen-Urstad, A. & Heintz, F. (2021) Notified but Unaware. Third Party Tracking Online, *Critical Analysis of Law* 8(1): 101-120.
Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16.3*, 28.
Lloyd, A. (2019). Chasing Frankenstein's monster: information literacy in the black box society. *Journal of Documentation,* 75(6), 1475-1485.
Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1-16.
Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.
Miller, T., Hoffman, R., Amir, O., & Holzinger, A. (2022). Special Issue on Explainable Artificial Intelligence (XAI). *Artificial Intelligence*, 103705.
Miller, T., Howe, P., Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.
Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288).
Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Conformity Assessments and Post-market Monitoring: a guide to the role of auditing in the proposed European AI Regulation. *Minds and Machines*, 1-28.
Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, 58(1), 504-509.
Pasquale, F. (2015). *The Black Box Society*. Harvard University Press.
Qi, Z., Khorram, S., & Fuxin, L. (2021). Embedding deep networks into visual explanations. *Artificial Intelligence*, *292*, 103435.

Ribera, M., & Lapedriza, A. (2019). Can We Do Better Explanations? A proposal of user-centered explainable AI. In *IUI Workshops* (Vol. 2327, p. 38).
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.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.
Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham L. Rev.*, *87*, 1085.
Stanfill, M. (2015). The interface as discourse: The production of norms through web design. *New Media & Society*, *17*(7), 1059-1074.
Strauß, S. (2021). "Don't let me be misunderstood": Critical AI literacy for the constructive use of AI technology. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, 30(3), 44-49.
Strathern, M (2000) The Tyranny of Transparency. *British Educational Research Journal* 26(3): 309-321.
van Nuenen, T., Ferrer, X., Such, J. M., & Coté, M. (2020). Transparency for whom? assessing discriminatory artificial intelligence. *Computer*, *53*(11), 36-44.
Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019). (Un) informed consent: Studying GDPR consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security* (pp. 973-990).
Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97-112.
Vredenburgh, K (2022) The Right to Explanation*. *Journal of Political Philosophy* 30(2): 209-229.
Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.
Zalnieriute, M. (2021). "Transparency-Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism. *Critical Analysis of Law*, *8*(1): 139-153.