AI and Ethics (2023) 3:585–600
https://doi.org/10.1007/s43681-022-00190-4

ORIGINAL RESEARCH

Moral transparency of and concerning algorithmic tools

Paul Hayes[1,2] · Ibo van de Poel[3] · Marc Steen[4]

Received: 16 February 2022 / Accepted: 6 June 2022 / Published online: 20 June 2022
© The Author(s) 2022

Abstract

Algorithms and AI tools are becoming increasingly influential artefacts in commercial and governance contexts. Algorithms and AI tools are not value neutral; to some extent they must be rendered knowable and known as objects, and in their implementation and deployment, to see clearly and understand their implications for moral values, and what actions can be undertaken to optimise them in their design and use towards ethical goals, or whether they are even suitable for particular goals. Transparency is a term with variable uses and interpretations, a problem which can challenge its use in design and policy. Here, we attempt to further clarify transparency. We argue that transparency is the state of affairs that obtains when relevant and understandable information about some X is available and accessible to some target audience (A), so that this information is sufficient for A for the purpose (P). Moreover, we connect this conceptualisation with transparency's moral value, where P is to provide an account about X's supportive or conflicting relationship with relevant values and goals. Such teleological ends in our context here can be the ability to account for the degree to which an algorithm, process or organisation respects certain values and is conducive to (social) goals.

Keywords AI · Algorithms · Transparency · Ethics · Values

1 Introduction

Algorithms and Artificial Intelligence (AI) are gaining great purchase in a variety of industrial, commercial, and governance contexts and are viewed as important sources of actionable information, vested sometimes with great authority. In many of their contexts they can contribute towards processes and actions that produce significant impacts on individuals.

To some extent they must be rendered knowable and known as objects, and in their design, implementation and deployment phases, to see clearly and understand their implications for moral values, and what actions can be undertaken to optimise them in their design and use towards ethical goals, or whether they are even suitable for particular goals.

In this paper, we closely examine the concept of transparency and its attendant uses with regards to algorithms before proposing an account of moral transparency. We note that complete transparency of an algorithm as an object itself (in its totality of constituting parts) is not always necessary or sufficient, but to support accountability and responsibility, a variety of aspects of the algorithm and its embedding are necessary. Therefore, we are not merely concerned with transparency of the algorithm, as such, but also transparency concerning the algorithm or AI tool (its larger environment).

In what follows, we conceptualise transparency as a state of affairs conducive to the production and acquisition of information about some X, arguing that such a state of affairs has the properties of information that is available, accessible, understandable, and relevant for its various purposes. We will explain these properties in detail. Moreover, we understand transparency teleologically, inasmuch as it is of some X, for some audience (A), and for a particular purpose (P). Furthermore, we connect this understanding to the ethical

✉ Paul Hayes
paul.hayes@tudublin.ie

Ibo van de Poel
I.R.vandePoel@tudelft.nl

Marc Steen
marc.steen@tno.nl

[1] SFI ADAPT Research Centre and Graduate School of Creative Arts and Media, Technological University Dublin, Dublin, Ireland

[2] Trilateral Research Ltd., Waterford, Ireland

[3] Ethics and Philosophy of Technology, Values Technology and Innovation, TU Delft, Delft, The Netherlands

[4] Human Behaviour and Organisational Innovations, TNO, The Hague, The Netherlands

importance of transparency, that is, how it can cast light on the ethical acceptability of X and how such knowledge can empower human agency and verify or highlight value supports or conflicts and allow us to respond to them. Therefore, more specifically, we will argue that:

> *Moral* transparency is the state of affairs that obtains when relevant and understandable information about some X is available and accessible to some target audience (A), so that this information is sufficient for A for the purpose (P) of providing an account about X's supportive or conflicting relationship with relevant values and goals.

We open in Sect. 2 by laying out our understanding of algorithms and artificial intelligence. In Sect. 3, we remind the reader of the value implications inherent in the design and use of different algorithms and AI tools. In Sect. 4, we lay out extensive discussion of the various elements of transparency, including reference to its epistemic nature and describing its core features, especially in relation to algorithms and AI tools. Finally, in Sect. 5 we provide an account of moral transparency, produced from a synthesised understanding of transparency's moral and epistemic uses.

2 Algorithms and artificial intelligence

An algorithm can be defined as a finite set of steps that are carried out to solve a problem, and "[i]f we adopt a data-centric view, we use an algorithm to transform some data, which describe a problem, to some form that corresponds to the problem's solution" [1]. Algorithms are basic mathematical artefacts, and are not fundamentally autonomous. They are however often embedded in software tools and computational systems (themselves embedded in wider social, political and economic systems)—they can be programmed and coded to eventually be efficiently executable by computers [1, 2].

We understand AI tools as those where problems are solved in software tools or programmes that build and train their models through extraction of rules from training data using Machine Learning (ML) algorithms (either through supervised processes where labelled data are used, or unsupervised processes where it is not) [3]. Linear regression is one example of a supervised learning algorithm, the goal of which is the accurate estimation of an output value based on some input values [3].

Pattern recognition then is usually the task associated with ML, and succinctly according to Mark Coeckelbergh, "[a]lgorithms can identify patterns or rules in data and use those patterns or rules to explain the data and make predictions for future data" [4].

Another technique associated with AI is Deep Learning which can, according to Panos Louridas, be explained as "[n]eural networks that consist of many hidden layers, arranged such that succeeding layers represent deeper concepts, corresponding to higher abstraction levels" [1]. Deep Learning algorithms in particular may be especially opaque due to their complexity. The process of the data input and output through to decision in such multi-layered instances of ML can render them as black boxes that elude understanding [4–6].

The advantage of ML algorithms is that they can extract actionable insights from large amounts of data [7], and thus they can effectively automate large-scale cognitive workloads towards solving problems.

Having explored what we mean by algorithms and artificial intelligence, we will now move on to explore in more detail their value-laden nature and the necessity of transparency with regards to uncovering these value implications.

3 The value-ladenness of algorithms

Algorithms and AI tools are very influential artefacts across many of the environments in which they are deployed [4]. With their potentially actionable insights producing opportunities for the management and allocation of resources, and the general tailoring of policy towards addressing some problem(s), as well as their general capacity to capitalise on and contribute to the information economy, they have arguably (and justifiably) begun to dominate discussion of values and ethics in information and communication technology. Their place in informing and driving decision-making, sometimes perhaps uncritically, has led to the coinage of neologisms such as "algocracy" [8, 9], conveying an acute awareness of a potential drift towards governance through algorithms.

Algorithms are not value neutral nor without adverse implications for values, despite their perceived impartiality they exist within complex socio-technical systems and influence (and are influenced by) relations between composing human agents and patients [10–13]. They have been subject to criticism and investigation with regard to nature of the training data they use, and how such data, should it reflect discriminative practices, can help perpetrate negative or "pernicious" feedback loops where previous patterns are reinforced by the algorithm [11, 14].

Algorithms have evidenced potential embedded racial bias, as an investigation of equivant's COMPAS recidivism predictive tool by ProPublica argued [15]. ProPublica found that Black defendants were 77% more likely to be flagged as higher risk of committing violent crimes in the future than Whites, and 45% more likely to commit any kind of crime, by the algorithm—statistics not borne out in reality [15].

Such embedded algorithmic bias can translate into racial discrimination if it (mis)informs human action (discriminative parole or sentencing decisions for example).

Even in private sector hiring, the utilisation of algorithms can potentially disproportionately and adversely impact groups and individuals based on personal characteristics. The case of Amazon's experimental candidate evaluation algorithm is well known—an algorithm that was designed for the purposes of scoring job applications which was found to favour applications from men because much of the training data consisted of historical applications that were indeed mostly from men [16].

The adoption of algorithms by both public and private actors has resulted in artificially bolstered gate-keeping of access to both public and private resources, potentially deeply impacting the economic welfare of vulnerable people whom are scarcely positioned to challenge their weighty decisions that can include access to housing, healthcare and credit [17].

Algorithms implicate other values beyond fairness, including (and certainly not limited to) privacy. An enduring example of algorithms' capacity to impact privacy is that of the case involving an algorithm used by Target that correctly inferred the pregnancy of a teen girl in America based on her purchases, which resulted in baby coupons being sent to her home and revealing the pregnancy to her father [18].

There are aspects then of algorithms that can predispose them towards misuse, and problems in design and planned deployment or implementation can have significant consequences for people even as those using them may not completely understand why or how they are reaching important decisions, and those subject to those decisions may not even be aware, as in the case of the pregnant girl, that such decisions are being made before they are made.

4 Transparency

The use of algorithms in multifarious contexts can present an epistemic challenge for individuals and societies both as end-users and data subjects; end-users may be insufficiently informed to reasonably use them towards their intended goals even as these artefacts create new standards and ways of coming to decisions (and regardless of the provenance and reliability of their training data or the sensitivity of their application domain) [19, 20].

Given the significant value-ladnenness of algorithms and their use cases, and the need for verification of their ethical suitability for a particular goal, and their decision-subjects' capacity to respond to them, transparency holds significant instrumental value. It can also bolster other important values, including autonomy (knowledge of an algorithm's limitations may make it less a risk of driving one's judgements); accountability (knowledge of fault and causation can aid in appropriate allocation of blame); and fairness (knowledge of training data composition and sources may help to determine, for example, the presence of bias) [21]. There needs to be transparency about algorithms and the contexts in which they are embedded so that their potential or actual impacts can be monitored, challenged, or corrected. Without transparency about actual or likely harm, values cannot be upheld.

Depending on the discipline or professional area, accounts of transparency differ quite substantially [22, 23]. Often, such as in the areas of computer ethics, it involves ideas of proactive intentional disclosive communication, and enhanced accessibility and visibility of selective information by an organisation to inform information receivers (stakeholders) of relevant facts [22]. In other areas, such as software engineering and computer science, the concept has sometimes translated into something quite different, for instance referring to the invisibility of processes in a network [22, 23]. Transparency implies that something can be seen through—however here we are not interested in transparency as invisibility of an object [24], but clarity of an object unobscured by distortions and obstacles. We are interested in transparency as it relates to knowledge (see [24]), as a state of affairs conducive to, ideally, the construction and/or acquisition of knowledge.

The account we will give of transparency here will be a synthetic one, cognisant of its uses in the academic literature, and will as a result be a broadly applicable one. A generally applicable account of transparency is necessary because it is not only the algorithm as a mathematical construct that should be known, but also, the context in which it is embedded—often we must be able to see, unobscured inasmuch as possible, and understand its whole socio-technical assemblage (see [25]).

4.1 Opacity of algorithms and epistemic opacity

Machine Learning (ML) algorithms are often black boxes, the inner workings of which cannot be observed or understood, and even their implementation and deployment within their organisational contexts can be less than transparent [5, 26, 27]. This opacity can emerge for three reasons.

The first is secrecy, information about an algorithm may be restricted to protect its owner's intellectual property rights, or to prevent manipulation or gaming by prospective algorithmic subjects attempting to evade analysis [12, 28, 29]. Related to this, is the fact for any other number of reasons, the deployment and utilisation of algorithms across a variety of contexts may not be sufficiently publicised and persons can find themselves unknowingly subject to automated decisions [17].

The second is illiterate opacity, which stems from persons (probably most persons) being insufficiently knowledgeable or skilled to understand algorithms on a technical level [28]. The third is intrinsic opacity, which emerges from the complexity of dynamic ML algorithms that operate with scale and speed, and mathematical rather than semantic ontology, that may be even beyond the epistemic capability of domain experts to understand [28, 30]. As noted by Mittelstadt, Russell, and Wachter [31], "[w]hat distinguishes machine learning is its arbitrary black-box functions to make decisions. These black-box functions may be extremely complex and have an internal state composed of millions of interdependent values".

This opacity poses obvious difficulties. It can be difficult to determine or understand an algorithm's casual role in some event or state of affairs, or even if it was involved in it at all if the veil of secrecy is thick enough that even its existence is unknown to most (consider the slew of surveillance tools exposed by the Snowden revelations as an extreme example). Algorithms cannot be effectively challenged or indeed corrected without sufficient knowledge. They can influence the decisions of persons who do not fully understand their limitations.

Work in the philosophy of science provides additional insights on opacity which reflect on these above described problems and their epistemological aspects. This is relevant to our discussion here and warrants some exposition, as this philosophical work on opacity gives more form to what we expect of its opposite, transparency. In the influential work of Paul Humphreys [32], he elaborates on a definition of epistemic opacity in response to the novel epistemic challenges presented by computer simulations, which is:

> … a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process.

This definition speaks to the problems with ML algorithms that operate with scale, speed and beyond the semantic capabilities of human agents—such algorithms by this definition are almost inescapably opaque.

Claus Beisbart takes a different approach to the concept of opacity, preferring an ordinary language derived conception as a disposition to resist epistemic access (which includes both knowledge and understanding) [33]. Beisbart's argument at its most basic is a simple one, that something is opaque if it is difficult to know [33]. A core feature of Beisbart's argument is that a lack of knowledge or understanding from the human agent's perspective is not the crux of opacity, rather opacity stems from the barriers to obtaining knowledge and understanding [33]. Beisbart emphasises the importance of relevant features to promote knowledge and understanding of something, thereby addressing a potential gap in Humphreys' interpretation (as argued by Beisbart "[p]rocesses may become the object of different epistemic projects, depending on what the precise aims of the investigation are, and these aims determine what is relevant") [33]. Besibart's full explication of opacity is [33]:

> 1. The application of a method is opaque to the extent to which it is difficult for average scientists in the default setting to know and to understand why the outcome has arisen.
> 2. A method is opaque to the extent to which its typical applications are opaque.

These insights are invaluable to understanding the problem of opacity, transparency, and complex algorithmic artefacts. In particular, we are inclined to largely (but not strictly entirely) agree with Claus Beisbart's understanding of opacity and build upon this in particular in what follows. This approach is elegant in adhering to the ordinary language usage of opacity and we believe satisfactorily responds to some lack of clarity surrounding Humphreys' formulation (for instance, open questions relating to relevance), and furthermore it is useful in positing that a crucial element of opacity is not that it relates strictly to the pre-existing knowledge or understanding of epistemic agents, but barriers to coming to this knowledge or understanding. In our work that follows, a major concern is understanding these barriers so that knowledge can be acquired, and importantly that can ultimately be acted upon in morally meaningful ways.[1]

4.2 Transparency in relation to openness, interpretability, and explainability

4.2.1 Transparency and openness

Before proceeding yet further, we would like to make some additional remarks on what transparency is not, given the apparent conflation of transparency with simple openness in much of the literature, as well its perhaps unclear relationship with explainability and interpretability. If opacity is a main threat to transparency, it might seem attractive to think of transparency as openness. Indeed, a more or less common understanding of transparency is that it involves the provision or production of information (disclosure) to some agent outside of the immediate sphere of the disclosing agent that supports decision-making [23, 34–38].

[1] However, we still remain interested in individual epistemic capabilities of different agents which responds to the teleological aspect of our conceptualization of transparency.

For example, Patrick Lee Plaisance primarily defines transparent behaviour as honest forthrightness, and argues that [39]:

> ...*transparent behaviour* can be defined as conduct that presumes an openness in communication and serves a reasonable expectation of forthright exchange when parties have a legitimate stake in the possible outcomes or effects of the communicative act.

In a similar vein, Christopher Hood argues that transparency is "…the conduct of business in a fashion that makes decisions, rules and other information visible from outside" [40]. Turilli and Floridi[2] point out that in [23]:

> …the disciplines of information management studies, business ethics and information ethics, "transparency" tends to be used to refer to forms of information *visibility,* which is increased by reducing or eliminating obstacles. In particular, transparency refers to the possibility of accessing information, intentions or behaviours that have been intentionally revealed through a process of disclosure.

However, an approach to transparency excessively wed to intentional disclosure is problematic. Transparency and openness are neighbouring concepts [41], but they should not be conflated. Openness is primarily an attitude and related to intentional action, while transparency is a state of affairs.

Whilst disclosure of factual data for which they are directly responsible by an honest organisation or individual will often, but not necessarily, be a pre-requisite for transparency, intentional disclosure does not guarantee useful information. Moreover, information can become available without the consent of an agent we might expect to be normally or primarily responsible for its communication, or could leak accidentally.

Data may require interpretation and elaboration before it can be a useful tool for decision-making [34]. David Heald, citing Larsson [42], offers this distinction [36]:

> ....transparency extends beyond openness to embrace simplicity and comprehensibility. For example, it is possible for an organisation to be open about its documents and procedures yet not be transparent to relevant audiences if the information is perceived as incoherent. Openness might therefore be thought of as a characteristic of the organisation, whereas transparency also requires external receptors capable of processing the information made available.

A transparency which privileges openness might also privilege seeing over understanding, to the extent that related duties may be discharged simply from disclosure of potentially insufficient or even non-interpretable information, as Ananny and Crawford [25] argue, "[s]eeing inside a system does not necessarily mean understanding its behaviour or origins". Indeed, an understanding of transparency as openness, were it to make agents duty-bound simply to provide information (without selectivity, perhaps even intentionally so), can result in a situation where, "[i]ncreasing transparency can produce a flood of unsorted information and misinformation that provides little but confusion unless it can be sorted and assessed" [43].

Moreover, again, conflating transparency with openness fails to account for involuntary disclosure. When involuntary disclosures (leaks) occur from an organisation and we learn more about that organisation and its practices (again for example, Edward Snowden's disclosures about the NSA's surveillance apparatus), we might normally say that some transparency has been attained and yet the organisation itself was by no means open, nor forthright.

Another issue is a somewhat semantic one, and arises when we consider transparency of an artefact. Whilst an algorithm or AI can certainly be "open" to public inspection, that is, available and accessible, an artefact cannot necessarily be open in the sense of being forthright and honest. Beyond that, as we have said, if an artefact is available and accessible that does not mean it is understood and it seems unreasonable to say that something which is esoteric is transparent (at least not to a particularly useful degree), it is still possibly opaque.

For transparency to be a useful concept, it needs to incorporate features beyond openness, it needs to be teleological and therefore relevant to something, and entail the communication of useful information.

4.2.2 Transparency, explainability, and interpretability

Beyond openness, there is some lack of clarity surrounding the relationship and distinction between transparency, explainability, and interpretability. We argue that whilst these are neighbouring (if not to some degree integrative) concepts, they are also not necessarily the same and there are distinctions that should be initially clarified here.

This triumvirate of connected concepts appears often, especially in technical literature and literature relating specifically to algorithmic transparency and transparency on a technical level (also see especially the xAI literature). In this literature, essentially, transparency is defined variously as making the internal workings of an algorithm apparent (even by design) or, somewhat distinctively, comprehensible (or has the potential to be understandable), or at least its outcomes [31, 44–47]. Explanation can refer to ways of

[2] Their work remains cognisant of the multifaceted nature of transparency, however, noting that information provided should be "…true semantic content that can be used for epistemic purposes" [23].

transmitting information about the algorithm, particularly in terms of reasons, decisions, and evidence, in support of building trust and understanding of its appropriate use, and has been considered an interface between person and algorithm [31, 44, 47]. The EU High-Level Working Group on AI links technical explainability to the traceability and understandability of decisions to humans, and also emphasises the importance of the explainabiliy of decision-making processes and organisational embededness [48].

Finally, interpretability can also refer to the comprehensibility of a model and its output and how it behaves [31], or the capacity to provide human understandable interpretations that are in distinction to explanations, not necessarily meaningful to users [44].

To be clear here and reiterate, at a basic level we understand transparency as a state of affairs conducive to, ideally, the construction and/or acquisition of knowledge about some X. In this case, a transparent algorithm is one about which sufficient information exists or can be generated and shared that can answer relevant questions about it (transparency then, is decomposable to different questions or corresponds with different levels of abstraction). Explanation is a relevant feature of transparency, representing the meaningful communication of relevant information that contributes to more transparency about our X for a given purpose, or indeed arises from already transparent situations. Interpertability should be understood as being synonymous with comprehensibiliy and communicability of information, and then is also a feature of transparency and one which can also further contribute to that state.

We make these distinctions to ensure that transparency serves, for the most part, its classical and generalisable usage and does not refer solely to the comprehensibility of a model, but a situation where a model, or related aspects, can be comprehended either in whole or in part (transparency can be granular), and to ensure no conflations between the neighbouring and interdependent concepts of explainability and interpretability.

4.3 Transparency and knowledge

We have seen in previous sections that opacity is a threat to transparency (4.1) but that openness alone is not enough to overcome it, and even is not technically always necessary from directly responsible agents (4.2). What we need to add to the availability of information is it should allow its receivers to answer relevant questions about the object of concern [22]. Transparency then should be understood as an epistemic concern and in opposition to the concept of epistemic opacity as we explored earlier. We believe that transparency arises where something is knowable, and knowable by virtue of, as Beisbart fundamentally suggests in his account of opacity, a lack of barriers to coming to some knowledge. Additionally, we generally move beyond requirements for only epistemic access to some X, but believe that certain conditions (or a state of affairs) must be such that the information environment around our object of inquiry actually supports the inquiry and its goals.

The knowability of some X can be objective, as argued by Beisbart who argues for a standard of something being knowable and/or understandable by scientists [33]. This is an important minimum standard for a realistic conception of transparency; however we must acknowledge different epistemic capabilities and the need, in practice, to customise information to meet the needs of different question-askers with sensitivity for their varying epistemic needs and abilities.

We will shortly return to discussion about traits of a transparency that support knowability or even understanding. In the following (4.4), we will unpack transparency more in the context of the algorithm, and what different varieties of transparency can apply to this, and what we can learn from this to synthesise a new understanding of moral transparency.

4.4 Transparency of algorithms in context

As algorithms are the product of many decisions throughout a multi-stage process, and are deployed in live and dynamic contexts with direct implications for values, when we refer to transparency of algorithms we can be referring to a large collection of possible objects and circumstances. We are, it might be said, not only interested in transparency of algorithms, but transparency concerning them.

For this reason, it is helpful to point towards how questions of algorithmic transparency can be organised more clearly, and are relevant to technical, governance, and other contexts (see [41]). To do this, we endorse an approach of growing appeal in the ethics and philosophy of technology, which is to use Levels of Abstraction (LoAs) [49, 50].

Algorithms can be viewed and analysed at different LoAs. From a technical perspective, Giuseppe Primiero usefully divides computational systems into the following LoAs [2, 51]:

- Intention
- Specification
- Algorithm
- High-level programming language instructions
- Assembly/machine code operations
- Execution

Analysis of an algorithm (as a computational system) at each given LoA can answer important morally relevant questions—including whether there is bias in training or input data [4] for instance; whether a model is accurate;

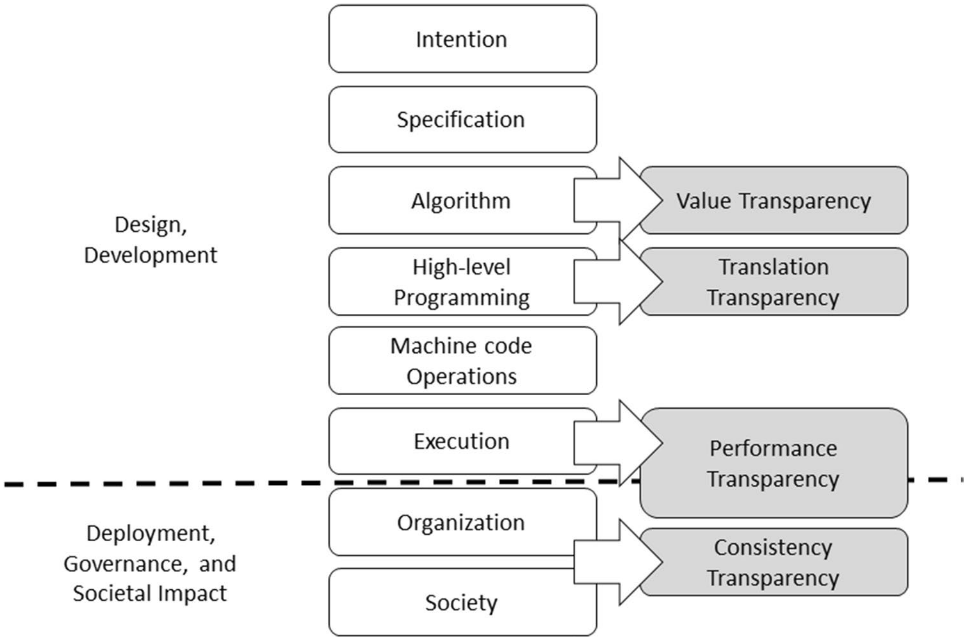

Fig. 1 Levels of abstraction and transparency of design, governance and societal impact (adapted from [51, 52])

whether the goals of the algorithm are ethical (intention); and whether it is free of bugs that threaten accurate and consistent performance (programming through to execution). The detail of such LoAs, that is, how much information must be made available to satisfactorily answer a question, again, depends on the audience. No single LoA however will help us identify all potential ethical issues or problems or respond to them, even if it might be useful in particular cases for individuals—algorithms need to be viewed from a variety of LoAs from the technical right through to their political or social context [41], or the LoAs relating to governance and societal impact, for a full and proper assessment of their justifiability.

When we speak of transparency of an algorithm from the perspective of a level of abstraction relating to its inception and design, with the overarching normative goal of evaluating their ethical acceptability, we are interested in what Michele Loi, Andrea Ferrario, and Eleonora Viganò call design transparency, or, "…the adequate communication of the essential information necessary to provide a satisfactory design explanation of such a[n algorithmic] system" [52].

For Loi, Ferrario, and Viganò,[3] such design transparency, in total, consists of [52]:

- Value Transparency—or transparency of the intended goal of the algorithm, including reasons and motivations to determine that the goal is indeed valuable.
- Translation Transparency—or transparency of the translation of the goal into machine language.
- Performance Transparency—or the provision of performance and impact metrics (e.g., "…classification accuracy and a meaningful comparison of group-related false-positive and false-negative rates").

Finally, Loi et al. introduce the concept of Consistency Transparency, which is proof that algorithms make their predictions by the same rules even if those rules cannot be "observed in operation" [52].

Figure 1 maps broadly a scheme of Levels of Abstraction, showing the movement from LoA to the varieties of transparency identified by Loi et al. [52]. By analysing an algorithmic system at different LoAs, or by making an algorithmic system knowable through these LoAs, varieties of transparency can be achieved that can help us know whether an algorithm is ethically fit for purpose, or perhaps what elements of its design or environment need to be changed.

We have added two additional LoAs beyond those of Primiero in Fig. 1, the organisation and society; both of which can certainly be decomposed into smaller frames of analysis again [51]. Ultimately, we do not want only transparency of an algorithm's design; we want its performance and consistency metrics applied to determine its societal impact. We want to be able to determine for example that it has not resulted in false positives, or whether it has resulted in increased discrimination against minorities. We also need to see how it has changed the environment in which it is

[3] For an alternative account of a different set of varieties of transparency see discussion of functional, structural, and run transparency by Kathleen A. Creel [53]. These varieties also ultimately fit within different technical LoAs, but in combination are important in understanding how an algorithm does or might interact with its wider socio-political or socio-technical environment.

used and the rules governing its use—we need information about its end-users to ensure that an algorithm's environment facilitates its ethical, responsible, and accountable use.

4.5 A general conceptualisation of transparency

We will now offer our general conceptualisation of transparency that integrates the various issues discussed in the previous sections. This general conceptualisation is based on three core ideas, namely (1) that transparency is not an attitude but a state of affairs; (2) that we should understand transparency teleologically as always being of something (X), for some audience (A), and for some purpose (P); and (3) that transparency requires information that meets a number of attributes (which we will shortly discuss) so that it can communicate knowledge that satisfies P, and ideally leads to understanding by A. We claim that this conceptualisation of transparency can be used across disciplines and across different levels of abstraction and organisational embedding of algorithms.

The first idea directly follows from our discussion in Sects. 4.1–4.3. We made three overarching points, taking some effort to argue that transparency is not simply openness, but is a largely epistemic and more multifaceted concept. First we stated openness is an attitude and can therefore only apply to intentional agents, not to algorithms. Second, openness supposes voluntary disclosure, while transparency can also result from involuntary disclosure of information. Third, information need not only become available and be accessible, but also need to be relevant and understandable; it needs to upgrade to knowledge (which is a state of affairs not an attitude).

The second idea poses that we should understand transparency always as being of some X for some agent A for some purpose P. While this idea might not be very controversial, we believe it is advantageous to come to a general characterisation of transparency that cuts across different LoAs and (hence) between different disciplines. In Fig. 1, we highlighted Loi et al.'s distinction of four main types of transparency (value transparency, translation transparency, performance transparency and consistency transparency) [52], which all fit this general characterisation. In fact, the labels (values, translation etc.) refer to examples of the purpose P for which we need transparency; as Fig. 1 also suggests at different LoAs, we are interested in the transparency of different objects X like e.g., an algorithm, a computer program, a decision procedure or the functioning of an organisation. What we further add to this is that transparency should be understood as being both objective (the possibility of being knowable by experts) and subjective or relative for some audience A. This implies that what counts as being accessible or understandable, for some information required for transparency, can be relative to the capacities and needs of the relevant agent, and cannot be always simply be determined in the abstract.

This brings us to the third element of our general conceptualisation: which is that transparency necessarily requires the presence of certain attributes to upgrade to knowledge or even understanding for the relevant audience A.[4] In the following sections, we discuss each of these attributes in more detail. After that we specifiy our proposed definition of transparency.

4.5.1 Availability

Availability is probably best considered as the existence of data or information pertaining to X. This is a critical requirement of transparency, as discussion of information, knowledge, access, understandability and relevance are all moot if there is no data or information available to develop and communicate. If there is no data, nor information, no questions about a particular X are answerable. Our X could be the many things, from the input data, to the model used, or information pertaining to how a decision was made or who was involved and how much control they had, or how it impacted an individual, and may be sought by A (a data/decision subject) for the P of contesting a decision.

4.5.2 Accessibility and findability

We define accessibility here as the *possibility of receiving available data or information by relevant audiences*. For instance a civil society organisation may ask for information pertaining to a recidivism prediction algorithm's model for the P of assessing its fairness, but the intellectual property holder may decline the request. In such a case, our X is available, but not accessible.

Findability too is important in this context. If information cannot easily be found, this also implies that it is not easily accessible. Important information could, for example, be released as a footnote in a digital report that is not easily yielded in web-search results.

4.5.3 Understandability

Understanding and understandability are a complex facet of transparency, and will require a little more exploration here than other attributes. First, it is important to find the demarcation between understanding and knowledge, which may not always be obvious. In the literature, there have been different thoughts on the relationship between knowledge

[4] Yu-Cheng Tu does an excellent job of identifying many of these attributes of transparency, which we adopt here as critical properties, albeit with variances in our interpretation of their meaning [22].

and understanding. Kvanvig [54], as described by Pritchard [55] argues that understanding is "…an epistemic standing that is closely related to knowledge," and one which is, "distinctively valuable". Furthermore, in the philosophy of science, it has been argued that the dominant thought has been that understanding is an equivalent of knowledge, that [55]:

> ….understanding why such-and-such is the case is equivalent to knowing why such-and-such is the case, where this is in turn equivalent to knowing that such-and-such is the case because of such-and-such.

From our perspective, we should understand there being a significant difference between knowing and understanding, whereby knowing is a state of possessing knowledge (whether we consider that a justified true belief or other interpretation), whilst understanding implies a greater level of command over the how and why questions of this knowledge.[5] For example, testimony can carry knowledge, but the knowledge obtained by an epistemic agent from this testimony may not yield a deep understanding of the issue, object or process in question (see Beisbart [33]).

Stephen R. Grimm [57], for example, describes the work of Zagzebski [58], of which he says:

> …on her view, understanding is fundamentally a matter of grasping how various pieces of information relate to one another; it is a matter of making connections among them, of seeing how they hang together.

So a not uncommon interpretation of understanding is the notion of "grasping", either at the relationships between interrelated information, as Luciano Floridi may say [56], or "…seeing connections among one's beliefs" [57].

Understandability, then, should be considered to be the state to which some phenomenon lends itself to understanding by an epistemic agent. If understanding is the grasp of connections, then we argue that understandability is the possibility of grasping those connections. Objectively, we might say that something is understandable if an expert can grasp those connections. Subjectively, something is understandable relative to the epistemic ability of the person receiving information.

We argue for a conception of transparency as a state of affairs ultimately conducive to acquisition of knowledge, and perhaps counter-intuitively have put in place the requirement of understandability prior to this knowledge, which may seem to be putting the cart before the horse. This is why it is also important to make the distinction between understandability and understanding. We propose that an object, the X which is the object of inquiry of our transparency (an AI model, perhaps), be understandable to promote transparency, under the assumption that knowledge derives from something that can be explained (and can be bolstered by explanations between agents). Such knowledge then can contribute to understanding, and then further knowledge.

The degree to which something is understandable may be more or less limited; transparency (and opacity) can come in degrees determined by this fact, yet it is a determinant of something being transparent, and the more understandable some X is, the more conducive it is to knowledge that can be acted on appropriately.

If something is understandable, then it should be possible to explain it to different audiences in relation to different levels of abstraction. xAI, the field of inquiry about explainable artificial intelligence, calls for "everyday" or "partial" explanations of an algorithm's functionality and behaviour in specific cases and is concerned with various methods for making AI tools more explainable and conveying relevant explanations to appropriate persons [31]. This movement then is concerned with the objective understandability of AI tools and their aspects, as well as, within its relevance criterion, subjective or relative understandabiliy vis-à-vis particular epistemic goals. Explanation is the method by which understandable information is transferred to a relevant audience. In the following subsection, we will explore what exactly we mean by relevance.

It should be noted that the xAI approach promotes and is promoted by transparency, however our approach to transparency is broader, and whilst encapsulating, referential to and endorsing xAI, should nevertheless be understood as being a broader umbrella and compatible with other uses (for example, the transparency of organisations, methods, and procedures).

4.5.4 Relevance

Yu-Cheng Tu usefully outlines this attribute of transparency, and though it is in the context of software engineering, Tu's definition provides an instructive starting point nonetheless [22]. Tu defines relevance as "…the degree to which the information obtained by stakeholders answers their questions" [22]. Transparency of X implies a well bounded question about that X which determines what kinds of information are necessary for answering or explaining it, or coming to some knowledge or understanding that can be put to some

[5] For more similar discussion, see Floridi [56], who gives the concept of knowledge (and information) an extensive treatment but differs by holding knowledge and understanding to a closer status, which is to say that knowing implies understanding, and where a lower threshold is met an agent may simply be "informed". We agree that information upgrades to knowledge, but for us the more immediately crucial distinction is between knowledge and understanding, which plays an important role in discussion of algorithms and AI, especially xAI. From our perspective, knowledge broadly implies a body of information about something, and understanding a significant insight into that body of information.

purpose. The kind of information made available should also be done with the epistemic capability of the question-asker in mind. Relevance then, as Beisbert suggests, is determined by the goal of the inquiry (and separately, who the inquirer is) [33].

So, information available and accessible must be able to satisfy with some degree of precision a question asked about X, and if such information, sufficiently relevant to the question, is not available, ideally (and arguably obligatorily depending on the purpose for which the information is sought) it should be produced and disclosed.

Epistemic capability of question-asking agents and understandability of information imparted is related to relevance, and is of key importance (see [31]). A technical or scientific explanation may not be relevant to their needs [31]. The algorithm's output in this situation remains opaque to that person subjectively—the information may not be strictly relevant for the person's epistemic needs, and beyond their epistemic capability. This is not to say that more complex information should not be available for consumption and further processing by experts, with more information intensive needs or even greater access needs, who will have interest in low levels of abstraction. An agent or patient therefore need not be given all of the available information, particularly in unwieldy or complex formats. Attributes of explanation (as a communicative and teleological act) are selective and teleological, they fulfil epistemic need without overwhelming epistemic capability, at which point the imparted information may serve to obscure rather than illuminate [31, 59, 60].

Determining what kind of information to share, and in what format it is shared, is contingent on epistemic capability and need. The subject of an algorithm's decision may be satisfied with a selectively chosen and clear everyday explanation. A scientific expert investigating some aspect of the algorithm may require a higher threshold of much more complex information, or a scientific or technical explanation. Relevance requires a determination of what a question-asker needs to know, and the elaboration or management of information that they can actually consume and hopefully upgrade to knowledge or understanding, depending on their capability. As argued by John Zerilli et al. design and physical level explanations may be excessive and irrelevant to those simply seeking reasons for algorithmic decisions [60].

Simple (relatively) explanation then can carry relevant information, imparting even knowledge that is arrived at from more thorough investigation of X. Relevance shows that questions about our X can be satisfied without a deep understanding of X as it decomposes in levels of abstraction, that is, our X does not need to be completely transparent to an agent, merely transparent enough to come to knowledge to respond to X in appropriate ways.

To illustrate this better, a loan applicant (or their advocate) (A) might ask for the decision reason (X) of a credit-scoring tool for the purposes (P) of contesting it. Relevance is determined by the epistemic need of A the applicant or their advocate, and the information communicated should be selected such that it supports their knowledge and understanding of X and capacity to contest any decision. Such information should be tailored to the epistemic ability of A, and could consist of a contrastive or counterfactual explanation or scientific one depending on the precise circumstances of the case [61].

5 A proposed definition of moral transparency

5.1 Moral transparency

So far, we have synthesised an understanding of transparency from different sources and applications and have framed it as a state of affairs conducive to the acquisition of knowledge with the normative and teleological role of being for some agent and for some purpose. We explored this with an interest in the opacity that often accompanies algorithms, and the need to eliminate or work around this opacity to support the rights of those affected by such algorithms within socio-technical assemblages and societal or moral values, and the responsibilities of those who use them. To that end, we will now undertake the task of securing the moral import of transparency by adjusting our definition to emphasise its moral role, where that P is explicitly to serve ethical or moral ends.

The various elements that we have discussed can now be combined in our proposed definition of moral transparency:

> Moral transparency is the state of affairs that obtains when relevant and understandable information about some X is available and accessible to some target audience (A), so that this information is sufficient for A for the purpose (P) of providing an account about X's supportive or conflicting relationship with relevant values and goals.[6]

Here we deliberately made the purpose for which we require transparency more specific, to validate that our X is ethical with a view to supporting morally relevant decision-making, for example, whether that is the kind of remediation that can come from accountability or the decision not to use an algorithmic system with knowledge that it is harmful. To that end, we believe that transparency is best understood as serving the (general) purpose of accountability and

[6] Then the capacity of A to provide an account of X (and its relationship with its values and goals) also requires some level of knowledge or understanding of X.

responsibility with respect to values and goals and the kinds of decisions that can follow this.

To this extent, and as recognised by Mark Coeckelbergh [26], the moral imperative of transparency can be understood to derive from the notion of responsibility, both in the sense of its being backward-looking and forward-looking. By backward-looking responsibility, we mean "….the (moral) obligation to account for what you did or what happened (and your role in it happening)" [62]. This is in contrast to forward-looking responsibility, which generally "…refers to the obligation to see to some state-of-affairs and to ensure discharge of entailed duties, which attach to one's relationship to some object or role" [21, 63].

Transparency is an important instrumental value that positively supports other values, and is necessary to support accountability (providing explanations for what we have done to those affected).

We need to be in a position to compile and provide information about our decisions and actions to affected or interested parties (or sometimes "moral patients") about how we or our AI tools came to these decisions or actions [21, 26]. This duty of transparency via accountability as answerability can further be justified or explained by an appeal to non-instrumentalism and the respect of human agency. That is, we must respect the agency and free will of those subject to our actions and decisions [39, 64], which naturally entails honestly informing them of the relevant circumstances leading to concrete impacts on their lives. From the backwards-looking perspective, transparency can be used to help validate respect for values—that a system has not resulted in untowards implications for fairness, autonomy, or privacy.

Moreover, we need to possess sufficient knowledge of the tools we use (and their physical environments) to ensure we wield them responsibly [26]. Part of this forward-looking responsibility rests on contemplating the impacts of the use of our tools and the potential consequences of our actions, and ensuring that they are suitable for their goals and purposes and no (or minimal) unintended or adverse effects arise from their design and deployment. Those who design and deploy AI tools then are responsible for making sure that their tools and their applications uphold or bring about values with minimal tension (for instance, economic welfare, security, fairness, autonomy). To do this, they must be informed sufficiently such that they can design or adapt them appropriately to their use case. We can see that transparency, as well as serving the important goal of accountability (arguably a key goal of xAI), is also an intrinsic feature of value-sensitive or ethics-by design approaches to technological development. Knowledge of a tool (whether existing or proposed) [26], its organisational embedding, and its operational context is promoted through transparency, and such knowledge empowers responsible actors to better meet the demands of our moral or societal values throughout software development cycles up to final use.

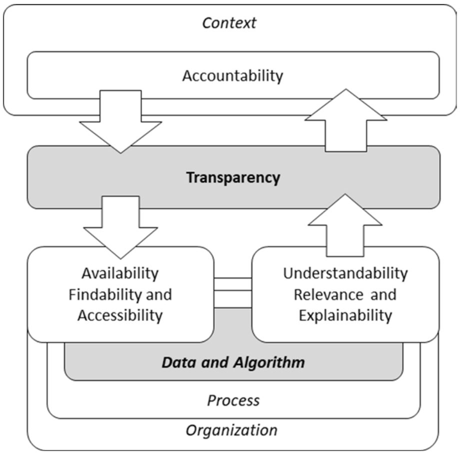
Context
Accountability
Transparency
Availability Findability and Accessibility
Understandability Relevance and Explainability
Data and Algorithm
Process
Organization

Fig. 2 Transparency of algorithms in context

From the forwards-looking perspective, transparency supports relevant actors in ensuring that a system can or will continue to uphold and/or respect values.

5.2 Implementing moral transparency

Figure 2 provides an illustrated account of some of our discussion so far and helps us to better describe our approach to moral transparency. Consider a situation where a civil society organisation (CSO)(A) asks the end-user of a credit scoring algorithm to validate the fairness of their algorithm. In this case, our moral transparency X is a reasonably comprehensive overview of the algorithm's design, implementation, and deployment, with a view to establishing its fairness (P), that is, that there is no bias and discrimination present in the design and use of the tool. The request may be comprehensive, but the information to be supplied need only be enough to validate the fairness of the tool.

To support transparency on fairness, the end-user[7] may wish to supply a variety of data that corresponds to different LoAs of the algorithm from input and training data (and other information presented in Fig. 1 as it relates to the question), organisational composition and processes for actioning outputs, and aggregate results of those outputs (for example, comparisons of credit scores by personal characteristics) and explanations for those outputs or any discrepancies—all tailored to the expertise present in the CSO or at least those persons the CSO may subsequently share the information with.

[7] Do note that such questions may be answered by agents that do not have direct responsibility for an algorithm, such as whistleblowers or journalists.

The transparency in this case is generalised, however we may also ask for individualised transparency, that is, an individual may ask the same end-user for an explanation concerning the fairness of a decision for them, and the end-user would be expected to, depending on the epistemic need and ability of the individual, provide information and reasons in support of the validity and fairness of the decision including input data and organisational decision (and appeal) procedures.

A convincing potential relationship between transparency providers and receivers has been outlined by Coeckelbergh, who argues that transparency is owed by responsible agents to moral patients, that is, the recipients or those affected by their actions and decisions [26]. This provides a useful start in defining transparency duties, but one we would like to make some effort here to define further.

We suggest that such an approach is generally correct but that the moral agent and patient relationship is not always clear in socio-technical systems. What is perhaps clearer is the notion of *relevant relations* between agents (and/or patients) and things [65]. Those responsible for transparency should be those with a relevant responsibility relation for some object (the algorithm's designer or end-user), and duties of transparency (in terms of to whom) should depend on a relevant relation to the agent holding a right to transparency (that may not be a moral patient, per se, but instead a potential advocate such as a civil society organisation) [21]. An Agent B[8] may hold a duty of transparency about something to another subject that is not strictly, or very apparently, a patient, for example the supplier of an AI tool should make various aspects of that tool transparent to Agent C, the end-user. An external entity (a civil society organisation or a journalist) can also provide transparency of the internal and organisational aspects of an AI tool, potentially without the cooperation of those responsible for its design and use, and they would hold such a responsibility for such transparency as a core element of the mission of their institution (see [65]).

All this is to say that we endorse a relational account of responsibility and transparency, and add to the warning that there is a challenge in disentangling these relations and it is worth noting that they are not strictly bi-directional nor strictly only between agents and patients. Fundamentally, our audience (A) may not be a patient in the strictest sense, when considering the vulnerability this might entail, but may simply be another responsible actor with transparency needs to bolster its own responsibilities.

In Fig. 3 we lay out a non-exhaustive framework of an ecosystem of transparency, that is, agents (and audiences) within the socio-technical system of an AI tool with information needs and duties to ensure relevant values are being upheld. In what follows, we will describe an at least provisional list of the types and kinds of information that each party should be rendering available (or requesting) in formats determined by the epistemic need and capacities of transparency recipients.

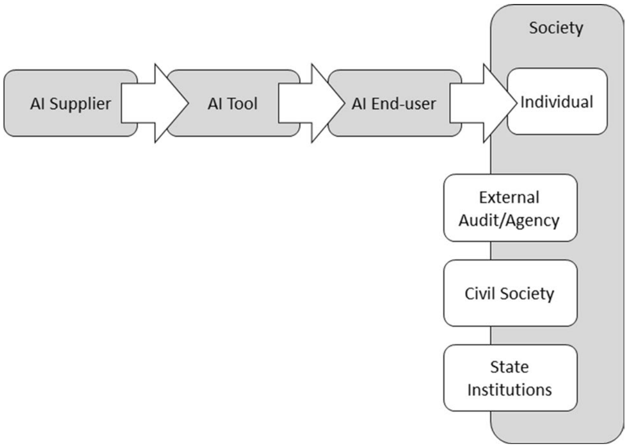

Fig. 3 Ecosystem of actors and artefacts in algorithmic transparency

AI Supplier: the supplier (representing the intellectual property holders and designers etc.) of an AI tool can be expected to release a variety of types of information necessary to help validate the compatibility of the tool with moral values and to help use the tool responsibly. Such information could pertain to the entity itself, including its corporate structure and the composition of the workforce (is it inclusive and diverse), whether it is a public or privately held organisation and where it is located. Such information can help us understand the standpoint (see [66]) of the creators of the tool, and whether, for example, they are located in a jurisdiction that respects the rule of law, thereby helping us understand likely fairness and legality implications of the tool.

AI tool: There is a wealth of information that should be made available about the AI tool, by persons who own it, designed it, use it and those with responsibilities for providing information in service of the public good (civil society). Useful for validating value support or tensions (and responding to those with responsibility) are the tool's goals and purposes; input data; training data; output data; programming language and code; the model and how it reaches it decisions to the extent that is possible; information pertaining to the development cycle (e.g., whether there was stakeholder input and feedback). Such information can be used, for example, to validate or assess implications for fairness and privacy.

AI end-user: Like the AI Supplier, the end-user should provide information as appropriate about its internal composition; the purposes and justification for which it uses the algorithm; the decision procedure and human involvement

[8] We have forgone Agent A in order to avoid conflation with Audience (A).

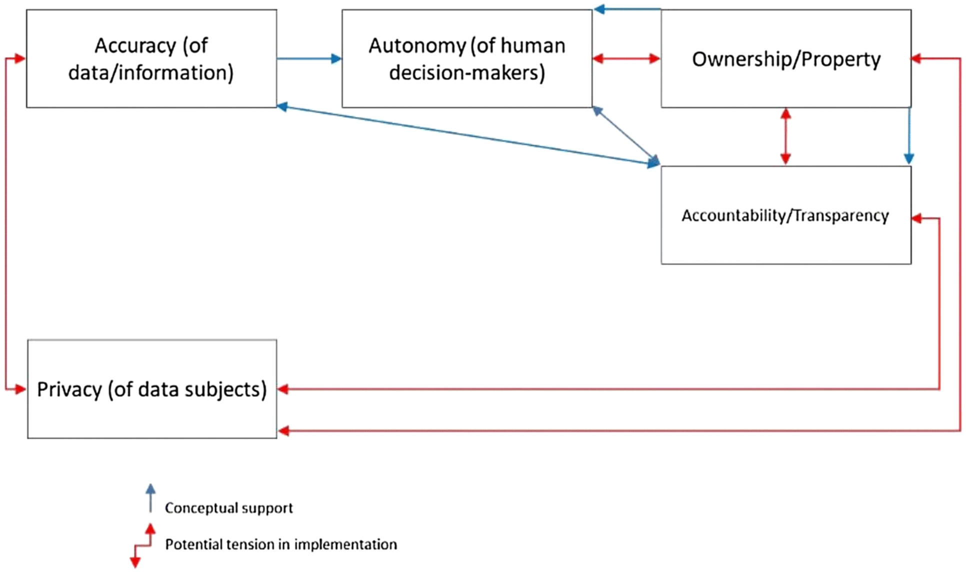
Accuracy (of data/information)
Autonomy (of human decision-makers)
Ownership/Property
Accountability/Transparency
Privacy (of data subjects)
Conceptual support
Potential tension in implementation

Fig. 4 Transparency and value tensions (adapted from [21])

in it; appeal procedures; other relevant internal governance information; results arising from the adoption of the tool; and it should list the actual supplier of the tool as well as the procurement procedure. Such information can help us validate or understand implications for fairness, and the autonomy of end-users.

External agency/auditor: An external agency or auditor should provide information as appropriate about its composition, rules of operation, and methodologies for assessing AI tools. See [67] for more substantial discussion on this general topic.

Civil society: Civil society should provide information about its mission, composition and methodologies for assessments it makes of AI tools. Its mission may in many cases be one anchored to moral transparency, that is, it validates whether an algorithm conforms to moral values (see Propublica's work regarding COMPAS for example).

State institutions: State institutions could represent auditors, AI end-users or even AI suppliers as the case may be and should provide information pertinent to their exact role.

The individual and society: We cannot hold the individual or society strictly as transparency duty holders, but they will require much of the preceding information to understand and respond to the consequences of the use of an AI tool for them. Those designing and using AI tools do however have an obligation to understand the ethical, legal, and social impact of their tools and should make efforts to understand the problem space and those who will be impacted by them, through direct engagement and communication if necessary.

5.3 Transparency, value tensions, and its limits

In previous research, we set about mapping value supports and tensions in the context of justice and security. With regards to transparency, we have noted that it can come into tension or conflict with other relevant values held by society (see Fig. 4 below, reproduced from the aforementioned article).

As evident from Fig. 4, we have argued that transparency, when at least unmoderated and in different permutations, can come into tension with values including privacy, accuracy, and autonomy. Transparency of an algorithm can risk privacy, for instance, if it could reveal personal details arising from input data, or other data about those subject to algorithmic decisions [68]. Transparency can come into tension with ownership and property to the extent that revealing too much about the model underlying algorithms and AI tools to a wide audience could put at risk the owner's intellectual property if it could be reverse engineered or duplicated [68]. Not depicted in the Figure is also the potential tension between accuracy and transparency [46], whereby more complex algorithms that are arguably more accurate are more conducive to opacity than transparency.

In our formulation of transparency, we are focussed on its instrumental potential to support other values, and we recognise that wholesale transparency, where every question is answerable to everyone, is neither desirable nor necessary to achieve moral transparency, or simply the validation of value and supporting achieving it. Instances where rendering something transparent results in harm (revealing personal details) would not be called for by moral transparency.

It is sufficient for enough information about an algorithm to be made available to verify that it will do no (disproportionate depending on the context) harm to its subjects, and such transparency may still be curtailed even to specified and confidential audiences such as a transparent and reliable auditing agency with sufficient enforcement powers (see [67]). Algorithms based on deep learning may never be completely transparent, or more specifically their internal processes, but this does not mean that they are so inscrutable as to preclude our moral transparency.

This also precludes the *absolute* necessity of explainability and interpretability of algorithms and their decisions, as this might only be required to the extent that it helps validate our values (although this militates against the use of opaque algorithms that defy transparency in high risk situations [69, 70]). It need not be morally obligatory then to design a transparent algorithm, so long as it can at least be demonstrated that an algorithm will not cause harm where it is deployed. An algorithm should also, we agree, not be held to a higher standard than a human would be [60]—we only need sufficient information and reason from human decision-makers in accountability and transparency procedures, not a God's eye insight into their cognitive processes.

The restriction of the provision of information to different agents, especially the public, is not new. Both the GDPR (see Article 23 for example) and the European Convention on Human Rights (see Article 10), for example, allow for derogation or limitation on the rights to freedom of information, normally where it is in service of security, the public interest, or scientific research for example. This too can apply to duties of transparency relating to algorithms, that is, those with transparency responsibilities can justify limiting the access and availability of relevant information to particular audiences where it is demonstrably necessary and proportionate (protecting personal information for example). What is important is that curtailments of transparency are well documented and justified, and that at least a sufficiently empowered agency remain capable of more thorough access to the justifiably confidential aspects of an algorithm, at least to the extent that an algorithm is involved in sensitive contexts like credit-scoring or parole decisions. To that extent, qualified transparency is also compatible with moral transparency.

6 Conclusion

Transparency is a vital instrumental value that plays an important role in supporting other values. We must have access to information, and ideally knowledge or even understanding, about value-laden objects and systems.

We have unpacked a synthetic account of transparency that builds on the literature in an effort to support a conceptualisation of transparency that can work more consistently across disciplines and subjects. We have moved the focus of transparency away from attitudinal attributes such as openness, to a more comprehensive list of attributes, such as availability, accessibility, understandability, and relevance because they support the acquisition of knowledge that can satisfy the teleological ends of transparency. Such teleological ends in our context here ultimately are the ability to account for the degree to which an algorithm, process or organisation respects certain values and is conducive to (social) goals and ultimately respond accordingly. When the goals of transparency are morally relevant, we are then talking about a particular kind of transparency, moral transparency—a concept itself motivating normative requirements.

Ultimately, many elements of algorithms, AI tools, and their environments should be known or knowable by the organisations using them and citizens subject to their decisions both directly and indirectly. To ensure values are being upheld, or will be upheld, the whole assemblage of algorithm through to society will need to be analysed at different levels of abstraction. Moral transparency requires proactive epistemic action. Algorithm designers should reasonably strive for explainable AI and all persons involved in design and deployment should investigate the risks inherent in their systems and make the outcomes of such investigations available to the public. There should be continuous information and knowledge construction processes producing morally relevant knowledge, that can be acted upon to support moral values, about algorithms and their risks and impacts (actual and possible).

Acknowledgements This research was funded by the Start Impulse Program of the Dutch National Science Agenda (NWA), under The Netherlands Organisation for Scientific Research, VW Data P4 (#400.17.605).

Funding Open Access funding provided by the IReL Consortium.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Louridas, P.: Algorithms. MIT Press, Cambridge (2020)
2. Angius, N., Primiero, G., Turner, R.: The philosophy of computer science. In E. N. Zalta (Ed) The Stanford Encyclopedia of Philosophy (Spring 2021.). Metaphysics Research Lab, Stanford University (2021). Retrieved from https://plato.stanford.edu/archives/spr2021/entries/computer-science/. Accessed 9 May 2021
3. Alpaydin, E.: Machine learning: the new AI. The MIT Press, Cambridge (2016)
4. Coeckelbergh, M.: AI Ethics. MIT Press, Cambridge (2020)

5. Hälterlein, J.: Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. Big Data Soc. **8**(1), 20539517211003120 (2021). https://doi.org/10.1177/20539517211003118
6. Umbrello, S., van de Poel, I.: Mapping value sensitive design onto AI for social good principles. AI and Ethics **1**(3), 283–296 (2021). https://doi.org/10.1007/s43681-021-00038-3
7. Kelleher, J.D., Tierney, B.: Data science. The MIT Press, Cambridge (2018)
8. Aneesh, A.: Virtual migration: the programming of globalization, Illustrated Duke University Press, Durham (2006)
9. Danaher, J.: The threat of algocracy: reality, resistance and accommodation. Philos. Technol. **29**(3), 245–268 (2016). https://doi.org/10.1007/s13347-015-0211-1
10. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. Law Rev. **104**, 671–732 (2016)
11. O'Neil, C.: Weapons of math destruction: how big data increases inequality and threatens democracy, 1st edn. Crown, New York (2016)
12. Ferguson, A.G.: The rise of big data policing: surveillance, race, and the future of law enforcement. NYU Press, New York (2017)
13. Kitchin, R.: Thinking critically about and researching algorithms. Inf. Commun. Soc. **20**(1), 14–29 (2017). https://doi.org/10.1080/1369118X.2016.1154087
14. Richardson, R., Schultz, J., Crawford, K.: Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice (SSRN Scholarly Paper). Rochester, NY: Social Science Research Network (2019). Retrieved from https://papers.ssrn.com/abstract=3333423. Accessed 28 Feb 2019
15. Angwin, J., Larson, J., Matu, S., Kirchner, L.: Machine bias. ProPublica. text/html (2016). Retrieved 19 Oct 2018, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
16. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (2018). Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 16 Apr 2022
17. Hao, K.: The coming war on the hidden algorithms that trap people in poverty. MIT Technology Review. (2020). Retrieved 16 Apr 2022, from https://www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/
18. Hill, K.: How target figured out a teen girl was pregnant before her father did. Forbes. (2012). Retrieved from https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=6d8242936668. Accessed 5 may 2022
19. Amoore, L., De Goede, M.: Governance, risk and dataveillance in the war on terror. Crime Law Soc. Chang. **43**(2), 149–173 (2005). https://doi.org/10.1007/s10611-005-1717-8
20. Amoore, L., Raley, R.: Securing with algorithms: knowledge, decision, sovereignty. Secur. Dialogue **48**(1), 3–10 (2017). https://doi.org/10.1177/0967010616680753
21. Hayes, P., van de Poel, I., Steen, M.: Algorithmsand values in justice and security. AI & Soc. **35**, 533–555 (2020). https://doi.org/10.1007/s00146-019-00932-9
22. Tu, Y.-C.: Transparency in software engineering (Thesis). ResearchSpace@Auckland (2014). Retrieved from https://researchspace.auckland.ac.nz/handle/2292/22092. Accessed 19 Oct 2018
23. Turilli, M., Floridi, L.: The ethics of information transparency. Ethics Inform. Technol. **11**, 105–112 (2009)
24. Menéndez-Viso, A.: Black and white transparency: contradictions of a moral metaphor. Ethics Inf. Technol. **11**(2), 155–162 (2009). https://doi.org/10.1007/s10676-009-9194-x
25. Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. **20**(3), 973–989 (2018). https://doi.org/10.1177/1461444816676645
26. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci. Eng. Ethics **26**(4), 2051–2068 (2020). https://doi.org/10.1007/s11948-019-00146-8
27. Valentino-DeVries, J.: How the police use facial recognition, and where it falls short. The New York Times. (2020). Retrieved from https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html. Accessed 8 May 2021
28. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. **3**(1), 2053951715622512 (2016). https://doi.org/10.1177/2053951715622512
29. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**(2), 2053951716679679 (2016). https://doi.org/10.1177/2053951716679679
30. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., Vinck, P.: Fair, transparent, and accountable algorithmic decision-making processes. Philos. Technol. (2017). https://doi.org/10.1007/s13347-017-0279-x
31. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279–288) (2019). New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287574
32. Humphreys, P.: The philosophical novelty of computer simulation methods. Synthese **169**(3), 615–626 (2009). https://doi.org/10.1007/s11229-008-9435-2
33. Beisbart, C.: Opacity thought through: on the intransparency of computer simulations. Synthese (2021). https://doi.org/10.1007/s11229-021-03305-2
34. Etzioni, A.: Is Transparency the best disinfectant? J Polit Philos **18**(4), 389–404 (2010). https://doi.org/10.1111/j.1467-9760.2010.00366.x
35. Fleischmann, K.R., Wallace, W.A.: A covenant with transparency: opening the black box of models. Commun. ACM **48**(5), 93–97 (2005). https://doi.org/10.1145/1060710.1060715
36. Heald, D.: Varieties of transparency. In C. Hood, D. Heald (Eds) Transparency: the key to better governance? (pp. 25–43). Oxford: Oxford University Press for The British Academy. (2006). Retrieved from https://global.oup.com/academic/product/transparency-the-key-to-better-governance-9780197263839?q=9780197263839&lang=en&cc=gb. Accessed 19 Oct 2018
37. Hulstijn, J., Burgemeestre, B.: Design for the values of accountability and transparency. In J. van den Hoven, P. E. Vermaas, I. van de Poel (Eds) Handbook of ethics, values, and technological design: sources, theory, values and application domains (pp. 1–25) (2014). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6_12-1
38. Vaccaro, A., Madsen, P.: Firm information transparency: ethical questions in the information age. In: Berleur, J., Nurminen, M.I., Impagliazzo, J. (eds.) Social informatics: an information society for all? In remembrance of Rob Kling, pp. 145–156. Springer, US (2006)
39. Plaisance, P.L.: Transparency: an assessment of the kantian roots of a key element in media ethics practice. J. Mass Media Ethics **22**(2–3), 187–207 (2007). https://doi.org/10.1080/08900520701315855
40. Hood, C.: Accountability and transparency: siamese twins, matching parts, awkward couple? West Eur. Polit. **33**(5), 989–1009 (2010). https://doi.org/10.1080/01402382.2010.486122
41. Larsson, S., Heintz, F.: Transparency in artificial intelligence. Internet Policy Review, 9(2). (2020)Retrieved from https://policyreview.info/concepts/transparency-artificial-intelligence. Accessed 8 May 2021

42. Larsson, T.:. How open can a government be? The Swedish experience. In V. Deckmyn & I. Thomson (Eds) Openness and Transparency. European Institute of Public Administration.
43. O' Neill, O.:. BBC—Radio 4—Reith lectures 2002—a question of trust - lecture 4—Trust and Transparency. (2002)Retrieved 13 Apr 2020, from http://www.bbc.co.uk/radio4/reith2002/lecture4.shtml
44. Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M.: Explainable artificial intelligence: an analytical review. WIREs Data Min. Knowl. Discovery **11**(5), e1424 (2021). https://doi.org/10.1002/widm.1424
45. Deloitte.: Transparency and responsibility in artificial intelligence a call for explainable AI. (2019). Retrieved from https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf. Accessed 17 June 2022
46. Hagras, H.: Toward human-understandable, explainable AI. Computer **51**(9), 28–36 (2018). https://doi.org/10.1109/MC.2018.3620965. (**Presented at the Computer**)
47. Schraagen, J.M., Kerwien Lopez, S., Schneider, C., Schneider, V., Tönjes, S., Wiechmann, E.: The role of transparency and explainability in automated systems. Proc. Human Factors Ergon. Soc. Annu. Meeting **65**(1), 27–31 (2021). https://doi.org/10.1177/1071181321651063
48. EU HLEG AI.: Requirements of trustworthy AI. FUTURIUM—European Commission. Text. (2019). Retrieved April 20, 2022, from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1
49. Umbrello, S.: Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach. Ethics Inf. Technol. (2021). https://doi.org/10.1007/s10676-021-09588-w
50. Floridi, L.: The ethics of information. OUP Oxford, Oxford (2013)
51. Primiero, G.: Information in the philosophy of computer science. In: Floridi, L. (ed.) The Routledge handbook of philosophy of information, pp. 90–106. Routledge, London (2016)
52. Loi, M., Ferrario, A., Viganò, E.: Transparency as design publicity: explaining and justifying inscrutable algorithms. Ethics Inf. Technol. (2020). https://doi.org/10.1007/s10676-020-09564-w
53. Creel, K.A.: Transparency in complex computational systems. Philos. Sci. **87**(4), 568–589 (2020). https://doi.org/10.1086/709729
54. Kvanvig, J.L.: The value of knowledge and the pursuit of understanding, 1st edn. Cambridge University Press, Cambridge (2003)
55. Pritchard, D.: Knowledge, understanding and epistemic value. R. Institut. Philos. Supplements **64**, 19–43 (2009). https://doi.org/10.1017/S1358246109000046
56. Floridi, L.: Semantic information and the network theory of account. Synthese **184**(3), 431–454 (2012). https://doi.org/10.1007/s11229-010-9821-4
57. Grimm, S.R.: Is understanding a species of knowledge? Br. J. Philos. Sci. **57**(3), 515–535 (2006)
58. Zagzebski, L.T.: Virtues of the mind: an inquiry into the nature of virtue and the ethical foundations of knowledge. Cambridge University Press, New York (1996)
59. Miller T.: Explanation in artificial intelligence: insights from the social sciences. arXiv:1706.07269 [cs] (2017). Retrieved from http://arxiv.org/abs/1706.07269. Accessed 22 May 2019
60. Zerilli, J., Knott, A., Maclaurin, J., Gavaghan, C.: Transparency in algorithmic and human decision-making: is there a double standard? Philos. Technol. **32**(4), 661–683 (2019). https://doi.org/10.1007/s13347-018-0330-6
61. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. J. Law Technol. **31**(2), 841–887 (2018)
62. van de Poel, I.: The relation between forward-looking and backward-looking responsibility. In: N. A. Vincent, I. van de Poel, J. van den Hoven (Eds) Moral responsibility: beyond free will and determinism (2011) (pp. 37–52). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_3
63. van de Poel, I., Royakkers, L.: Ethics, technology, and engineering: an introduction, 1st edn. Wiley-Blackwell, Malden (2011)
64. Audi, R.: The good in the right: a theory of intuition and intrinsic value. Princeton University Press, Princeton (2005)
65. Hayes, P.: An ethical intuitionist account of transparency of algorithms and its gradations. Bus Res. **13**, 849–874 (2020). https://doi.org/10.1007/s40685-020-00138-6
66. D'Ignazio, C., Klein, L.F.: Data Feminism. Cambridge. (2020) Retrieved from https://bookbook.pubpub.org/data-feminism
67. Pasquale, F.: The black box society: the secret algorithms that control money and information. Harvard University Press, Cambridge (2016)
68. de Laat, P.B.: Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos.Technol. **31**(4), 525–541 (2018). https://doi.org/10.1007/s13347-017-0293-z
69. Robbins, S.: AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI & Soc. **35**, 391–400 (2020). https://doi.org/10.1007/s00146-019-00891-1
70. Robbins, S.: A misdirected principle with a catch: explicability for AI. Mind. Mach. **29**(4), 495–514 (2019). https://doi.org/10.1007/s11023-019-09509-3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.