An ethical intuitionist account of transparency of algorithms and its gradations

Hayes, Paul

DOI
10.1007/s40685-020-00138-6

Publication date
2020

Document Version
Final published version

Published in
Business Research

Citation (APA)
Hayes, P. (2020). An ethical intuitionist account of transparency of algorithms and its gradations. *Business Research*, *13*(3), 849-874. https://doi.org/10.1007/s40685-020-00138-6

Business Research (2020) 13:849–874
https://doi.org/10.1007/s40685-020-00138-6

ORIGINAL RESEARCH

An ethical intuitionist account of transparency of algorithms and its gradations

Paul Hayes[1]

Received: 19 December 2019 / Accepted: 4 December 2020 / Published online: 23 December 2020

Abstract To make evaluations about the morally relevant impacts of algorithms, transparency is needed. This paper lays out discussion of algorithms and transparency in an explicitly moral analysis with a special focus on the domain of justice and security. The paper provides an account of the moral import of transparency, defined itself as an instrumental value denoting a state of affairs conducive to acquisition of knowledge about some X. A normative account of transparency is outlined relying on an intuitionist framework rooted in the works of Ross and Robert Audi. It will be argued that transparency can be derived as a subsidiary (prima facie) principle from other duties including beneficence and justice and that it is groundable in the value of knowledge. Building on this foundation, the paper examines transparency and duty conflict with a special focus on algorithms in justice and security, recognising that complete transparency can be impossible where duties conflict. It is argued that as a subsidiary (prima facie) principle, transparency is overridable but ineradicable, which is to say that sufficiently justifiable reasons for secrecy or opacity can licence limiting transparency, that is, there may be occasion where full transparency is not our final duty.

Keywords Values · Algorithms · Intuitionism · Transparency · Justice and security

1 Introduction

In an age empowered by information technology and the generation of endless data, Big Data is now being mined for patterns that can provide useful insights for action across multitudes of contexts, from commercial to governance. Algorithms occupy a

✉ Paul Hayes
hayesp.research@gmail.com

[1] Values, Technology and Innovation, TU Delft, Delft, The Netherlands

very important place in this age of ubiquitous information and information technologies, intended to provide these key insights. Algorithms are, however, not neutral and value free, but rather value laden and can be instilled with and reflect biases, and can help realise or undermine our values in significant ways (O'Neil 2016; Hayes et al. 2020).

Since algorithms are value laden, and can have significant consequences for individuals, it is essential that we can know them and understand their context to some extent, at least in a manner that allows us to respond to them in appropriate ways, individually or as a society. Transparency is necessary to come to this knowledge.

The most recent research domain of the present author is that of justice and security, and the ethics and transparency of algorithms therein. The potentially severe consequences of poorly designed or inappropriately implemented or deployed algorithms in this context, undoubtedly anathema to human flourishing, provide a particularly urgent and sobering case for transparency of algorithms. Whilst promising insights that can help us understand and respond to crime (and possibly even save lives), algorithms also threaten to compound negative contacts with agents of justice and security, and exacerbate issues of discrimination and racial inequality (Ferguson 2017b), where in the worst circumstances minorities like Black Americans already face disproportionate violence and victimisation at the hands of the police (Zimring 2017).

The work presented here, noting the urgency of transparency of both algorithms and the contexts in which they are embedded, primarily explores the normative dimension of transparency, what it is, why it is ethical, and how it may conflict with other values and duties, using the case of algorithms in justice and security to illustrate points and provide examples of application and clarification. This paper argues that transparency obligations can usefully be understood as a subsidiary principle of prima facie duty in the ethical intuitionist tradition, and as such may be overridable but not ineradicable in situations of sufficiently weighty countervailing duty claims (Audi 2005, p. 24). Furthermore, transparency of something may not be complete, but graded, where facets of the thing in question are not, at least widely, disclosed, because of countervailing duty claims.

The work presented here builds on a Kantian normative conception of transparency of it being a value endorsing openness and honesty, as required by a prohibition against lying to respect human beings, or more precisely, their "… rational agency and free will" (Plaisance 2007, p. 188). Notably, this paper also builds on the work of the European independent High-Level Expert Group on Artificial Intelligence's (HLEGAI) (2019) *Ethics Guidelines for Trustworthy AI*, which endorses transparency as an important ethical principle, though one which is not absolute. Similarly, it builds on the work of Turilli and Floridi (2009, p. 107) in exploring the wider issues of frictions in transparency (duty conflicts by my account, and dependence/regulation relations by theirs).

It builds on and moves forward with a Kantian conceptualisation by connecting it with and advancing a Kantian Intuitionist account of transparency inspired by Robert Audi. Using the ethical theory intuitionism, it also supplies the influential work of the HLEGAI with some independent theoretical validation. A somewhat parallel account of transparency to that of Turilli and Floridi (2009) is presented, though one which seeks to emphasise the import of transparency by further arguing that it is both a value and an ethical principle, with its basis found ultimately in a plurality of principles of prima facie duty and ontically grounded in the intrinsic value of knowledge.

This paper is structured as follows. Section 2 provides a brief descriptive account of transparency so that the reader may understand what kind of state of affairs I refer to when discussing duties relating to bringing about such a state of affairs.

Section 3 introduces some examples of algorithms in justice and security, thereby outlining the unique importance of transparency and knowing the capability and consequences of algorithms, and providing background information for points of illustration throughout the paper.

Section 4 introduces ethical intuitionism and a normative account of transparency as a subsidiary principle of prima facie duty, derivable from a plurality of Rossian principles of duty and ontically grounded in the intrinsic value of knowledge.

Section 5 further explores the nature of a subsidiary principle of prima facie duty of transparency, examining the questions of whom are the duty and rights holders, the limits of transparency, and potential conflicts. This utilises examples of algorithms in justice and security to illustrate points and examine applications.

Section 6 addresses whether or not the creation and deployment of intrinsically opaque artefacts (algorithms) is permissible under a subsidiary principle of duty of transparency.

Section 7, building on the limits and boundaries of transparency explored in the preceding sections, explores the morality of one proposed type of trammelled transparency, qualified transparency.

Finally, Sect. 8 briefly makes the necessary contribution of reaffirming the inherently moral and value-laden nature of transparency, rejecting the argument that it is neither a value nor a true ethical principle.

2 What is transparency?

In this paper, transparency in its descriptive sense (not normative, which follows) will be defined as a state of affairs conducive to the acquisition of knowledge about some X (for instance, an algorithm) characterised by availability, accessibility, findability and understandability/explainability of relevant information, a synthetic definition (Hayes et al. 2020).

Briefly, the properties of these characteristics can be elaborated as follows:

- Availability—information about X exists.
- Accessibility—the information is accessible to an agent seeking it.
- Findability—the information is appropriately catalogued and publicised such that it can be found by an agent seeking it.
- Understandability/explainability—agents must be able to grasp the available information, and be able to impart it to others.

- Relevant information—the information available about X is sufficient to answer questions about it (see Tu 2014, pp. 30–32). Transparency can have teleological and relational elements in that "…some X will be rendered transparent to some person(s) Y for purpose Z (whether that is auditing, or informed decision-making, etc.)" (Hayes et al. 2020, p. 15).

This synthetic conceptualisation of transparency arises from an acknowledgement of transparency's key features across the relevant literature, noting in particular that its uses normally entail the disclosure of information by a responsible agent that is relevant for decision making (Fleischmann and Wallace 2005; Heald 2006; Turilli and Floridi 2009, p. 106; Vaccaro and Madsen 2009; Etzioni 2010; Hulstijn and Burgemeestre 2014, Tu 2014, pp. 27–32).[1] In addition, this descriptive conceptualisation of transparency, emerging from a larger project of which this paper forms an integral piece, is intended to argue for a widening of the concept beyond mere openness in communication, acknowledging that openness and disclosure are insufficient for an effective account of transparency, given the risks of, for example, information glut.

If a concept of transparency is to be enduringly useful and remain relevant, particularly in an age of Big Data and incredibly complex machine-learning algorithms, it must move beyond visibility and re-centre towards knowledge as a goal. Something can be seen, and not understood, and therefore, there may be little value in it being transparent (see Ananny and Crawford 2018). There can be a preoccupation with the conditions of availability and accessibility in discussions of transparency, which are certainly necessary but largely insufficient conditions to support a state of affairs useful for empowering agents with information conducive to knowledge creation, and to which they can respond.

Here, it is proposed that something should be seen, and ideally understood. As such, transparency connotes the availability, accessibility, and findability of information as part of its visibility component (see Tu 2014, pp. 27–32), and understandability or explainability of this information as a part of the understandability component (see Tu 2014, pp. 27–32; Miller 2017; Mittelstadt et al. 2019). If something is visible, and understandable, it promotes knowledge, which is an important intrinsic moral value. The moral value of transparency is, therefore, tied instrumentally to its potential for leading to knowledge creation and dissemination.

When something is transparent, it can be known, which is good in itself, but also if it can be known, it can be acted upon or responded to in appropriate ways. Therefore, knowledge itself has instrumental features which are commonly necessary to promote the values traditionally associated with transparency, for instance, human dignity.

This conceptualisation of transparency acknowledges that, effectively, transparency of something is graded or exists on a spectrum from more (fully transparent) to less (completely opaque), as determined by the presence and depth of the listed characteristics of transparency. As the paper will explore, not all possible

[1] This particularly builds on the work of Yu-Cheng Tu (2014, pp. 27–32) in adopting a multi-component view of transparency incorporating accessibility, understandability, and relevance, though the precise definitions of each offered here differ somewhat.

agents will be entitled to a fully transparent state of affairs about some object. One of the goals here is to explain why this is so.

The work presented here supports and expands on the outputs emerging from XAI, the movement which explores and promotes methods for explaining artificial intelligence (which will be returned to in Sect. 6), which similarly demonstrates the importance of an understandability characteristic above and beyond a visibility characteristic in transparency, at least where meaningful decisions are made by opaque systems. Applied particularly to the context of algorithms, when the conditions ranging from availability to relevant information are fulfilled they can support the design and implementation of algorithms that are traceable and explainable, and by virtue of this enable individuals to come to know and respond to them (High-Level Expert Group on AI 2019, p. 18).

Where this conceptualisation of transparency is distinct from XAI is that it is ultimately intended for general use across contexts, and not only in reference to the operations of algorithms but (especially in the domain here) information pertaining to the environment in which the algorithm is embedded.

3 Transparency and algorithms in justice and security

Ethical challenges emerge from Big Data and, separately but often relatedly, the use of algorithms. Algorithms are value laden and are used in sensitive governance contexts such as the domain of justice and security, where they can have significant impacts on those who are subject to their analysis, or simply live in a geographic area that is subject to algorithmic analysis. This section will highlight some examples of algorithms in justice and security, outlining their risks and emphasising the importance and challenge of transparency in an age of algorithmic governance.

Such algorithms are not neutral and are often value laden, contrary to the rationale for using them in the first place (Hayes et al. 2020; Barocas and Selbst 2016; O’Neil 2016; Ferguson 2017b; Kitchin 2017). They are popularly critiqued based on ethical grounds and questions arise concerning the nature of the training data they use, and how such data, should it be based on discriminative policing and police practices, can help perpetrate negative or “pernicious” feedback loops where previous patterns of corrupt policing are reinforced by an algorithm’s output (O’Neil 2016, p. 87; Richardson, Schultz and Crawford 2019). These claims have been empirically tested. Lum and Isaac (2016) for example tested the PredPol algorithm (a geo-spatial risk or hotspot-based predictive algorithm) to test drug-related datasets of different origins (health data versus arrest records) and found that distribution of risk based on the latter set was concentrated in low-income areas. To act on this output, assuming it is biased, would result in unfair policing more likely to target disadvantaged groups.

In fact, algorithms can have very racially biased results as an investigation of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism predictive tool by ProPublica, also in 2016, found (Angwin 2016). COMPAS predicts a defendant’s likelihood of reoffending and is used in the judicial process in the US 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 (Angwin 2016). In fact, Black defendants were incorrectly predicted to reoffend at twice the rate as White defendants (44.9% versus 23.5%) (Angwin 2016). What is more, as highlighted by D'Ignazio and Klein (2020) and reported by Angwin (2016), the algorithm's input data were found to use variables that are a close proxy for race.

Algorithms also have overt implications for privacy, identity, and can mediate their subjects' contacts and experiences with the police. The Chicago Strategic Subjects List (SSL) is an algorithm deployed by the Chicago Police Department and ranks individuals by risk of involvement, whether as victim or perpetrator, in gun violence, assigning them a score of between 1 and 500 (Ferguson 2017b, p. 37). This raises questions about whether it is appropriate to attach such ambiguous and yet meaningful data to an individual, which appears to be at odds with their autonomy and moral agency. There are also questions of privacy in their deployment, with one of its operational uses being a custom visit by the police with members of the community and a social worker (Ferguson 2017b, p. 38).

An example of algorithmic software provided by Palantir to Police Departments in the US further demonstrates the potential erosion of privacy that can occur from algorithmic design and deployment. It is not only criminals who may be included in data collection practices, as Sarah Brayne (2017, p. 992) observes, police are increasingly using data on persons with no prior police contacts. Brayne (2017, pp. 992, 994) offers the example of network analysis offered by Palantir's platform, which has access to disparate data sources. The Palantir network analysis shows associational webs of entities relating to a person who has had prior police contact, including people and vehicles or phones (Brayne 2017, p. 992). Some of the individuals appearing in the associational network have not had prior contacts with the police and are included in a database simply by association, and may be colleagues or family (Brayne 2017, p. 992). Brayne (2017, p. 992) calls this a network of incidental persons a secondary surveillance network.

In all these cases, there are implications most pressingly for members of minority or disadvantaged groups who are likely to be disproportionately represented in police databases, which also form the foundation of the training data sets used (Ferguson 2017b, p. 73; Richardson, Schultz and Crawford 2019). There is also the risk of increased police contacts or undesirable encounters with the general justice and security system which could emerge from increased suspicion of individuals, and therefore, an increased risk of dangerous encounters with police—for instance in high-risk areas or between members of a police force and persons on the SSL (Ferguson 2017b, pp. 79, 85, 95).

Based on these challenges, it is clearly important for us to know many aspects of an algorithm's design, implementation, and deployment both collectively and as a society, and as individuals whose rights might be more immediately impacted. We need to see, and know, about these algorithms and the organisations and processes in which they are embedded to challenge them or uphold our rights where undue interference is threatened. Through knowledge, we can hold designers and end-users to account and challenge them to do better, to design and use algorithms in a way that serves the public good and supports our moral values. For this reason, we need transparency, ideally both of algorithms on a technical level, and of the environments in which they are embedded. Transparency serves the production of knowledge that can enable the verification of value or disvalue, and accountability (Hayes et al. 2020). In relation to risk ranking, for instance, Nicholas Diakopoulos (2015, p. 400) succinctly raises the exact kind of questions we expect to be able to answer about algorithms, "…is that risk being assigned fairly and with freedom from malice, abuse, or discrimination."

Yet, there are numerous challenges to the transparency we seek. Opacity dominates many algorithmic contexts, both by design and by necessity—they are so-called black boxes (Diakopoulos 2015, p. 404). Such opacity can be intentional, where intellectual property holders prevent information from being widely disseminated to protect that intellectual property from duplication (Mittelstadt et al. 2016, p. 6). From the examples here, the algorithms are usually held as private property and exploited for the financial gain of their IP holders. Some of them elect to be rather open (like PredPol), and others prefer to maintain more secrecy (equivante's COMPAS, ProPublica's investigation notwithstanding). In addition, should precise design decisions and the inner-workings of algorithms become exposed, their subjects may be able to form evasive strategies (Ferguson 2017a, p. 1187), which could undermine their efficacy.

Another problem, arising with algorithms and Big Data more broadly, is their general level of understandability and that they may be incomprehensible both to the general public and experts alike, particularly in the case of machine-learning algorithms (Burrell 2016, p. 2; Mittelstadt et al. 2016, p. 6; Lepri et al. 2018, pp. 619–620).

Therefore, algorithm designers and users have (at least prima facie) good reasons for not wanting to disclose certain information—to protect their property interests and to not undermine the operational impact of the algorithm. In addition, should it be that these parties are open to disclosure, the dissemination of relevant information may not even be particularly useful if its audience cannot comprehend it.

Transparency is important, and not without challenges, as there are different facts and countervailing considerations to consider. The remainder of this paper will unpack a theory of a principle of duty of transparency to help clarify how relevant stakeholders can deliberate on what information should be made known, and to whom, and whether or not a duty of transparency prohibits the creation of obscure artefacts like black box algorithms. Algorithms in justice and security will provide an application case for the unpacked theory.

4 Ethical intuitionism and transparency as a subsidiary principle of duty

In the preceding, it was argued that transparency is a state of affairs that is conducive to the acquisition of knowledge. The remainder of this paper will examine more closely the normative dimension of transparency as supplemented by moral theory, arguing that it should be understood as a subsidiary principle of duty.

A pluralist deontological approach, ethical intuitionism, is adopted in this examination of transparency and what it might entail normatively. Before justifying transparency as an ethical principle, some explanatory work about ethical intuitionism will first need to be undertaken.

Ethical intuitionism proposes that there is a plurality of intuitively (non-inferentially) knowable, self-evident, basic moral principles of prima facie duty.[2] An early list of these principles, proposed most notably by Ross (2003), an early (to late) twentieth Century British philosopher and amongst the first major proponents of intuitionism, were, broadly; fidelity, reparation, gratitude, beneficence, self-improvement, justice, and non-maleficence.

Robert Audi (2005) applied his understanding of Kant's categorical imperative (specifically the second formulation of the categorical imperative—that persons should be treated as ends in themselves and not merely as means), to systematise intuitionism's principles, and to elaborate an enhanced version of Ross' list. Under Audi's (2005) argument, this formula provides a grounding for Rossian duties. He begins his development of Kantian intuitionism by stating that Rossian duties can be argued to be derivable from Kant's categorical imperative, that they can be deduced "…from careful application of the categorical imperative to everyday life" (Audi 2005, p. 102). The Rossian duties, so understood, tend towards respecting human dignity (a concept inclusive of "…at least rationality, the capacity for normative judgement and moral agency, a kind of sentience, and other values warranting respect for persons" (Audi 2005, p. 157)); they broadly prohibit the treatment of persons merely as means and promote respecting them as ends (Audi 2005, p. 106). By Audi's (2005) account, Rossian duties represent articulated standards in the endorsement of Kant's humanity formula. Audi (2005, pp. 105–106) proceeds to argue that Rossian principles of duty are constraints on the imperative's use—"…interpretations of it that do not yield them, and applications of it that are inconsistent with them, are prima facie defeated by that fact". Kant's categorical imperative can provide an explanatory account of Rossian principles (Audi 2005, p.110), using human dignity as the basis underlying the categorical imperative, and which forms the basis of Rossian duties.

This relationship is not parasitic of one by the other, as it might seem. Audi (2005, pp. 109–112) explains that they stand in "mutually clarifying" relations. A thorough account is provided of their relationship and how they mutually support

[2] There are some things to note here, self-evidence and intuition do not preclude justification (Audi, 2005). Although the theory is not dependent on justification, intuitions are not indefeasible and are not impervious to being proven wrong (Audi, 2005). They are also knowable, but not necessarily known, they arise from an appropriate understanding of propositions (Audi, 2005). As argued by Audi (pp. 39–40), on intuitions:

> The intuitionist thesis that some knowledge of what we ought to do is intuitive and non-inferential implies neither that it is not reflective nor that it cannot be supported by argument or refuted by relevant considerations to the contrary.

And on self-evidence, as per Ross (Audi, 2005, p. 43):

> His view does not imply that ordinary moral agents know or would accept self-evidence of its principles, nor even that moral theorists can know their self-evidence *non-inferentially*. It is first-order moral propositions such as the principle that there is a prima-facie duty to keep promises, and not second-order thesis that such principles are self-evident, which are the fundamental kind of thing we must know intuitively if a Rossian intuitionism is to succeed.

each other, though the salient points as argued by Audi (2005, pp. 109, 110, 112) are thus:

> …[o]ur justification for accepting the categorical imperative can be enhanced by our justification for accepting the principles of duty… and our justification for accepting them may be enhanced by our awareness of the support they receive from "above"—from the imperative—as well as by our awareness of their being intuitively confirmed from "below"—in application to concrete moral cases about which we have clear convictions… we have a plurality of moral obligations expressible in Rossian principles of prima facie duty, and although these are non-inferentially and intuitively justifiable, they are systematizable by, and stand in mutually clarifying relation to, the categorical imperative.

Audi (2005, pp. 188–195) augmentation of, or upgrading of the theory to Kantian intuitionism, resulted in an enhanced list of principles of duty more rooted in the notion of respect for persons, and concerned with avoiding treating persons merely as means, but as ends:

1. **Prohibition of injury and harm**: This prohibits the multitude of harms and injuries that can be visited upon persons, whether that manifests as physiological, psychological, social or in another way (Audi also considers deprivation of freedom).
2. **Veracity**: Simply put, this is the obligation not to lie.
3. **Promissory fidelity**: This is the obligation to fulfil one's promises.
4. **Justice:** An injustice would tend towards being active deprivations of values such as liberty or pleasure. This duty entails treating persons in a manner that does not deprive them, as well as actively working towards rectifying and preventing injustice. I would also argue this should subsume fairness, the fair treatment of persons or allocation of goods and opportunities in the Rawlsian sense, which Audi (2005, p. 173) also does examine briefly by way of middle theorems (explained below).
5. **Reparation:** This is the obligation to make amends for a wrong doing.
6. **Beneficence:** This is the duty to contribute to the flourishing or well-being of persons, and effectively the values that are present in their lives.
7. **Gratitude:** This is the duty to effectively (and proportionally) reward good deeds done to us through word or deed, as appropriate.
8. **Self-improvement:** This is a duty towards our own flourishing, to build our capabilities, including moral and intellectual capacities, that is, intellectual, social, aesthetic, and virtues.
9. **Enhancement and preservation of freedom:** This duty requires, primarily, the removal of restraints to freedom, and secondly enhancing opportunities.
10. **Respectfulness:** This is the duty to *treat* people respectfully.

Such principles of duty are prima facie as they may and often do come into conflict (or at least, as Audi (2005, 2015) and Stratton-Lake (2003, p. xxxviii) make clear, reasons may come into conflict). A surgeon who is presented with conjoined twins, whom can only save one in an operation that will surely leave the other dead, will face competing reasons towards duties of beneficence and the prohibition of harm and injury, and the same is true of the classical trolley case. There may also be competing obligations towards one duty, such as where one makes conflicting promises (to say, attend two concurrent birthday celebrations). The final duty is that which corresponds to the action one must take. An overridden prima facie duty may trigger another duty, such as reparation (Ross 2003, p. 28).

Audi's Kantian intuitionism is value based: as evident from the preceding, the value of human dignity in particular plays a central role to Audi principles of prima facie duty, stemming from his deployment of the categorical imperative. Values (intrinsic values) are, he acknowledges, what make life choiceworthy (Audi 2015, p. 53)—they are the good that Ross' theory strove for (in his case, knowledge, virtue, and pleasure (Ross 2003, pp. 23, 140)). Values are the good that are (ideally, but not always) brought about by right acts. Ross (2003, p. 155) argued that an action was good where it sprang from a certain motive. Intrinsic values can provide reason for action (Audi 2015, p. 53), and "…action in accord with… principle is at least a partial realization of that value" (Audi 2005, p. 141).

Audi (2015, p. 100) argues that the basic bearers of intrinsic value are internally experiential, so goods such as Ross' three main intrinsic values like knowledge, virtue, and pleasure. Returning finally to transparency, it is not experiential, but is a state of affairs with direct connection to intrinsic values (with knowledge being the first focus here) and is one also of a distinct utility. Transparency is not so much experiential in the sense of intrinsic values, but its presence can help bring them about. It is an instrumental value—it is not good for its own sake, but for bringing about the intrinsically good. We do not experience transparency in the sense that we experience the knowledge that we hope emerges from it. There is indeed a goodness in this, in bringing about an intrinsic value (primarily when done from the right motivation, which distinguishes a good act from simply the good), therefore, it remains morally important. Transparency can, as Audi might say, contribute to a choiceworthy life (if rather indirectly).

Just as, in a sense, Audi's principles of prima facie duty are derivable from the categorical imperative, we can derive further principles from these duties that are also ontically groundable by values (prominently, knowledge) (see Audi 2015, p. 141). Such principles, subsidiary to principles of prima facie duty, are more specific and applicable (codes of conduct in professional ethics are an example) and are called middle theorems (Audi 2005, pp. 165–171). In our case, a middle theorem of transparency can be derived from principles of beneficence, self-improvement, veracity, justice, and arguably fidelity. Transparency as a normative rule is derivable from these, in combination as:

- Creating and sharing knowledge is an act of beneficence, and knowledge is often required for self-improvement.
- True information is a requirement of transparency as described (veracity) (see Floridi 2012).
- The deprivation of knowledge would be unjust.

- We expect openness and honesty in governance; it is arguably an implicit promise (fidelity) that comes with the assumption of certain kinds of power.[3]

Transparency as a subsidiary principle is a finer and more applicable refinement of these principles of prima facie duty. It is ontically groundable as it serves other values, most pertinently knowledge. As a subsidiary principle, it is also overridable but ineradicable.

To articulate transparency as a principle more clearly, we might say that it is:

> the duty to render an object or entity and related objects or entities (as necessary) knowable.

The implication of this is that a relevant agent (the developers of an algorithm, a journalist investigating an algorithm, a law enforcement agency using one), with a sufficiently relevant relation to an object and related objects in question (an algorithm, its uses by an organisation, etc.) should create and process and make accessible information about that object for a relevant audience to which they also bear some existing or prospective relation (see Kaspar 2012, pp. 101–104).

The descriptive account of transparency outlined in the preceding further gives shape to the principle of transparency. Information should be made available, findable, accessible, understandable and should be relevant to answering possible queries.

The next step is to further elaborate on transparency as a subsidiary prima facie principle of duty and illustrate its application with the help of some examples of algorithms in justice and security, a motivating concern for this paper.

5 Transparency as a subsidiary principle of prima facie duty

Transparency as a principle, by default, obliges those in the position to do so (depending on their relation to an object and related objects of concern) towards production of relevant information and dissemination of such information to individuals with whom they have a relation, and in appropriately understandable forms as required by those audiences. For if available information cannot be consumed in its raw form (or even found, for that matter), it is not sufficient to produce knowledge (see Floridi 2012). As a broad principle, transparency illuminates the simple good in acts of information and knowledge production. By virtue of this good, transparency implies duties. Transparency motivates (or gives positive reasons for) broad disclosure of information of an object with which a duty bearer has some significant relation, and its responsible presentation so that it is sufficiently accessible, findable, understandable and relevant.

So far, a broad sketch has been drawn of this principle, leaving several important questions in need of further elaboration, including pressingly:

[3] For more in-depth discussion of power in a state context, see (Fox-Decent, 2011).

- Who exactly are the duty bearers of transparency?
- Who holds transparency rights and what kinds of information might they expect to receive?
- What about the limits of transparency and what about potential duty or value conflicts (for example, transparency and privacy?)
- How might such conflicts be resolved?

Each of these questions will require varying degrees of elaboration, and will thus be dealt with in the following four sections.

5.1 Transparency duty bearers

The duty bearers of transparency are those responsible for some object or task by virtue of their relation to it (so, their role) and other agents (rights holders). Duty bearers may be individual, or collective (see the work of Miller 2009). A corporate entity is responsible for its intellectual property and as a result holds collective prima facie duties of transparency about this object, as well as being transparent about its own constitution as it is itself an important object of analysis.[4] All things being equal, for example, we might argue that PredPol has a prima facie duty to disclose information in relation to their predictive risk algorithm. Indeed, PredPol has been an ideal example of an agent discharging this duty, being one of few such organisations to have released its algorithm for peer review (Lum and Isaac 2016, p. 17).

At this point, it is important to also note that transparency of the object in question is not the sole responsibility of its creator. The object does not exist in a vacuum, and there are others who hold a relation to the object and to other agents who will additionally hold duties of transparency. The police who use the algorithm, all things being equal, would be expected to produce information regarding their use of it, and its results (for example, changes in crime figures following its adoption) for the general public.

Finally, there may be a third party that holds a potentially more oblique duty of transparency here. Some objects, as argued, are value laden, and though they may not bear a direct relation to certain agents, they attach themselves to them by virtue of the overall goods that the agent is responsible for (for instance, accountability). Therefore, as a value-laden object of public interest, as it impacts public values in serious ways, a prima facie duty of transparency is held by (for one example) journalistic institutions, the core responsibility of which is to shed light on matters of public interest (see Diakopoulos 2015). A useful example here is ProPublica, who conducted the earlier referenced rigorous investigation of the COMPAS algorithm and unearthed significant racially charged false positives in its results (Angwin et al. 2016; D'Ignazio and Klein 2020).

[4] D'Ignazio and Klein (2020) make an important point about self-disclosure. Those who work on (and perhaps with) algorithms should, by their argument, self-disclose their positions and acknowledge their own privileges and not only that, but seek and invite the standpoints of others, those whose voices are probably not adequately represented in the design process and who are most likely to suffer the adverse consequences of an algorithm's deployment.

There are two types of actions that duty bearers will undertake in the course of their duties which are inherently ethical; proactive and reactive information production and management. Duty bearers in organisational contexts are expected to record and produce (and catalogue and publicise) information about their activities. This can be regarded as a proactive duty towards transparency. Duty bearers will also hold reactive duties, that is, responding to queries from transparency rights holders and producing information that can answer their questions.

5.2 Transparency rights holders

In terms of what should be transparent to whom, a prima facie duty of transparency gives positive reasons for duty bearers to render all objects they hold a relevant relation to (and sometimes themselves) transparent to individuals with whom they have a relevant relation. Knowledge is an intrinsic value, and a public good, which all things being equal, means that acts leading to such are *right*. This is obviously not always the case; nevertheless, moral reasons motivate knowledge-promoting actions.

The general public can claim (overridable) rights to information pertaining to objects for which duty bearers are responsible. The general public will have a right to different categories of information relating to the design and deployment of hotspot algorithms such as PredPol. Local police departments should make this information available (their proactive duty) and respond openly and candidly to related queries to the extent that no other duty conflicts are entailed (for example relating to intellectual property or privacy) as is their reactive duty.

Ultimately, the strength of claims to transparency and access to certain types of information are dependent on the precise nature of the relationship between the duty bearer and rights holder, and might be moderated by, for instance, the power relations between the two (those subject to algorithmic decisions by coercive authorities such as the police will have a strong claim to transparency). There are additional concerns other than knowledge that may give a claim to certain information additional weight, for example, where it is necessary to prevent some harm or for redress (see Robbins 2019a).

The untrammelled disclosure of information from duty bearer to rights holder is not always necessary. In fact, there are certain types of information which should probably always be made known (the existence of an algorithm and its role in some processes or outcomes), and other types of information which may not always be necessary for at least all potential rights holders (a proprietary model or specific outputs such as individuals' personal information). Value or duty conflicts give justification for curtailing transparency duties. This will be explored more in the following sections. That is to say, the obligation to share particular kinds of information with particular agents can be defeated.

5.3 Transparency limits and duty/value conflict

So far, a duty of transparency threatens to be superogatory or unachievable given the manpower and effort entailed by activities of information production, recording, and meaning-making, as well as from the inherent conflict that arises between transparency and other values. The duty, in practice, demands positive efforts and a reasonable commitment to actions within what it is in one's power to achieve. Transparency is clearly resource intensive, requiring substantial effort and manpower to create conditions conducive to knowledge for audiences of varying epistemic capability. If pressed on what information to make transparent, given limited resources, a duty-bearing agent will have to prioritise the most objectively important information, that is, information necessary for justice or to avoid harm (see for example Robbins 2019a, b).

Our duties, bound as they are to values, may conflict, as acknowledged in the HLEGAI's Guidelines for Trustworthy AI (2019, p. 13). Duties of transparency can conflict directly with the value, for instance, of privacy. This is similarly illustrated by Turilli and Floridi (2009, p. 107) who argue that although certain ethical principles are dependent on information transparency (accountability, safety, welfare, and informed consent) others give cause to regulate the extent or depth of information transparency (privacy, anonymity, freedom of expression, and copyright).

If the Chicago Police Department were to lay bare every facet of the Strategic Subject List without discretion, it might entail revealing the identities of those on the list, at great cost to their privacy and probably their overall welfare (one can imagine the difficulties of finding a job given the public allegation of being a potential victim or perpetrator of gun violence). Therefore, in this example, we can see more specifically a conflict between a principle of transparency and the prohibition of harm and injury, as the revelation of the identities of persons on the SSL would come at no small expense to their privacy and likely their overall well-being.

In another example, transparency can conflict with intellectual property. If we imagine equivante's COMPAS algorithm being exposed in granular detail, to the point that it was reproducible by competing organisations, the value of property would be undermined, and again, there may be a conflict between, more granularly, transparency duties and the prohibition of harm and injury.

So the duty of transparency, insomuch as it relates to making an object knowable, is not always a zero-sum do or do not affair, and there are multiple considerations against what is to be made transparent, and what harm or benefits might arise, and even the plausibility of the satisfactory commitment to transparency given the potential monetary and time investment implied. Therefore, transparency gives rise to positive reasons for action, but other considerations, relevant to other values and duties, give negative reasons for action. As argued by Patrick Lee Plaisance (2007, p. 192), "…insisting on a transparency unmediated by other values can certainly become destructive and self-defeating".

The duty of transparency is not a blanket and binding requirement to bring about a state of affairs of complete knowledge to everyone about all things for which one has sufficient relation. There are considerations which defeat rendering aspects of things known, at least to all. Practically and at a zoomed in level, transparency duties can be conceived as relating to decisions about types and categories of information to be made available that contribute towards states of knowledge about something. Over the course of making some object knowable, elements of information that contribute to the totality of knowledge about something, may, for sufficiently justifiable reasons, remain opaque. As such, it is not always a duty bearer's final duty to disclose every possible facet of information relating to X to particular agents.

5.4 Resolving duty/value conflicts

Ross' (Audi 2005, p. 160) preference for solving issues of conflict was practical wisdom, a concept most often associated with virtue ethics but nevertheless accepted in intuitionism, a theory that sometimes holds that virtues themselves are an intrinsic value and an important element of human moral development. Practical wisdom, a notably Aristotelian (2004) concept, conveys a sense of experiential knowledge of what is good and right. Practical wisdom is learned from experience (not automatically endowed), implies the ability to recognise the most morally salient features of a situation and recognition of the most appropriate action to take in light of those features (Hursthouse and Pettigrove 2018; Vallor 2018).[5] In the context of transparency, this implies purposive and deliberative reflection on the facts of a scenario, and a decision made that respects values and duties at stake, and finally recognition of one's final duty.

For our purposes here, practical wisdom by itself may not provide sufficient guidance on the resolution of duty conflict, as it presupposes sufficient moral knowledge and experience. Fortunately, Audi's Kantian development of intuitionism provides some additional conceptual resources to contribute towards decision making in moral dilemmas.

Audi (2015, pp. 172–177) provides eight weighting principles to help resolve conflicting prima facie duties. Not all will be relevant or explored in detail here, but they are nevertheless listed as:

1. Treatment of persons.
2. Moral diversity.
3. Distributive scope.
4. Equality.
5. Priority of the worse off.
6. Reducing alienation.
7. Coordination values.
8. The principle of secular rationale.

[5] For a rather extensive treatment of both ethical pluralism and practical wisdom, see (Ess, 2020).

Here, I want to focus on the weighting principle of treatment of persons,[6] which is specifically formulated by Audi (2015, p. 177) as:

> If two options we have are equally well supported by conflicting Rossian obligations, then if one option is favoured in terms of our (a) avoiding treating persons merely as a means or (b) treating persons as ends (or both), then that option is preferable, other things equal, with (a) having priority (other things equal) over (b) if (a) supports one option and (b) the other.

My view is that this principle, and the Humanity Formula broadly, especially reflect the particular importance of duties of non-harm and beneficence. To properly treat an agent as an end in themselves not merely as a means naturally entails respecting their agency and humanity, and refraining from acts that threaten their welfare and enjoyment of a choiceworthy life, that is, acts that might harm them. An important negative injunction is entailed. The importance of beneficence is also especially reflected, as the inherent good in supporting the agency of individuals is made obvious in recognising human beings as ends in themselves.

When the choice in the context of this research is between being secretive about some piece of information or being transparent, an excellent first step in reaching a decision might be in discerning which of the two options avoids treating persons merely as a means, and properly treats them as ends in themselves. I believe that, broadly speaking, the decision made should be one which does the least harm to individuals (especially relevant to avoiding treating persons merely as means), and best contributes to a choiceworthy life.

Take for one example, a scenario where a journalist or citizen asks a local police department about the vulnerabilities of its automated number plate recognition system (ANPR), that is, what kind of factors short of number plate removal from the vehicle might cause the system to fail to identify a vehicle's information in a database. Whilst the police hold a duty of transparency and it would normally be ideal to contribute to a state of knowledge about their algorithm, they can fairly argue that not sharing this information is necessary for maintaining the efficacy of a system that could potentially, say, identify fleeing criminals and could otherwise be used for enforcing the law (traffic fines). Denying this information probably does not cause harm to the question asker, they are still fairly treated by an account of the Humanity Formula, and ultimately the secrecy is to preserve the efficacy of a system potentially contributing to duties of beneficence and values of security in enforcement of the law. In the situation where the hypothetical system is deployed ethically, the denied request for information may properly treat persons as an end, as the purpose of the secrecy is to maintain the efficacy of a system contributing toward a choiceworthy life for human beings.

In another example, a subject of a recidivism risk assessment algorithm might wish to know the inputs considered in the calculation of the risk score and ask for this from the developer of the system. This claim for transparency is stronger, as

[6] The others are equally valid, however space is of course limited, and I believe that this one in particular carries a lot of explanatory power that can provide an initial set of compelling justifications in resolving duty conflicts.

they may need this information to challenge the algorithm's assessment and an inability to do this may play a causal role in an unfair prison sentence. Therefore, in the previous example secrecy was justified through non-harm via the principle of treatment of persons. Here transparency is justified in the same vein—this information is required to fairly treat the person as an end and in so doing to help them avoid coming to harm. The transparency here is not good only because it contributes to knowledge, but arguably knowledge essential to uphold human dignity.

The treatment of persons' principle then may help yield clarity on final duty, the most appropriate action to take in light of all relevant considerations. This theoretical treatment also lends support to the importance of non-harm in decisions of explanation, as described in some length by Scott Robbins (2019a) and which I will soon revisit.

As an important note to add to the forgoing, elsewhere with a colleague I have argued in favour of the integration of the conceptual resources of an ethic of care into decisions between parties of asymmetrical power relations, and here I maintain the importance of seeking the (situated) knowledge of those who may usually be overlooked in design decisions and suffer negative consequences from this (Hayes and Jackson 2020, citing Haraway 1988; D'Ignazio and Klein 2020). The specific learnings from care ethics are well accommodated within a careful intuitionist framework emphasising the importance of justice, beneficence, and non-harm, and one which can additionally (albeit not explored here) perhaps be enhanced with due consideration for the affective (Roeser 2011, 2017). Partiality (see for example, Randall 2020) and emotion need not be enemies of ethics.

To understand what actions may directly or indirectly cause or be conducive to harm to others, one must engage in dialogue with stakeholders, especially society's most vulnerable, to understand risks of harm, and what actions and consequences, for that matter, may be experienced as harmful. Not only should systems of power and their constituent parts be transparent to some degree to subjects of this power, but these subjects and their needs should be visible and known to those exercising power.

6 On the duty to make transparent algorithms and the role of explanation

If there is a subsidiary prima facie duty to render something knowable, it would seem important to address whether exactly, or to what extent, this duty may extend to creating an *artefact that can be known*.

Recall that some algorithms [artificial intelligence in the form of machine learning (ML) and deep Learning (DL) etc.] are intrinsically opaque—they are dynamic and utilise large amounts of data in ways that can be difficult or impossible to track. Some models are simply not interpretable; their operations are not understandable, or explainable, to human beings (Gilpin et al. 2018). As such, some algorithms are not (at least easily) explicable—they challenge the ability of human agents like designers and end-users to provide complete and accurate accounts of their operations, and much more so to non-expert audiences (see Adadi and Berrada 2018; Gilpin et al. 2018). Given this problem, it may reasonably be asked whether or not opaque algorithms should even be created or used when they so greatly challenge the possibility of transparency.

This is an issue that has given birth to movements such as XAI (explainable AI) which:

> …proposes creating a suite of ML techniques that 1) produce more explainable models while maintaining a high level of learning performance (e.g., prediction accuracy), and 2) enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners (Barredo Arrieta et al. 2020, p. 83).

A major goal of XAI is to provide explanations for decisions made by algorithms, to justify those decisions, control AI,[7] improve the models, and to discover new knowledge (Adadi and Berrada 2018, pp. 2, 5–6).

Such activities are inherently right, serving the advancement of human knowledge and understanding of algorithms, and ought to be considered the duty of those designing algorithms, as well as those sufficiently placed to inspect and evaluate them.[8] Certainly, a reasonable commitment should be made to advancing knowledge and understanding of the algorithms for which one is responsible based on a duty of transparency. That duty grows yet stronger in other cases, where the good is not located simply in the contribution to knowledge, but the contribution to knowledge essential for appropriate treatment of persons (informing someone of reasons for an algorithmic decision so that they may respond appropriately, and whereby otherwise harm might arise).

Of course, problems remain; such endeavours may stop short of producing explanations adequately understandable by end-users or algorithmic subjects. More problematic is that explicability might preclude the efficacy of the algorithm itself (Adadi and Berrada 2018, p. 6; Robbins 2019a).

I tend to agree with Scott Robbins' (2019a) assessment of explicable AI, and believe that an appropriate intuitionist analysis supports and validates the same conclusion, that opaque AI should not be used in situations where there is a risk of harm nor similarly created with intended application for situations where they are conducive to harm.

Explicable AI is an upper bound for transparency, and one with a very particular referent (AI) and a high degree of knowledge about that referent and only an absolute requirement where there is a definite risk of harm. AI may still be transparent to some degree where it is not explicable (information about its uses and organisational context might be known), but its reasoning cannot be understood and explained to justify potentially weighty decisions that may result in death or prolonged incarceration, for example (Robbins 2019a, p. 500). The issue is that where an AI is deployed in a context where it risks harm, and its decision cannot be

[7] See also Meaningful Human Control (Santoni de Sio and van den Hoven, 2018; Mecacci and Santoni de Sio, 2019).

[8] See (Diakopoulos, 2015).

explained or justified and as a result responded to, the subject of its assessment and an end-users' action is being acted upon arbitrarily, and the denial of explanation does not properly treat them as persons.

For this reason, AI should for example not be designed with intended application in judicial settings (recidivism risk assessment), or in settings where it may greatly increase the risk of violence by police authorities against individual algorithmic subjects (e.g., risk-of-violence scoring, at least where police widely have access to lethal weapons), unless reasons can be produced that justify its decision and enable those algorithmic subjects to respond to them.

Therefore, more generally, whilst transparency demands positive efforts and a reasonable commitment to actions within what it is in one's power to achieve, if the domain of interest is one entailing a significant risk of harm, and if the transparency pertains to something that can perpetuate harm, but no reasonable commitment can shed sufficient light on the object, then those responsible for it should not create or use that object.

Other AI applications may be less sensitive and harmful, and thereby not warrant this upper bound of transparency (explicable AI), but nevertheless, still warrant some degree of transparency. Model-centric explanations should minimally be considered by relevant duty holders for maximum disclosure, where "[t]he information that [is] provided relates to the data the algorithm was trained on, how the algorithm was tested for bias, the intentions of the designers, performance metrics, etc." (Robbins 2019a, p. 506). As Robbins (2019a, b) argues, such information is necessary for society to determine for itself whether or not the risks of the algorithm are acceptable. The acts this entails are right and transparent acts, with a basis in the value of knowledge, and through informing individuals with model-centric explanations they are properly treated as persons and empowered to respond to and challenge authority—they are not simply subjects of power nor merely used as a means to it. They are treated as agents and are not merely acted upon.[9] This capacity to challenge authority, and the algorithmic systems it deploys, is essential where the nature of datafication of individuals itself poses unresolved ethical questions about identity and selfhood (see Fitzpatrick 2020), with important implications for human dignity.

7 Qualified transparency

Whilst in particular cases the duty of transparency towards particular individuals (for instance, members of the general public) may be overridden in light of relevant considerations, as explored in the preceding, this may create a new opportunity even where it forecloses another. Where designers or end-users of an algorithm cannot impart comprehensive information about their algorithm to the public, and there remains a need for accountability, this can motivate the establishment of new entities and relations. In this case, designers and/or end-users would have a duty

[9] For more on a Kantian explanation of the relationship between sovereign and subject, see again (Fox-Decent, 2011).

(and one that could not easily be overridden) to disclose their algorithm (and information pertaining to its use) to an expert agency. Here, the algorithm is transparent in a relevant way, albeit not entirely to the general public.

Such an idea is not new, and has in the past been variously described as indirect transparency or bureaucratic transparency, where information is shared only with experts (Hood 2007, p. 194). More recently it has come to the fore again in the works of Pasquale (2010, 2016), who defends the practice of qualified transparency (largely in relation to internet service providers, but no less conceptually relevant here). Secrecy is not inherently bad, despite a "moral attachment" to disclosure (Birchall 2011, p. 64), there can be legitimate need for it (Plaisance 2007, p. 188; Pasquale 2010, p. 110).

Yet we still need knowledge, even if it is just in the form of "assurance" that we are not being subject to unethical, unjust, forms of surveillance or secret judgements that affect our interactions with public agents (Robbins and Henschke 2017) (or that the state will not become authoritarian, as argued by Robbins and Henschke (2017, p. 584)).

Pasquale (2010, 2016, pp. 161–171) argues that the intellectual property rights of service providers (in the specific context of his work) can be protected as well as the rights of the general public where disclosures are made to an entity that would, essentially, examine and evaluate algorithms, and consider complaints, whilst limiting disclosure themselves of information that would be harmful to the service provider or public. Of such entities, Pasquale argues (2016, p. 165):

> Agencies ought to be able to "look under the hood" of highly advanced technologies like algorithms at the heart of the Google search engine and the data they process. This might involve hiring computer scientists, programmers and other experts capable of understanding exactly how algorithms changed over time, and how directives from top management might influence what is always portrayed as a scientific, technical, and neutral process.

Such independent third parties would be subject to the prima facie duties of transparency too. Pasquale (2016, pp. 164–165) points out the necessity of third party entities providing information about their own methodologies for assessment and judgement of algorithms. Such entities must also be sufficiently powerful so as to be able to hold the subjects of their scrutiny accountable (Pasquale 2016 regularly uses the example of the Federal Trade Commission), particularly when the general public may not be in possession of sufficient information to do this. Qualified transparency without the possibility of accountability could be an empty exercise (Ananny and Crawford 2018; Kemper and Kolkman 2018).

O'Neill (2004, p. 272) for one endorses this manner of transparency (which she refers to as accountability in this case) as superior to more general forms of transparency in the institutional context:

> Professional and institutional performance can be assessed by those who are both sufficiently informed to judge what they assess and sufficiently independent to judge it objectively. Those assessments can be communicated

to the wider community by those sufficiently literate to communicate intelligibly with relevant audiences.

The qualified transparency solution may be a fitting one, in at least some cases, for the domain of justice and security. Algorithm developers and end-users have compelling reasons for wanting to preserve some degree of secrecy which, as was earlier shown, protects the integrity of the algorithm, their own capacity to benefit monetarily from it in the case of the designers/creators, and the privacy of its subjects. The relevant agents may be justified in preparing and sharing relatively little information with the general public, but could still be open to thorough scrutiny by a third party that could carry out investigations of the algorithm and its design, implementation, and deployment, and report its results to the public. In this case rather limited knowledge is shared with the public, and yet it is still knowledge, and most importantly it is knowledge about something to which the public can respond.

Qualified transparency represents an acceptable discharge of duties that can consider the factual circumstances of a case. It still results in knowledge but recognises that in some cases complete knowledge, by all possible agents, can be harmful. It is a manifestation of final duty, the correct and obligatory acts undertaken in light of all considerations. Appropriate institutions are required for qualified transparency, and so it makes societies responsible for the establishment of such institutions where they are not present (Miller 2009).

8 Transparency: pro-ethical or ethical, value or not?

As the preceding has been concerned with establishing a normative account of transparency (applied here in the domain of algorithms in justice and security) that emphasises its ethical nature, it would be remiss to close without acknowledging and addressing claims that transparency is neither a value in itself nor an ethical principle.

It has been argued that transparency is not an ethical principle, but a pro-ethical principle as disclosed information may be ethically neutral or "…ethical principles can be impaired if false details (misinformation) or inadequate or excessive amounts of information are disclosed" (Turilli and Floridi 2009, p. 106), and further the view has been adopted that as its value is instrumental, it is not a value in itself (Robbins and Henschke 2017, p. 585). O'Neill (2002) holds a similar view, warning that an increase in transparency may result in excessive and unsorted information; she is largely sceptical of its ethical value in practice. I argue that it is a mistake to view transparency as a pro-ethical principle and not simply an ethical one, nor a value, and its classification as pro-ethical is somewhat curious.

Transparency was described from two perspectives here, first as a state of affairs, and second as a prima facie principle of duty to realise that state of affairs as it provides the foundation for other values, and that it is derivable from and subsidiary to other principles of duty.

Transparency as a state of affairs is valuable, but not for its own sake. I do not consider it experiential, nor consequently intrinsically valuable. As a state of affairs that supports other instrumental (accountability for example) and intrinsic (knowledge) values, transparency can at times be indispensable. It is true that specific instantiations of transparency are neither good nor desirable, such as where irrelevant personal information about some non-public figure is leaked. It can stand in conflict with other values (privacy, in this example). Often when disvalue arises in particular instantiations of a state of affairs, it is because that state of affairs conflicts with another good or desirable state of affairs, as in the privacy example. Since values can conflict does not preclude them from being values. Liberty, for example, might uncontroversially be considered a value, and yet, it can instantiate in undesirable ways—a person can use unconstrained liberty to commit any manner of heinous acts. This calls for the careful management of states of affairs with appropriate (good or right) acts, a central point of ethics, and does not disprove value, merely that there is a plurality of them and that they can conflict or come into tension.

In that sense, to say that transparency is a tool and not a value is to take values in somewhat of a vacuum from any normative substance that tends to come with them. Here, I have supplied a description of transparency and wrapped it in an ethical intuitionist theory (and certainly other approaches may be similarly fitting) with normative implications. This demonstrates what transparency ought to be, and what acts it should entail. I recognise that as a value it supports prima facie duties, and only after careful consideration are our final duties clear. Our final transparency related duties are always ethical, they consider all the relevant reasons for or against their doing.

This also demonstrates that transparency, although subsidiary, is an ethical principle, and one which should not result in misinformation, excessive information, or other kinds of potential harms. Transparency requires the production of information that is available, findable, accessible, understandable and relevant, to a potentially rather open audience only by default, but the default stance is prima facie, and can be overridden. This prescribes effective information management. In addition, as the default stance is prima facie and what shape our final duty takes depends on the relevant circumstances and countervailing reasons, our final duty will always be to only render appropriate information available, findable, accessible, and understandable to the correct agents. Any failure to discharge these duties does not disprove that transparency is an ethical principle, it proves it, that it is effectively a rule that can be broken resulting in some harm. It is not simply pro-ethical. Furthermore, discharge of transparency related duties are right acts, and there is inherent goodness in them when done from the right motivation.

In summary, transparency is an instrumental state of affairs that when instantiated in a manner cognisant of final duty, is always good, is always valuable. It is a value, and it is an ethical principle. It is an instrument, but it is not a blunt one. Transparency, properly conceptualised, is ethical and can direct ethical action. As argued by Birchall (2011, p. 65), "…rather than there being something awry in the concept of transparency itself, action, reform, engagement are called upon to valorize and renew it."

9 Conclusion

In the preceding, I have elaborated on a normative theory of transparency, and defended its instrumental value and its status as bona fide ethical principle. I have appealed to ethical intuitionism to supply transparency with an explicitly normative backbone, framing it securely as an ethical principle that demands that relevant agents actively engage in processes of discovery, sharing, and responding, with relevant and understandable information. The focus of analysis of this account of transparency presented was algorithms in justice and security. Some transparency of algorithms is clearly an ethical requirement of responsible data innovation in justice and security, and I agree with Robbins (2019a) that an upper bound of transparency, explicable (or explainable) AI, is an essential requirement where algorithms present a significant risk of harm.

Reasons can be offered for offering more or less information to different agents. Duties to share different kinds of information with particular agents can be defeated, depending for example on the risk of harm sharing or withholding such information might entail. Our state of knowledge about an object (an algorithm) can be fairly graded, at least to different agents involved.

A form of qualified transparency can be morally justified, and may pose a solution where information cannot be widely disseminated but there is nevertheless a need for accountability. Any curtailment of transparency in this regard, however, must be somewhat transparent itself and justified accordingly.

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). The author gratefully acknowledges VW Data P4 colleagues Ibo van de Poel (TU Delft), Marc Steen (TNO), Tjerk Timan (TNO), and Remco Boersma (Dutch Ministry of Justice and Security) for collaboration and long conversations about transparency and algorithms in justice and security. The author also gratefully acknowledges the anonymous peer reviewers and managing editor for their insightful comments. Any errors remain the author's.

References

Adadi, A., and M. Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Ananny, M., and K. Crawford. 2018. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20 (3): 973–989. https://doi.org/10.1177/1461444816676645.

Angwin, J., et al. 2016. *Machine Bias*, *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 19 Oct 2018.

Aristotle. 2004. *The Nicomachean Ethics*, New Ed edition, ed. H. Tredennick (trans: Thomson, J.A.K.). London, New York: Penguin Classics.

Audi, R. 2005. *The good in the right: a theory of intuition and intrinsic value*. Princeton: Princeton University Press.

Audi, R. 2015. *Reasons, rights, and values*, 1st ed. Cambridge: Cambridge University Press.

Barocas, S., and A.D. Selbst. 2016. Big Data's disparate impact. *California Law Review* 104: 671–732.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Birchall, C. 2011. Transparency, interrupted: secrets of the left. *Theory, Culture & Society* 28 (7–8): 60–84. https://doi.org/10.1177/0263276411423040.

Brayne, S. 2017. Big data surveillance: the case of policing. *American Sociological Review* 82 (5): 977–1008. https://doi.org/10.1177/0003122417725865.

Burrell, J. 2016. How the machine "thinks": understanding opacity in machine learning algorithms. *Big Data & Society* 3 (1): 2053951715622512. https://doi.org/10.1177/2053951715622512.

D'Ignazio, C., and L. Klein. 2018. *Data Feminism*. MIT Press. https://bookbook.pubpub.org/data-feminism. Accessed: 17 Sep 2019.

Diakopoulos, N. 2015. Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3 (3): 398–415. https://doi.org/10.1080/21670811.2014.976411.

D'ignazio, C. and L.F. Klein. 2020. *Data Feminism*. Cambridge, MA: MIT Press.

Ess, C.M. 2020. Interpretative Pros Hen Pluralism: from computer-mediated colonization to a pluralistic intercultural digital ethics. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00412-9.

Etzioni, A. 2010. Is transparency the best disinfectant? *Journal of Political Philosophy* 18 (4): 389–404. https://doi.org/10.1111/j.1467-9760.2010.00366.x.

Ferguson, A.G. 2017a. Policing predictive policing. *Washington University Law Review* 94 (5): 1109–1189.

Ferguson, A.G. 2017b. *The Rise of Big Data policing: surveillance, race, and the future of law enforcement*. New York: NYU Press.

Fitzpatrick, N. 2020. The Data City, the idiom and questions of locality. *Etica & Politica* XXII (2): 19–32.

Fleischmann, K.R., and W.A. Wallace. 2005. A covenant with transparency: opening the black box of models. *Communications of the ACM* 48 (5): 93–97. https://doi.org/10.1145/1060710.1060715.

Floridi, L. 2012. Semantic information and the network theory of account. *Synthese* 184 (3): 431–454. https://doi.org/10.1007/s11229-010-9821-4.

Fox-Decent, E. 2011. *Sovereignty's promise: the state as fiduciary*. Oxford, New York: OUP Oxford.

Gilpin, L. H., D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining explanations: an overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. https://doi.org/10.1109/DSAA.2018.00018.

Haraway, D. 1988. Situated knowledges: the science question in feminism and the privilege of partial perspective. *Feminist Studies* 14 (3): 575–599. https://doi.org/10.2307/3178066.

Hayes, P., and Damian J. 2020. Care ethics and the responsible management of power and privacy in digitally enhanced disaster response. *Journal of Information, Communication and Ethics in Society* 18 (1): 157–174. https://doi.org/10.1108/JICES-02-2019-0020.

Hayes, P., I. van de Poel, and M. Steen. 2020. Algorithms and Values in Justice and Security. *AI & SOCIETY* 35 (3): 533–555. https://doi.org/10.1007/s00146-019-00932-9.

Heald, D. 2006. Varieties of transparency. In *Transparency: the Key to Better Governance?*, eds. Hood, C., Heald, D., pp. 25–43. Oxford: Oxford University Press for The British Academy. https://global.oup.com/academic/product/transparency-the-key-to-better-governance-9780197263839?q=9780197263839&lang=en&cc=gb. Accessed: 19 Oct 2018.

High-Level Expert Group on AI. 2019. *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top. Accessed: 5 Aug 2020.

Hood, C. 2007. What happens when transparency meets blame-avoidance? *Public Management Review* 9 (2): 191–210. https://doi.org/10.1080/14719030701340275.

Hulstijn, J., and B. Burgemeestre. 2014 Design for the values of accountability and transparency. In *Handbook of ethics, values, and technological design: sources, theory, values and application domains*, ed. J. van den Hoven, P.E. Vermaas, I. van de Poel, pp. 1–25. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-6994-6_12-1.

Hursthouse, R., and G. Pettigrove. 2018. Virtue ethics. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. Winter 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/. Accessed: 22 Oct 2019.

Kaspar, D. 2012. *Intuitionism.* New York: Bloomsbury Academic.

Kemper, J., and D. Kolkman. 2018. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*. https://doi.org/10.1080/1369118X.2018.1477967.

Kitchin, R. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20 (1): 14–29. https://doi.org/10.1080/1369118X.2016.1154087.

Lepri, B., et al. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31 (4): 611–627. https://doi.org/10.1007/s13347-017-0279-x.

Lum, K., and W. Isaac. 2016. To predict and serve? *Significance* 13 (5): 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x.

Mecacci, G., and F. Santoni de Sio. 2019. Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-019-09519-w.

Miller, S. 2009. *The moral foundations of social institutions: a philosophical study*, 1 edition. Cambridge, New York: Cambridge University Press.

Miller, T. 2017. Explanation in artificial intelligence: insights from the social sciences. . http://arxiv.org/abs/1706.07269 [cs]. Accessed 22 May 2019.

Mittelstadt, B.D., et al. 2016. The ethics of algorithms: mapping the debate. *Big Data & Society* 3 (2): 2053951716679679. https://doi.org/10.1177/2053951716679679.

Mittelstadt, B., C. Russell, and S. Wachter. 2019. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. New York: ACM (FAT* '19). https://doi.org/10.1145/3287560.3287574.

O'Neill, O. 2002. BBC - Radio 4 - Reith Lectures 2002 - A Question Of Trust - Lecture 4 - Trust and Transparency. 2002. http://www.bbc.co.uk/radio4/reith2002/lecture4.shtml.

O'Neil, C. 2016. *Weapons of math destruction: how big data increases inequality and threatens democracy*, 1st ed. New York: Crown.

O'Neill, O. 2004. Accountability, trust and informed consent in medical practice and research. *Clinical Medicine (London, England)* 4 (3): 269–276. https://doi.org/10.7861/clinmedicine.4-3-269.

Pasquale, F. 2010. 'Beyond innovation and competition: the need for qualified transparency in internet intermediaries. *Northwestern University Law Review Chicago* 104 (1): 105–173.

Pasquale, F. 2016. *The Black Box Society: the secret algorithms that control money and information.* Cambridge, London: Harvard University Press.

Plaisance, P.L. 2007. Transparency: an assessment of the kantian roots of a key element in media ethics practice. *Journal of Mass Media Ethics* 22 (2–3): 187–207. https://doi.org/10.1080/08900520701315855.

Randall, T.E. 2020. Justifying partiality in care ethics. *Res Publica* 26 (1): 67–87. https://doi.org/10.1007/s11158-019-09416-5.

Richardson, R., J. Schultz, and K. Crawford. 2019. Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review* 94 (2): 192–233.

Robbins, S. 2019a. A misdirected principle with a catch: explicability for AI. *Minds and Machines* 29 (4): 495–514. https://doi.org/10.1007/s11023-019-09509-3.

Robbins, S. 2019b. AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & Society*. https://doi.org/10.1007/s00146-019-00891-1.

Robbins, S., and A. Henschke. 2017. The value of transparency: bulk data and authoritarianism. *Surveillance & Society* 15 (3/4): 582–589. https://doi.org/10.24908/ss.v15i3/4.6606.

Roeser, S. 2011. Moral emotions and intuitions. *Palgrave Macmillan UK*. https://doi.org/10.1057/9780230302457.

Roeser, S. 2017. *Risk, technology, and moral emotions*, 1st ed. New York: Routledge.

Ross, D. 2003. *The right and the good*, 2nd ed. Oxford: Oxford University Press.

Santoni de Sio, F., and J. van den Hoven. 2018. Meaningful human control over autonomous systems: a philosophical account. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2018.00015.

Stratton-Lake, P., ed. 2003. *'Introduction', in The Right And The Good*, 2nd ed. Oxford: Oxford University Press.

Tu, Y.-C. 2014. *Transparency in Software Engineering*. Thesis. ResearchSpace@Auckland. https://researchspace.auckland.ac.nz/handle/2292/22092. Accessed 19 Oct 2018.

Turilli, M., and L. Floridi. 2009. The ethics of information transparency. *Ethics and Information Technology* 11 (2): 105–112. https://doi.org/10.1007/s10676-009-9187-9.

Vaccaro, A., and P. Madsen. 2009. Corporate dynamic transparency: the new ICT-driven ethics? *Ethics and Information Technology* 11 (2): 113–122. https://doi.org/10.1007/s10676-009-9190-1.

Vallor, S. 2018. *Technology and the virtues*. New York: Oxford University Press.

Zimring, F.E. 2017. *When Police Kill*. Cambridge: Harvard University Press.