Artificial Intelligence and the Value of Transparency

Joel Walmsley

Department of Philosophy
University College Cork
Ireland

j.walmsley@ucc.ie

ORCiD: 0000-0003-2456-6685

Abstract:
Some recent developments in Artificial Intelligence—especially the use of machine learning systems, trained on big data sets and deployed in socially significant and ethically weighty contexts—have led to a number of calls for "transparency." This paper explores the epistemological and ethical dimensions of that concept, as well as surveying and taxonomising the variety of ways in which it has been invoked in recent discussions. Whilst "outward" forms of transparency (concerning the relationship between an AI system, its developers, users and the media) may be straightforwardly achieved, what I call "functional" transparency about the inner workings of a system is, in many cases, much harder to attain. In those situations, I argue that contestability may be a possible, acceptable, and useful alternative so that even if we cannot understand *how* a system came up with a particular output, we at least have the means to *challenge* it.

Keywords: Transparency, Explainability, Contestability, Machine Learning. Bias.

1. Introduction

Alongside, and arguably *because of*, some of the most recent technical developments in Artificial Intelligence, the last few years have seen a growing number of calls for various forms of *transparency*[1] within and about the field. For example, the 2019 report from the European Commission's High-Level Expert Group on AI—entitled *Ethics Guidelines for Trustworthy AI*—features the notion of transparency prominently, and the European Union's *General Data Protection Regulation* (GDPR) includes the stipulation that, when a person is subject to an automated decision based on their personal information, he or she has "the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."[2] In part, these calls respond to an epistemic limitation; machine learning techniques, together with the use of "Big Data" for training purposes, mean that many AI systems are both too complex for a complete understanding, and faster and more powerful than human cognition (at least, on the relatively narrow set of tasks for which AI is designed). Of course, in many cases, "complete understanding" is neither desired nor required; we are perfectly happy to interact with technology by adopting Dennettian[3] "intentional" or "design" stances (rather than the more complete but cumbersome "physical stance") so long as the system functions correctly, and the respects in which it is *not* transparent are roughly neutral along ethical, political or commercial dimensions. But given that we increasingly and preferentially trust AI systems, and that we do rely on them to make decisions, recommendations and predictions in a variety of socially significant and morally weighty contexts (for example, not just regarding what video to watch or what product to buy, but also whether

[1] Sometimes also discussed under the heading of "explainability," "explicability" or "understandability" (e.g., by Robbins, 2019) or with reference, also, to "accountability," "intelligibility" and "interpretability" (e.g., in Floridi et al, 2018)

[2] General Data Protection Regulation, Recital 71, available at https://gdpr-info.eu/recitals/no-71/

[3] See Dennett, (1971)

a person qualifies for job interview, a loan, or for parole), the call for transparency has acquired an ethical (and legal) dimension too. Furthermore, these dimensions intersect: both in the popular media, and within specialised technology circles, we find a steady stream of examples and anecdotes of problematic biases, prejudices and other errors that have been automated and reinforced—albeit, sometimes, unwittingly—because of our reliance on AI systems that we do not fully comprehend.

In this paper, I start by teasing out the epistemic and ethical considerations—and the *examples* within AI—that have led to this contemporary concern for transparency, by asking "Why *now*?" and "Why *care*?" I go on to note that there are actually several different conceptions of transparency that have been both invoked and sought, and I taxonomise and explore their main features. Although the appeal of transparency is obvious as a policy objective or an antidote to the problems often described, I argue that it is not always possible (or maybe even *desirable*) especially when it comes to the inner workings of such systems. Instead, I tentatively suggest that as an alternative when transparency cannot be achieved, the best we can do is to build AI systems so that their outputs can be *contested*. This can take several forms. On the one hand, fortunately, machine learning includes some technical tools—in the form of "reinforcement learning"—so that contestability can be included by design in the training process of such systems. On the other hand, I argue, we should also pay greater attention to contestability as a matter of policy about how AI systems are deployed. This combination would mean that that even if, in some cases, we cannot fully understand how an AI system works, the possibility of challenging its decisions may still allow us to achieve the ethical goals—such as justice, fairness or impartiality—that we value.

2. Transparency: Why now?

On the epistemic side of the issue, the demand for transparency stems from the intersection of three recent developments in computer science, AI, and human-computer interaction: machine learning (hereafter "ML"), big data, and the growing reliance on (or even *trust in*) algorithmic systems to make predictions, provide recommendations and take decisions. Let us examine each in turn.

It is useful to think of ML in contrast to classical programming, or what Haugeland (1985) called "Good Old Fashioned AI" or GOFAI. In the latter, *humans* write the programs and provide the input data, and the computer applies the former to the latter in order to come up with an output. It is striking that this general "input—processing—output" schematic can be applied to such a wide range of tasks, whether logico-mathematical calculation, puzzle-solving, chess playing, or dyadic conversation (to name just a few of the famous examples). GOFAI is, in a sense, transparent by design, since the programs are based on (but therefore also limited by) whatever the human programmers can think up. As Robbins (2019) points out, GOFAI systems effectively consist of explicitly coded rules, and so transparency amounts to the ability to inspect the code and derive the output: "Opacity with regard to

this type of automation would only occur if the institutions doing the automating did not want people to know how the decisions are being made" (p. 503) Thus, even if the intended end-user does not understand how a program works, the fact that its code was written by a human means that there's no fundamental, *in principle*, impediment to transparency. ML turns this process on its head, as it were; humans provide the data (which may or may not be labelled) and the computer itself comes up with a set of rules, or mapping functions, that describe patterns and correlations within that data set. The aim is often that those rules can subsequently be used to make predictions or recommendations about future or additional data points, because the ML system uncovers patterns that were not already apparent. Crucial to the success of ML is the fact that the system itself can modify these rule-like mappings (as additional data is acquired or generated) in order to improve accuracy: this is the "L" in machine *learning*. Because the rules track patterns of which we were not previously aware (and thus, one of ML's oft-cited attractions is its ability to facilitate new discoveries) and those rules are incrementally modified as more data is added, human developers can quickly lose an understanding of how the machine actually works.

In order that ML systems can formulate and update their rules, lots of data is required, and it just so happens that we now live in an era where not only is hardware fast and powerful enough to process enormous amounts of it, but also human users are willing to provide it for free (indeed, some authors, such as Smith (2019, p.52) have even described ML as a *post*-Big Data technical development). The big data revolution is often characterised as an explosion along the dimensions of the "three Vs" of velocity, volume and variety: data is being collected faster and faster, in ever-increasing amounts, and in a bewildering variety of forms (text, images, audio etc.) And the rise of interactive social media and the so-called "Web 2.0"—with billions and billions of tweets, likes, shares and search queries—has given rise to extremely large data sets of details about users and their connections, on which ML systems can be trained to uncover patterns, associations and trends. When coupled with data that is collected and curated for specific purposes as well, the patterns encoded in these rapidly expanding sets can subsequently be used to serve up better targeted adverts for products you might buy, to recommend music or movies or restaurants you might like, and to make predictions about your creditworthiness or likelihood of committing a crime (more on which later).

For now, the *epistemic* point is that—at the intersection of ML and "big data"—we have built automated systems that we don't fully understand, that are much faster and more powerful than the human mind, and which are trained on data sets that are too large for us to comprehend (even *with* the use of conventional tools such as spreadsheets and pocket calculators). Accordingly, a new phenomenon has arisen where we see AI experts and tech-sector insiders giving TED talks, and writing op-ed pieces with click-bait headlines making reference to "The Dark Secret at the Heart of

AI,"[4] likening it to alchemy,[5] sorcery,[6] and various other forms of "black box"[7] mystery.[8] Of course, it is tempting to invoke Arthur C. Clarke's (1972) famous "third law"—"Any sufficiently advanced technology is indistinguishable from magic"—and hope that this sense of scientific awe will dissipate. But the calls for transparency that respond to these epistemic limitations become more pressing when one notes a third, but somewhat less well-known, recent development in the field of human-computer interaction.

Jennifer Logg and colleagues (e.g., Logg, Minson and Moore, 2019) have recently documented a phenomenon that they call "algorithm appreciation," whereby, in a variety of situations, people seem to prefer advice when they think it is provided by a computer system, rather than by a human. This seems to reverse a tendency (first noted in the 1950s) towards *dis*trust of the output of algorithms in comparison to human judgment (dubbed "algorithm aversion" by Dietvorst, Simmons and Massey, 2015). In a series of experiments, Logg et al. asked participants to make a variety of judgments (e.g., a numerical estimate concerning a visual stimulus, and predictions about the popularity of songs, or the likelihood of romantic attraction between two individuals) and found that people were significantly more likely to adhere to advice about those judgments when they thought it came from an algorithm rather than from a person. This seems to demonstrate an increasing willingness to *trust* automated systems, even when we are not sure (or not concerned with) how they work. The many anecdotes of people who have ended up driving through fields and into rivers as a result of following the directions of a sat-nav or GPS system are testament to this new phenomenon.

The implications of the foregoing considerations should be clear; the fastest moving areas of contemporary AI are ones where humans are at an epistemic disadvantage when it comes to understanding the systems we have built. At the intersection of machine learning, big data, and algorithm appreciation, we have a situation where we don't fully understand the machines, they're faster, more powerful and more complex than us, but we trust them preferentially nonetheless. If our use of such systems were restricted to recommender systems for movies, music and restaurants, this might not be such a problem. But the epistemic limitations outlined here start to have much deeper

[4] https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

[5] https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy or https://www.youtube.com/watch?v=oXAEL8IjUlo

[6] https://www.sapiens.org/column/machinations/ai-as-magic/

[7] https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box

[8] http://oecdobserver.org/news/fullstory.php/aid/5543/A_mystery_in_the_machine.html

moral consequences when these AI systems are put to use in predicting social outcomes such as job performance, criminal recidivism or creditworthiness.

3. Transparency: Why Care?

Building on the three epistemic concerns outlined above, that there are three similar reasons that contribute to the ethical dimension of the call for transparency: the potential for "machine bias," worries about the use of opaque AI in what Floridi et al., (2018) have called "socially significant" contexts, and the possibility of perpetuating what Fricker (2007) has described as "epistemic injustice."

One consequence of the way that ML systems are trained is that they can very easily end up, literally, encoding more general biases and societal prejudices that are represented in the big data set. One notorious recent example concerns the way in which Google Translate handles languages in which the third person singular is gender-neutral, such as Turkish, Swahili and Finnish. Until recently, when translating from such languages into English, for example, the tool tended to return answers that conform to stereotypical gender roles and characteristics (for example, the Turkish "O bir doktor" was translated as "*he* is a doctor", whereas "O bir hemşire" returned "s*he* is a nurse"; "O bir mühendis" was translated as "*he* is an engineer", whereas "O bir aşçı" gave "*She* is a cook"). This has, appropriately enough, been dubbed "machine bias" but it's important to note that it's not so much that the AI *system* (or even its designer) is biased; rather the repetition of sexist stereotypes arises as an unintended artefact of the way in which the system is trained. Google Translate works by learning from the patterns found in millions of bilingual texts scraped from the world wide web, and since those patterns ("in the wild" as it were) tend to exemplify more general societal patterns, including prejudices about gender roles, characteristics and occupations, so the translation engine comes to replicate them.

Relatedly, in 2018, a news story broke about similar sexist bias in a recruitment tool that was developed by Amazon.[9] The goal had been to develop a system that could automatically scan the CVs of job applicants and come up with a short-list of candidates that should be hired. But because the system was trained on historical data consisting the CVs of many years' worth of previous applicants and employees, the majority of whom were men, the system effectively taught itself to prefer male applicants by penalising CVs that contained keywords relating to women. In this case, the problematic male-domination of the industry in question was echoed and amplified by the automated recruitment process. Again, this example of "machine bias" is not necessarily representative of any explicit prejudices of the human designers or recruiters (even if, perhaps, they should have known better);

[9] See https://www.bbc.com/news/technology-45809919

rather, the system's undeniably sexist recommendations are an unforeseen (and probably unintended) consequence of its training regime.

In both of these cases, it's important to note that the problem was identified *and fixed* with human intervention: Google Translate now offers multiple translations from gender neutral languages, and Amazon has scrapped its automated recruitment tool. This indicates that in some cases where transparency is not present, it is important at least that the users of AI systems have a way of challenging or contesting automated decisions. I will return to this issue in Section 5 below. The two examples mentioned here are not *particularly* opaque (it is relatively easy to see what went wrong on the basis of the training data set in both instances), but as we shall see, in cases where transparency is otherwise difficult to achieve, contestability may be the best way to ensure fairness or impartiality.

A more worrying use of ML systems to make morally consequential and socially significant decisions comes from the widely discussed "COMPAS" algorithm,[10] which assigns a risk score to prisoners based on its calculation of the likelihood of recidivism subsequent to release. The system and its output scores have been used in several US states, in decisions about bail, sentencing, and parole. The superficial appeal of using an automated procedure is clear; it can calculate a score faster than a human, it can include a larger number of complex contributory factors, and (so it is claimed) it can do all of this in a way that takes the decision out of the hands of humans who may otherwise be guided by their own *personal* biases and instincts. As Marcus and Davis (2019) put it, "the decisions that the program is making, being computed 'algorithmically,' have an aura of objectivity that impresses bureaucrats and company executives and cows the general public."

The use of COMPAS is understandably controversial, but four particular critiques are of special relevance to the present discussion of transparency. First, continuous with the previous examples of machine bias, COMPAS appears to display significant racial disparities in its risk assessment, possibly as a result of being trained on historical data that implicitly encode the notorious racial prejudice in the US criminal justice system. Angwin et al. (2016) found that African Americans were almost twice as likely as Caucasians to be labelled by COMPAS as a higher risk despite not in fact going on to re-offend, and Caucasians were more likely to be labelled a lower risk compared to African Americans, even though they did in fact to go on to commit other crimes. Second, Dressell and Farid (2018) have shown that COMPAS is not significantly better in its predictions than untrained humans anyway. So not only is COMPAS prejudiced, it's also not especially accurate. Third, since the COMPAS system is owned and developed by a for-profit company, the calculations and proprietary

[10] Developed by Northpointe (now renamed *Equivant*); the acronym stands for "Correctional Offender Management Profiling for Alternative Sanctions"

software used to derive a risk score are considered commercially sensitive trade secrets and therefore not publicly disclosed for audit. Finally, although judges in several states are allowed to use COMPAS scores to inform their sentencing decisions, in some cases, defendants have not been permitted to confront and cross-examine the system in the way that they would if it were a human witness against them: whether the use of COMPAS in this way is a violation of due process is a matter still under debate. Again, I will return to this last point in section 5, when we consider contestability and the possibility of challenging the outputs of AI systems even when they are not transparent.

It's worth pointing out here that the case of COMPAS also illustrates the importance of transparency for the familiar debate concerning an alleged trade-off between fairness and accuracy in predictive AI systems. It is widely held that fairness and accuracy are in tension with each other (i.e., that increasing one will decrease the other) because, to put it crudely, the demands of fairness might require imposing constraints on the set of possible outcomes suggested by the data alone (for example, if one attempts to equalise the false-positive rates between two groups of a protected category (such as race)—i.e., to ensure fairness of outcome—one may have to ignore or discount other patterns in the data such as differential base-rates of recidivism between those two groups.) There is a substantial literature on this discussion both in general (see Dutta et al., 2020) and with respect to COMPAS in particular (see Kleinberg et al., 2016 and Angwin & Larson, 2016), which I will not go into here. Note, however, that in order to make a judgment call about the relative balance we might wish to strike between fairness and accuracy, we actually have to know *how the system works* in the first place. Transparency seems to be a requirement for making the trade-off in either direct, In order to know whether we *are* trading fairness for accuracy of a prediction or judgment, we need to know the extent to which a protected characteristic (such as race or gender) actually contributes to a decision (e.g., about parole, credit-worthiness or employability). Again, I will also return to this topic in section 5.

One final consideration brings together the epistemic *and* ethical concerns about transparency. This is the possibility of what Fricker (2007) has called "epistemic injustice," when somebody is wronged "specifically in her capacity as a knower."[11] There are two types of epistemic injustice, according to Fricker, and the potential for *both* of them seems to be present in some computerised diagnostic decision support systems (particularly so-called "patient-facing digital symptom checkers").[12]

[11] Fricker (2007), p.20

[12] Here, I have in mind examples such as the "GP at Hand" system developed by Babylon Health, which provides some NHS services in the UK. See https://www.gpathand.nhs.uk/

On the one hand is what Fricker calls *testimonial injustice* where "prejudice causes a hearer to give a deflated level of credibility to a speaker's word."[13] This often manifests along depressingly familiar prejudicial lines of race, class or gender, and negatively influences credibility judgments simply in virtue of the group to which the speaker belongs. There is a risk, with decision "support" systems, that although they are officially designed as a tool to enhance professional judgment, they may in fact be regarded as *replacements* for the human decision-makers that they emulate. As Danaher (2019, p.9) writes, "Machines are now being designed, built, and implemented to replace, not simply complement, their human coworkers." In the case of medical diagnosis, for example, there is a risk of testimonial injustice because, when coupled with the increasing "algorithm appreciation" discussed in the previous section, a GP's opinion may be discounted or rejected *simply because the GP is human*.

On the other hand is what Fricker calls *hermeneutic injustice*, where a person is disadvantaged when it comes to making sense of their own experiences, or rendering them communicatively intelligible, because of a gap in interpretive resources. Taking the case of automated medical diagnosis as an example again, a person's ability to comprehend or communicate the nuances of a set of symptoms or health concerns may be too tightly constrained by the requirement that it be reported through the tedious connected boxes of a diagnostic flow-chart, or a frustrating hierarchy of pre-programmed menu options. Not only does this run the risk of inhibiting a person's understanding or communication of the experience that led them to seek the help of the app in the first place, but it may also negatively influence health *outcomes* (since it is well known that a patient's *active* engagement is correlated both with better diagnosis *and* prognosis).

Given the considerations canvassed here, it is hardly surprising that epistemic limitations and their ethical consequences have led to calls for greater transparency about contemporary AI systems, both as a general scientific goal, and as a policy objective. The fact that we often don't understand how ML systems work, but we rely on them nonetheless in a variety of socially and morally significant situations, warrants closer attention. But the kinds of transparency envisaged are just as diverse as the specific problems to which they respond. Accordingly, it is important to identify and discuss the details of these differences in order to assess the extent to which they are possible, whether they resolve the problems to which they respond, and if not, what alternatives might be available.

4. Varieties of Transparency

It may be convenient to divide the way in which the notion of transparency has been invoked into two major categories. One, we can think of as a kind of "outward" transparency, since it concerns the relationship *between* the AI system and things external to it (especially developers, users and the

[13] Fricker (2007), p.1

media). Here, we may be concerned with transparency about how and why the system was developed, or how it is described in both technical and popular presentations, or what the user knows about its deployment. The other category, we can think of as a kind of "functional" transparency, since it concerns the inner workings of the system itself. Here, we may be concerned with transparency about, for instance, how a system came up with a particular judgment or recommendation on a given occasion, or how the various factors, in general, are weighted and combined by the system. We need not draw the distinction between outward and functional transparency in a particularly strong way—since, as we have seen, both aspects may be involved in any given example—but it may be a useful starting point, and (as I shall argue) it also maps onto the ease with which we can achieve the transparency being sought.

On the side of "outward" transparency, the first sub-type operates at a kind of meta-level that has recently been invoked in the "values in science" debate. Here, the concern is not so much with transparency *as* a value, but rather with transparency *about* values. Elliott (2017), for example, describes transparency as a condition for the legitimate incorporation of values into science: since fact-value entanglements are practically unavoidable in many scientific domains, Elliott writes, "… the best we can do is to be transparent about our assumptions so that others can take them into account… scientists should strive for transparency about their cognitive attitudes towards theories and about the role of values in their descriptions of scientific information."[14] This kind of transparency *about* values may be straightforwardly applied to research and development in the field of AI.

The aforementioned European Commission guidelines for trustworthy AI, for example, make reference to outward meta-level transparency. "Explicability" is one of the four fundamental ethical principles outlined in the guidelines, and under that heading, explicit reference is made to the fact that the *purpose* of AI systems should be openly communicated and that explanations of "… design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency)." (p.18). The idea here is that those who develop, deploy (and materially *profit from*) AI systems should be open about the values and motivations that drive them, since these may be legitimate factors for consideration by the users of AI systems (or, indeed, those who are subject to decisions made by them).

A second kind of outward transparency is predominantly descriptive, and concerns the way in which an AI system and its capabilities are communicated or portrayed in the broader societal context (especially in popular presentations in the media). It would not be too much of a stretch to say that a great deal of hype currently surrounds any announcement of any new breakthrough in AI—partly

[14] Elliott, 2017, p.171

because it is often closely tied to commercial interests—and both programmers and journalists are sometimes guilty of overselling or exaggerating these developments. In October 2019, for example, *New Scientist* magazine tweeted an announcement about an article concerning new research in agent-based modelling (Lawton, 2019) with the sensationalist exaggerated claim "Predicting the future is now possible with powerful new AI simulations that can model e*very conceivable social interaction*" (my emphasis).[15] Marcus and Davis (2019) have documented a large number of similar cases of more-or-less irresponsible presentation and reporting of AI systems in a variety of domains.

So outward descriptive transparency simply amounts to an honest portrayal of what a system can (and cannot) actually do. Again, the European Commission guidelines explicitly recommend this kind of transparency as a requirement, stating that "… the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This should encompass communication of the AI system's level of accuracy, as well as its limitations." (p.18)

A related, third, kind of outward transparency concerns the extent to which a user is aware that they are dealing with an AI system rather than with a human. Increasingly, some companies use automated AI systems—effectively, chatbots—to handle and re-direct initial queries from customers and clients, and given their increasing sophistication, it may not always be clear to the user that they are in fact interacting with a computer rather than a human customer service representative. The airline Aer Lingus, for example, uses a bot to manage initial contact with customers via direct messaging on Twitter (asking basic factual questions about one's booking reference, flight number and dates of travel), without announcing that fact unless directly asked, as a kind of triage, before handing the conversation over to a human representative. The European Commission guidelines recommend against this, in favour of what we might think of as "user-facing outward transparency," stating (p.18): "AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such." Similarly, the UK's Engineering and Physical Sciences and Research Council, in collaboration with the Arts and Humanities Research Council, have recently drafted a list of "principles of robotics," Rule 4 of which states: "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." (see Boden et al., 2017).

The three types of outward transparency effectively consist of—and can be achieved with—honesty about the development, use and deployment of AI systems. By contrast (and perhaps more

[15] See here: https://twitter.com/newscientist/status/1180916793126326273

challenging) are several varieties of "functional" transparency that concern, as it were, the inner workings of an AI system.[16] It is useful to distinguish further between the explainability of a *particular* decision or action taken by an AI system (which we can label "token functional transparency") and the explainability of the AI *system* in general (which we can label "type functional transparency"). It is also worth noting that the type/token distinction in functional transparency also has a parallel in the kinds of problematic discrimination or unfairness that I examined earlier, but with a significant difference in ease of detection or diagnosis. A bias at the level of *types* (e.g., in the form of general racial discrimination) is at least in principle easier to detect, because it is susceptible to a statistical analysis (e.g., of the kind that Angwin et al., 2016, conducted) showing that one protected category of people has been treated worse than another. A token discriminatory decision (about an *individual's* parole eligibility or creditworthiness, for example) is harder to confirm since it can often be defended (or explained away) with some plausible argument or other. This point will become significant in section 5 when I turn to the question of contestability[17].

The distinction between kinds of functional transparency is often glossed over, in practice. For example, under the heading of "traceability," the European Commission guidelines call for both type *and* token functional transparency (p.18), recommending that "The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system." The focus here is especially on how to audit or trace processes when something has gone wrong, or when an error has been identified (more on which later), but for now we can illustrate the difference between types of functional transparency using an example mentioned earlier.

Consider the use of COMPAS to provide a risk score for an individual seeking to be released on parole. On the one hand, the individual under consideration may (legitimately) ask why, on this occasion, the system assigned them a risk score of 80%; this would amount to seeking an explanation for the system's particular (token) decision on this particular (token) occasion. Token functional transparency would require showing how the system computed the risk rating of 80% on this occasion, given the individual's answers to the 137 standardised questions on which the score is based (e.g., "Was one of your parents ever sent to jail or prison?" or "How often have you moved in the last

[16] The nascent sub-discipline of "explainable AI" (or xAI) is especially focussed on this kind of transparency.

[17] I thank an anonymous reviewer for drawing this point—about the parallel type/token distinction at the level of problematic discrimination—to my attention.

twelve months?")[18] On the other hand, the individual may also legitimately ask for a general explanation of how the system factors in the various questionnaire answers, and what the relative weightings of each category happen to be. Type functional transparency would amount to showing how, for example, questionnaire answers in relation to social environment or residence were weighted in comparison to questionnaire answers about education or family.[19]

As I mentioned above, current legislation tends to blur the distinction between type- and token functional transparency, and in some cases this is precisely because the two overlap or are intimately connected. For example, if one is able to give a satisfactory type-functional explanation of how an AI system assigns risk scores in general, then a transparent token-functional explanation (of a particular decision) could straightforwardly follow if we know the individual's questionnaire answers. In other cases, however, type- and token functional transparency should be kept apart. For example, even if a system is not type-functionally transparent (i.e., we don't have a complete explanation of how it assigns scores in general), we might still be able to give a token-functional explanation (of a particular decision) if we know that an individual scored especially highly on a handful of key measures. Similarly, even without considering *token*-functional explanations, we may wish to focus an ethical evaluation of a system at the level of type-functional transparency, by asking, for example, whether it's morally legitimate to include factors such as a person's postcode, or educational background. These are significant questions that deserve further study: for now, let us note that the distinction between type- and token-functional transparency can inform that debate.

Note, however, that because of the epistemic considerations I canvassed in Section 2 (especially concerning machine learning with big data) the varieties of functional transparency will be significantly harder to achieve that the varieties of outward transparency. We may, at most, be able to provide an impressionistic or qualitative account of the underlying calculation or relative weightings, constituting something like an incomplete Hempelian explanation sketch: a "more or less vague indication of the laws and initial conditions considered as relevant [that] needs 'filling out' in order to turn into a full-fledged explanation."[20] Of course, on Hempel's view, such explanation sketches are perfectly legitimate parts of *scientific* enquiry; indeed in everyday or common-sense folk psychological prediction and explanation, we are perfectly content with adopting a Dennettian intentional stance and making do with partial explanations that are replete with qualifications, hedges

[18] See Angwin et al (20116) for more detail and the actual COMPAS questionnaire used, here: https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

[19] See this recent comic for a similar—if light-hearted—idea of type functional transparency with respect to machine learning based hiring algorithms, like that discussed in Section 2: https://xkcd.com/2237/

[20] Hempel (1965) p.238

and *ceteris paribus* clauses. *Other humans* are not transparent in the way that is often recommended for AI systems, and so demanding full transparency for the latter might simply amount to holding them to an unrealistically high standard (see, e.g., Zerilli et al, 2018).

Further, in other cases, we might wish to avoid full functional transparency in decision-making systems for at least two reasons. First, if the system's inner workings become known, bad actors may exploit that knowledge to "game" the system. If one knows, for example, that an automated CV-scanning algorithm has been trained to prefer resumés from people who attended particular universities, one could insert the words "Oxford" or "Cambridge" into a CV in invisible white text, to cheat the screening process (see Buranyi, 2018). Second, as pointed out by Robbins (2019), part of the rationale for using AI to make decisions often stems from the fact that we *don't already know* what are the relevant considerations in the first place. As I mentioned earlier, one of the strengths of ML is that it can *discover* patterns, of which we were not previously aware, in big data sets; if we already knew which considerations should be used, we wouldn't need ML, and if we insist on ensuring type and token functional transparency in building ML systems, there may be no advantage to using them.

The point here is that because of the methods used in contemporary AI, some of the kinds of transparency that have been called for—especially type and token functional transparency—are either difficult or impossible, or else even undesirable. What alternatives can be found, in these cases? One set of possibilities would be simply to forbid the use of opaque systems in morally weighty or socially significant contexts, to restrict them to "low stakes" situations where we do not require (or care about) transparency (what Scott Robbins has called "boring AI"), or to allow users to "opt out."[21] But there is also a further alternative: as hinted above, and as suggested by, for example, Hirsch et al., (2017) and Mulligan, Kluttz and Kohli (2019), we could focus on designing and building AI systems where we have the possibility of challenging or contesting their (token) outputs. Almada (2019) has coined the expression "contestability by design" for this approach to indicate that it should not be a mere afterthought, and for the purposes of elucidation, it may be convenient to divide such suggestions into two categories; those at the level of *policy* concerning regulations for the deployment or use of AI systems, and those at the level of *function* concerning how the AI systems actually work. In the final section of this paper, I will explore the (albeit tentative) proposal that contestability—either at the level of policy, or at the level of technical implementation—may be an adequate possible substitute for functional transparency in cases where the latter is not possible. I should note that I do not wish to make concrete proposals either for how such AI systems should be designed and built, or for how their deployment should be regulated. Rather, my aim here will be to open up space for such a

[21] The European Commission guidelines, for example, recommend that "... the option to decide against this interaction, in favour of human interaction should be provided." (p.18)

discussion by showing that even if we cannot fully explain how a system works, contestability seems to be an alternative route to enabling the ethical features that we care about (such as fairness, justice, impartiality, and so on).

5. Contestability as an alternative to transparency.

Interestingly, several of the texts that call for transparency also contain a parallel reference to the possibility of challenging or disputing the output of an AI system. Indeed, in the above example of the COMPAS controversy, the fact that defendants were not permitted to cross-examine the system in the same way that they could with a human witness was just as problematic (in virtue of being a violation of due process) as COMPAS's lack of functional transparency. This suggests a possible connection between transparency and contestability; indeed the EC guidelines on trustworthy AI connect contestability with *both* outward and functional transparency, stating that the general principle of fairness "entails the ability to contest and seek effective redress against decisions made by AI systems and the humans operating them. In order to do so, the entity accountable for the decision must be identifiable, and the decision-making process should be explicable." (p.13). Similarly, the EU's GDPR regulations (quoted in Section 1) also stipulate that when an automated decision is made, or "profiling" is conducted on the basis of personal data, the individual concerned should be able both "to obtain an explanation of the decision reached after such assessment *and to challenge the decision.*" (My emphasis). Strictly, however, transparency is not a requirement for contestability and so the two can be decoupled: one need not know exactly how a decision was made in order to challenge it as erroneous, unjust or unfair. As a result, I want to suggest (albeit tentatively) that contestability may even serve as a reasonable alternative or proxy (or as a way of ensuring fairness), when transparency is not possible, and that this kind of contestability could operate both at the level of regulatory policy *and* at the level of technical design and implementation.

There are (at least) three ways in which contestability on non-transparent systems could work. The first—at the level of policy, and implicit in the EC and GDPR guidelines—simply involves ensuring that there is, as the expression goes, a "human in the loop": as Almada (2019) puts it, human review of automated decision making is seen as an "antidote to error" and human intervention can be seen as a harm-preventing form of quality-control. Effectively, this amounts to the recognition (as I mentioned before) that many of the AI systems canvassed above are properly regarded as decision-*support* systems rather than decision-*making* systems that are replacements for humans. So, for example, rather than the COMPAS algorithm determining a sentencing or parole decision itself, a judge or jury should use it as one source of (defeasible) evidence, alongside both their own judgment and contrary viewpoints from the defendant and their legal team.

There are, of course, attendant risks with adopting the human-in-the-loop approach as a policy-based form of contestability. For example, human review could mean that time-critical systems perform too slowly such that a major advantage of ML is lost, or perhaps human intervention could unwittingly introduce *other* biases (e.g., if a parole board, or defendant, is allowed to contest the recommendations of COMPAS, the extent to which it is successful will of course depend on the level of accuracy and other biases displayed by the parole board and/or the defendant themselves). These concerns are reminiscent of the debate concerning the alleged trade-off between fairness and accuracy that I discussed earlier, however we should note two further points that may at least give us some *prima facie* reason to pursue the alternative of contestability. First, as noted, although the trade-off between fairness and accuracy is a legitimate and serious concern, it is not even possible to make judgment calls about that balance without transparency. It is, however, possible to *contest* the output of a decision-making system—on the grounds of either accuracy *or* fairness—even when it is not functionally transparent, so this is one advantage of focussing on contestability for practical purposes. Second, for the time being, legislating for this kind of contestability would have the effect of locating the *responsibility* for a decision with the human-in-the-loop, even when the process that gives rise to the AI system's recommendation is not (functionally) transparent.

Further, at the levels of both policy and design/implementation, contestability can be built into the development phase of an AI system by seeking feedback from human users both with respect to the accuracy of the system and with respect to the various ethical dimensions (such as fairness or justice) about which we care. There have already been some studies in the field of human-computer interaction (see Binns et al., 2018) concerning the way in which users make such judgments about algorithmic systems (and how they compare to similar ethical evaluations of human decision-making). And this kind of data (derived from what are effectively focus groups of users) could be used, even in cases where the system's operations are not transparent, to ensure both the development of algorithms that are perceived as fair and of more general standards that developers must meet. This is one aspect, for example, of the design and testing process described by Hirsch et al (2017), in their development of a machine-learning system for psychotherapy.

Third, and importantly, there are technical mechanisms that permit contestability—as a kind of feedback loop—in the operation of the very same ML systems where transparency may be difficult to achieve. In particular, in ML systems that use *reinforcement learning*, one can devise a kind of "reward" signal and set the system's goal to maximise it, much like "operant conditioning" in behavioural psychology (see Sutton & Barto, 2018). A challenge to the system's output as the result of having it contested could serve as an error signal to the system, indicating that it has made a mistake or generated some form of problematic output. So instances where a decision is contested could be

coded as negative feedback, in order that the system can update its mapping function and avoid such mistakes in the future (see Kaas, 2020, for example).

Many ML systems—including search engines, translation programmes, and recommender systems—make use of user feedback in order to improve their performance either in general or for a particular user. The popular audio streaming service Spotify, for example, has a "radio" function in which it can make music recommendations based on a user's listening history, saved items, and other preferences. Suggested songs can then be "liked" or "disliked" and as the user "contests" the suggestions and thereby provides feedback, the system learns more about the user's preferences in order to gradually improve the recommendations (similar features may also be found in a variety of other popular platforms with built-in recommender systems, such as YouTube, Netflix, Amazon etc.) To be sure, what we have learnt from recommender systems for popular entertainment—where user feedback obviously makes them more successful—may not straightforwardly carry over to more complex and controversial cases like COMPAS. The latter is not only much more morally weighty and socially significant, but also (perhaps more importantly) it is adversarial in a way that most current recommender systems are not: the state that employs COMPAS to make sentencing recommendations, and the defendant thus sentenced, have very different goals. Nonetheless, the model of contestability-based feedback in recommender systems that use reinforcement learning does provide a limited and suggestive illustration of how ML already contains at least the technical tools for implementing contestability as a design feature.

Of course, it would be a significant challenge to extend this kind of feedback mechanism, through reinforcement learning protocols, to the kinds of ML systems discussed above. It might be too much, for example, simply to allow a defendant to give feedback on their COMPAS score (presumably to claim that it is too high or too risk-averse). But nonetheless (perhaps in conjunction with the human-in-the-loop contestability mentioned above) a feedback error signal *could* encode how often (or by how much) the algorithm's advice is overridden by parole boards or juries, or the frequency and distribution of false-positive or false-negative judgments, or whether a loan was successfully paid back on time despite predictions to the contrary.[22] If this general protocol could be extended to other ML systems, it may go some way towards allaying epistemic (and possibly ethical) concerns about a lack of transparency.[23] We might not know exactly how a system arrived at the decision it made, but we would at least have the technical capacity and a variety of ways to challenge its fairness or

[22] I am grateful to an anonymous reviewer for making this suggestion.

[23] Indeed, this is a further strategy that Hirsch et al. (2017) recommend for the design and pilot phase of ML systems in the field of mental health diagnostics. They frame it as a mechanism for improving the *accuracy* of such systems, but this use of feedback could clearly also underlie the development of systems that are (perceived as) *fair* or *just*.

accuracy—in conjunction with the other approaches I have suggested—so that contestability is central to the development and operation of AI ML systems rather than a mere afterthought.

6. Concluding remarks.

As we design and build ever more complex AI systems—especially those that use machine learning with big data sets—demands for epistemic transparency understandably grow. And the issue becomes more pressing given that we seem to preferentially trust and rely on these systems to make predictions and decisions in a growing range of socially significant and morally weighty contexts.

But the kinds of transparency that are called for and sought seem to be just as diverse as the examples that motivate them, and it may not always be possible (or even desirable) to achieve them all. We do need both transparency *about* values and transparency *as a* value, but the outward transparency that I've described above is much easier to provide than functional transparency about the inner workings of many AI systems. In the case of the latter, it may be that the best we can do is to design AI systems and policies for their deployment—using tools and techniques that are already present within the suite of ML resources—whose output can be contested or challenged. This way, even if we don't fully understand *how* a system makes a prediction, recommendation, or decision, we at least have something analogous to a right of reply, or to the due process of cross-examination so that if "computer says 'no'"[24] we can reply "Actually, 'Yes.'"

[24] See: https://en.wikipedia.org/wiki/Computer_says_no

References

Almada, M. (2019) “Human intervention in automated decision-making: Toward the construction of contestable systems. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL ’19), June 17-21, 2019, Montreal, QC, Canada*. ACM, New York, NY, USA. Available at https://dl.acm.org/doi/10.1145/3322640.3326699

Angwin, J., J. Larson, S. Mattu, L. Kirchner (2016) “Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *ProPublica*, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J., and Larson J. (2016) “Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say” *ProPublica*, 30 December 2016, https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). “ ‘It’s reducing a human being to a percentage’: perceptions of justice in algorithmic decisions.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Paper no: 377, p1-4. https://doi.org/10.1145/3173574.3173951

Boden, Margaret, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorrell, Mick Wallis, Blay Whitby & Alan Winfield (2017) “Principles of robotics: regulating robots in the real world,” *Connection Science*, 29:2, 124-129

Buranyi, Stephen (2018) “How to persuade a robot that you should get the job” *The Guardian* 4th March, 2018. Available at: https://www.theguardian.com/technology/2018/mar/04/robots-screen-candidates-for-jobs-artificial-intelligence

Clarke, Arthur C. (1972) *Profiles of the Future: An Inquiry into the Limits of the Possible* (2nd Edition) (Gateway Books)

Danaher, John. (2019) *Automation and Utopia: Human Flourishing in a World Without Work* (Cambridge, MA: Harvard University Press)

Dennett, Daniel C. (1971) “Intentional Systems” *The Journal of Philosophy* 68(4):87-106

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015) “Algorithm aversion: People erroneously avoid algorithms after seeing them err.” *Journal of Experimental Psychology: General*, 144(1): 114–126.

Dressel, Julia and Farid, Hany (2018) “The accuracy, fairness, and limits of predicting recidivism” *Science Advances* 4(1):eaao5880. DOI: 10.1126/sciadv.aao5580

Dutta, S., Wei, D., Yueksel, H., Chen, P-Y., Liu, S., and Varshney, K.R. (2020) “Is there a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing” *Proceedings of the 37 th International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020 (available at https://proceedings.icml.cc/static/paper_files/icml/2020/2831-Paper.pdf)

European Commission High-Level Expert Group on Artificial Intelligence (2019) *Ethics Guidelines for Trustworthy AI*. (Brussels: European Commission) Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Elliott, Kevin, C. (2017) *A Tapestry of Values: An Introduction to Values in Science* (Oxford University Press)

Floridi, Luciano, Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena. E. (2018) "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations" *Minds and Machines* (28): 689–707

Fricker, Miranda. (2007) *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford University Press)

Haugeland, John (1985) *Artificial Intelligence: The Very Idea*. (Cambridge, MA: MIT Press.)

Hempel, Carl (1965) "The Function of General Laws in History" in his *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press)

Hirsch, T., Mercen, K., Narayanan, S., Imel, Z. E., and Atkins, D. C. (2017) "Designing Contestability: Interaction Design, Machine Learning, and Mental Health" *DIS '17: Proceedings of the 2017 Conference on Designing Interactive Systems* pp.95–99 https://doi.org/10.1145/3064663.3064703

Kaas, M (Forthcoming, 2020) "Raising Ethical Machines: Bottom-Up Methods for Implementing Machine Ethics" in S.J. Thompson (Ed.) *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence* (IGI Global Press)

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016) "Inherent Trade-Offs in the Fair Determination of Risk Scores" Available at https://arxiv.org/pdf/1609.05807.pdf

Lawton, Graham (2019) "Simulating the World" *New Scientist*, 5 October 2019, 3250:38-41 (available at https://www.newscientist.com/article/mg24332500-800-predicting-the-future-is-now-possible-with-powerful-new-ai-simulations/

Logg, Jennifer M., Minson, Julia A. and Moore, Don A. (2019) "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment" *Organizational Behaviour and Human Decision Processes* 151:90-103

Marcus, Gary and Davis, Ernest (2019) *Rebooting AI: Building Artificial Intelligence We Can Trust* (New York: Pantheon)

Mulligan, D.K., Kluttz, D.N., and Kohli, N. (2019) "Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions." Available at SSRN 3311894

Robbins, Scott (2019) "A Misdirected Principle with a Catch: Explicability for AI" *Minds and Machines* 29(4):495-514

Smith, Brian Cantwell (2019) *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press)

Sutton, R.S., and Barto, A.G. (2018) *Reinforcement Learning: An Introduction* (2nd Edition) (Cambridge, MA: MIT Press)

Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2018) "Transparency in algorithmic and human decision-making: is there a double standard?" *Philosophy & Technology* 32:661-683