Andrada, G., Clowes, R. W., & Smart, P. R. (in press) Varieties of transparency: Exploring agency within AI systems. *AI & Society*. [https://doi.org/10.1007/s00146-021-01326-6]

Varieties of Transparency: Exploring Agency within AI Systems

Gloria Andrada,[1] Robert W. Clowes,[1] and Paul R. Smart[2]

[1] Instituto de Filosofia da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisbon, Portugal
[2] Electronics and Computer Science, University of Southampton, Southampton SO17 1DG, UK

Abstract
AI systems play an increasingly important role in shaping and regulating the lives of millions of human beings across the world. Calls for greater *transparency* from such systems have been widespread. However, there is considerable ambiguity concerning what "transparency" actually means, and therefore, what greater transparency might entail. While, according to some debates, transparency requires *seeing through* the artefact or device, widespread calls for transparency imply *seeing into* different aspects of AI systems. These two notions are in apparent tension with each other, and they are present in two lively but largely disconnected debates. In this paper, we aim to further analyse what these calls for transparency entail, and in so doing, clarify the sorts of transparency that we should want from AI systems. We do so by offering a taxonomy that classifies different notions of transparency. After a careful exploration of the different varieties of transparency, we show how this taxonomy can help us to navigate various domains of human–technology interactions, and more usefully discuss the relationship between technological transparency and human agency. We conclude by arguing that all of these different notions of transparency should be taken into account when designing more ethically adequate AI systems.

Keywords: Transparency; AI ethics; Agency; Philosophy of technology; Philosophy of mind; Philosophy of AI

1 The Problem of Transparency

AI systems now play central roles in structuring and regulating the lives of millions of human beings across the planet. There are a vast range of ways in which AI intervenes in our individual and collective lives. Reliance on AI also occurs at collective and societal levels, with AI systems playing increasingly central roles in regulating all manner of social processes, from the functioning of stock exchanges,[1] to traffic systems,[2] to managing the everyday life-chances of millions of individuals, affecting issues such as who gets a loan, who gets a job, and even who is sent to prison and who might get parole.[3] AI systems also play a role in extending and supporting a vast range of individual human cognitions: from how we find new candidates for life-saving drugs,[4] to how we navigate cities in cars or on foot, to how we find emotional support online.[5] Individuals also use AI technologies to regulate themselves and their activities, using self-tracking devices to monitor their health, their mood and their daily exercise,

[1] See Ferreira et al. (2021).
[2] Wang (2008).
[3] See Bernard Marr, *The Revolutionary Way of Using Artificial Intelligence in Hedge Funds*: https://www.forbes.com/sites/bernardmarr/2019/02/15/the-revolutionary-way-of-using-artificial-intelligence-in-hedge-funds-the-case-of-aidyia/#17eb640157ca (last Accessed: 13 June 2021). Also, in extreme cases, such systems may be used for the algorithmic regulation of society (Cristianini and Scantamburlo 2020).
[4] See: https://www.science.org/news/2020/08/ai-invents-new-recipes-potential-covid-19-drugs
[5] See: https://replika.ai/

or even to regulate their relationships with others.[6] Individuals also rely on search engines to gather the information they require in their daily activities, including recommendations from streaming services, to decide which music they are going to listen to, or which films they are going to watch, or even which partner they might date and eventually marry.

For all the potential of AI technology to improve human life in myriad ways, there is also a growing atmosphere of concern about the risks and dangers that such systems pose as they become ever more embedded in the everyday operations of human societies and individual lives. As AI systems become increasingly involved in public decision-making, "taking policy decisions or authoritative decisions regarding the rights and burdens of individual citizens" (de Fine Licht and de Fine Licht 2020, p. 917), more attention is being devoted to systems that may otherwise do their work according to purposes and mechanisms which are opaque.

Society and its institutions do recognize, and to some extent are trying to act on, the problems that arise in relation to AI technologies' deep penetration into the human social world and individual lives. One area of great anxiety, which has led to a search for a response, concerns the question of AI *transparency*. For instance, in the 2019 guidelines presented by the AI HLEG (the European Commission's High-Level Expert Group on Artificial Intelligence) *transparency* is identified as one of seven key requirements for AI technology.

As AI systems play increasingly central roles in our individual and collective activities, it is natural to see demands for greater visibility of the constituents of the underlying processes.[7] This is due to the fact that their opacity engenders a series of ethical and political problems concerning the nature of public and private decision making, which may be obscured (intentionally or otherwise) by the use of AI systems.[8] Where such systems affect important changes in the lives and opportunities of individuals and societies, it seems at least *prima facie* preferable for us to gain knowledge about these processes.[9] And it is precisely their apparent de facto opacity that has led to widespread calls for greater transparency in AI systems. This is acknowledged in the AI HLEG report, where the authors note:

> [D]ata, system and AI business models should be transparent: Traceability mechanisms can help [in] achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations. (AI HLEG 2019)

While these aims are no doubt worthy, it should be noted that no further definition of what "transparency" might be is given in the AI HLEG document. The problem here is that without specifying more clearly what is intended, it is hard to see how helpful these injunctions can be.

However, we will see throughout this paper that "transparency" in AI systems can be taken to mean different things, and that several concepts of transparency play important roles in clarifying and explaining different aspects of human interactions with AI. As we will show, there is a pressing need

[6] See Lupton (2016).

[7] See Floridi et al. (2018).

[8] For important discussions of these themes, however, see Coeckelbergh (2020), de Fine Licht and de Fine Licht (2020) and Walmsley (2020).

[9] For an overview, see Müller (2020). See also de Fine Licht and de Fine Licht (2020).

to clarify and explain how these different varieties of transparency interact, and especially how they shape, constrain and enable various forms of human action.

To get an initial handle on this, we should note that, on the one hand, "transparency" in AI systems can be taken to refer to a variety of related conceptual targets, including the algorithmic transparency of AI systems, the openness to discussion of AI-influenced decisions, and the challengability or explicability of the same. Although these all point to different aspects of AI systems, broadly, they all refer to the visibility or penetrability of the underlying processes in AI systems which determine or shape subsequent decisions. This group of notions has played a major role in AI and informational ethics, and are usually in the foreground of demands for greater transparency in AI systems.[10] Roughly, the problem is that algorithms, decision-making systems and neural nets are implicated in making decisions with respect to which individuals or groups would like greater oversight, accountability, ability to intervene, or simply understanding. We will refer to this general class of issues as issues of *reflective transparency*.

However, "transparency" in AI systems can, interestingly, also be taken to refer to a *prima facie* unrelated conceptual target that arises from a different tradition. This *other* form of transparency appears in the philosophies of mind and cognitive science, and also in the phenomenologically influenced tradition of philosophy of technology.[11] Here transparency relates to the experience of human–technology interaction, and more recently, to the *experience* of acting with AI-inflected technologies. "Transparency" in this respect refers to how an artefact or technology appears, or crucially disappears, in the user's experience. It is usually argued that for many forms of artefactually mediated actions and interaction, the technology must be, to some extent, transparent to the user, who should not pay attention to the interface, but rather to the task being performed. We will refer to such transparency as *transparency-in-use*.

While reflective transparency relates to our abilities to *see into* mechanisms and underlying processes to control them better, transparency-in-use relates to the experience of *seeing through* a technology to competently and fluently act while using it. There seems to be a tension between these two notions of transparency. The former requires some degree of conscious access to certain features of the AI system, while the latter requires our conscious thought and attention to drift further away from the system itself. Yet, as we will show, these two sorts of transparency are intimately related and entwined with central aspects of human agency and our ability to act with artefacts, especially to act with and control AI-mediated technologies. Understanding how these two notions fit together indicates some new ways forward on the question of building more ethically adequate AI systems. Yet, at present, these two notions of transparency are largely disconnected from each other in scholarly debate, and the gap between them has not been bridged. This is what we may call the *problem of transparency*.[12] Making progress on this problem is the main goal of this paper.

[10] In their influential work, Turilli and Floridi write: "In the disciplines of computer science and IT studies, however, 'transparency' is more likely to refer to a condition of information invisibility, such as when an application or computational process is said to be transparent to the user" (Turilli and Floridi 2009, p. 105).

[11] More on this in Sect. 3.

[12] Wheeler (2019) identifies the problem in the following quote: "Sometimes, technology is described as being transparent when a specified class of users is able to understand precisely how it functions. This is a perfectly reasonable notion of transparency, but note that a device which is transparent in this sense may be broken or malfunctioning, and so will not be transparent in the phenomenological sense, and that a device which is

To this end, this paper analyses the two main varieties of transparency,[13] not only to offer a conceptual reconciliation of the problems, terminologies and debates in which they are embedded, but also to show their vital importance to our understanding of how AI technology interrelates with human action. This is due to the fact that notions of transparency not only feature in two apparently separate debates about human technological interaction, but also illuminate different aspects of human agency.

Our plan is as follows. In the next section (Sect. 2), we discuss what we identify as *reflective transparency* and its various subtypes, which we call *information*, *material* and *transformational transparency*. This is the sort of transparency discussed in the ethics of AI. In the following section (Sect. 3), we discuss what we call *transparency-in-use*. This is the concept of transparency that is much used in phenomenology, philosophy of technology and philosophy of cognitive science, and it is used to conceptualize how many humans come to "act through" certain tools and artefacts. Next (Sect. 4), we introduce some relevant distinctions concerning the nature of transparency as a relational property; and we then proceed to show (Sect. 5) how these two different types of transparency can be viewed as parallel properties of human–technology interactions that *inter alia* constrain and may (perhaps surprisingly) enable some aspects of human agency. Since this aspect of the transparency debate has been so little explored, we offer a diagnosis of which aspects of AI systems may inhibit or enable different aspects of human agency. Finally (Sect. 6), we draw on these considerations to make some recommendations for more adequate and ethically sound AI systems focused on enabling human action.

2 Reflective Transparency in AI Technologies

As we just saw, AI systems play increasingly central roles in our societies. That is why it is natural to see demands for greater visibility of their constituents and effects. Demands for greater transparency are often made in relation to a series of related categories including information transparency, algorithmic transparency and data transparency, and while it is not clear that such demands always have the same target, there is an underlying sense that the workings, contributions to decisions and potential biases of AI systems need to be opened up to further analysis. Our discussion attempts to subsume many of these categories under the single heading of *reflective transparency*. This is due to the fact that what is being requested is insight *into* various aspects of a given mechanism of autonomous or artificial decision making and making some of its details or constituents open to further deliberation.

Let us expand on this. At a superficial level, the terms "transparency" and "opacity" are used in connection with AI technologies much as they are used with respect to other hidden-away processes such as covert political discussions. For example, we might read that a journalist "requests greater transparency in the activities of the select committee", when what the journalist is requesting is more

phenomenologically transparent-in-use may be impenetrable in its inner workings, and so will not be transparent in the 'open to understanding' sense. Therefore, there is a double dissociation between the two concepts" (Wheeler 2019, p. 859).

[13] Walmsley (2020) classifies different notions of transparency and distinguishes between "outward" transparency that targets various epistemic and ethical features of AI systems and functional transparency. We do not have sufficient space to address the difference between Walmsley's taxonomy and ours. However, we wish to highlight that these different varieties of transparency correspond to different forms of what we call reflective transparency. We thank an anonymous reviewer for bringing this to our attention.

insight into a given decision-making process. When we hear calls for transparency in corporate or governmental processes, these seem to concern the acquisition of greater knowledge of, or insight into, the reasons and processes that lead to decisions, or to policies being determined.[14] Transparency in AI technologies operates in a similar way. AI systems play ever greater roles in regulating our lives, often following hidden or inaccessible procedures. Calls for transparency in AI systems often require *seeing into* various aspects of equipment and the mechanisms of the device with which human beings interact, thereby providing greater scope to consciously understand or deliberate on the constitution or effects of those systems.[15] For theoretical purposes, we call this *reflective transparency*, and we differentiate three ways or types of *seeing into* the constitution and effects of a given AI system. We call these: information transparency, transformational transparency, and material transparency.

2.1 Information Transparency

Information transparency is the sort of transparency that has received the most attention in the ethics of AI and adjacent discussions (Diakopoulos 2020; Turilli and Floridi 2009; Weller 2019). In general, the term "information transparency" refers to the disclosure of information about an AI system, typically for the purpose of supporting judgements concerning the system's fairness, trustworthiness, safety, efficacy, accountability, and compliance with regulatory and legislative frameworks. Information transparency is thus deemed to be important because it supports the formation of cognitions (e.g., judgements, decisions, beliefs) pertaining to a system's suitability for use within a given social context.

Among the sorts of information targeted by the notion of information transparency is information about the nature of the algorithms that are used by a given AI system, as well as the kinds of data that were used in a training regimen. In general, however, there are no hard-and-fast rules as regards the kind of information that needs to be disclosed in order to satisfy demands for information transparency. The kind of information that needs to be disclosed will vary according to the kind of system-related evaluation that needs to be made. In addition, such evaluations are often performed by multiple stakeholder communities (e.g., developers, deployers, users, policy-makers, and so on), and each of these communities may require access to different bodies of information, even when such information is used for similar evaluative processes. When it comes to assessments of system trustworthiness, for example, the users of an AI system may require access to bodies of information that differ from those demanded by regulatory authorities.

Some forms of informational opacity may be the result of using difficult-to-analyse technologies such as neural nets, genetic algorithms or other machine learning or optimization techniques, where even the designers of the systems are unsure of the precise means by which certain system outputs are produced (see Zednik 2021). This does not mean that the designers or owners of such systems should necessarily be "let off the hook" when it comes to disclosing explanatorily relevant information; nevertheless, acquiring such information can present developers with a formidable technical challenge.

[14] See de Fine Licht and de Fine Licht (2020).

[15] This is sometimes referred to as "opening the black box". See Zednik (2021).

2.2 Material Transparency

The second subtype of reflective transparency is material transparency. This form of transparency is important, because the digital character of AI systems might sometimes lead us to forget their materiality.

We can fairly straightforwardly develop a contrast between the informational, or processing, or algorithmic parts of a system that are reflectively opaque to a user, and those other aspects of the system which, while still opaque, are not directly dependent on the algorithms, information processing system or data captured, but rather on their material realization. In this respect, reflective transparency, conceived as *seeing into*, would involve shedding light on or revealing such important material aspects of AI systems, as it does with their informational aspects. This includes the materiality of the hardware that realizes a given informational system, its production or maintenance requirements and the labour relations that accompany this. For instance, it has been argued that training some AI systems can emit as much carbon as five cars in their lifetime, but this is definitely not common knowledge.[16] Moreover, insight into the materiality of AI systems brings to light lots of factors that information transparency leaves out, such as energy use, as well as factors in relation to which AI technologies might be opaque (e.g. their manufacturing process, the division of labour, their ecological impacts, and so on).[17] These factors exist, and they are important aspects of opacity that are often in play when there are demands for greater transparency. Some of these may strictly have nothing to do with information, or hold only very indirect relations, but nevertheless we should take such claims seriously, given that it is reasonable to ask how much energy is expended on a given search; or how much energy is used to store data for a transaction; or what the implications of mining a resource are on the environment, or on the political conditions of a given country. All these are relevant issues when thinking about the ethics of AI.

2.3 Transformational Transparency

A third subtype of reflective transparency is transformational transparency, i.e., transparency concerning the neural and bodily transformations that pervasive use of AI technology elicits. In this respect, AI technologies are acknowledged to be opaque to the user in connection with the deeply transformative effect that our interaction with them has on our cognitive capacities.

Transformational opacity happens where there are hidden changes to the agent (the person) who uses an AI artefact that have important effects on the agent's cognitive, epistemic or ethical capacities. For instance, pervasive GPS use can have transformative effects on one's wayfinding abilities and techniques (Gillett and Heersmink 2019). These effects can be both positive and negative, but lay users do not have access to such data. Such effects could be very diverse, including broadly "capability echo"-type effects, where a user feels the absence of the system when not using it, perhaps as a loss of capability.[18] They could also involve things like memory dependencies, and potentially affect the

[16] See Karen Hao's work on the carbon footprint of deep learning: https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/ (last accessed: 3 December, 2020).

[17] See, for instance, the report that Amnesty International and Afre-watch have presented on child labour and cobalt mines: https://www.amnesty.org/en/latest/news/2016/01/child-labour-behind-smart-phone-and-electric-car-batteries/ (last accessed: 2 June, 2021).

[18] See Smart et al. (2017), pp. 77–78.

agent's employment prospects, cognitive abilities, or even her sense of self or personhood.[19] Here the point is that we come to rely on such systems, and companies may even deliberately entice us to rely on systems that change our cognitive (or ethical or epistemic) capacities.

Summing up, reflective transparency can point to the (in) visibility of certain informational or material aspects of AI systems, as well as to the (in)visibility of the effects that pervasive reliance on AI systems elicits in their users. As we will see later, the notion of reflective transparency (in its three subtypes) can help us refine what we require from transparency in AI systems, but before looking at this in more depth, we will next explore a different conception of transparency.

3 Transparency-In-Use in AI Technologies

As mentioned in the introduction, some notion of transparency is also much used in phenomenology, and in the philosophies of technology and cognitive science, where it is used to conceptualize how humans come to act through certain tools and artefacts.

What we will call *transparency-in-use* is the standard form of transparency that features in embodied approaches to cognition, and also in philosophical accounts that inherit concepts from the phenomenological tradition.[20] Transparency here evokes the idea that the agent *sees or acts through* the equipment, typically in cases where an artefact is being used to some end. Such phenomenological transparency, or transparency-in-use, relates to the *mastery* of the use of a particular artefact, or range of artefacts, by a skilful agent. When devices are somewhat transparent to the skilful agent, the agent's attention and conscious thought does not stop at the interface, but rather at the task in hand. This happens only when a degree of skill, or indeed mastery, is achieved. In other words, a technological device becomes transparent to the user once they have become sufficiently skilful in their interaction with it. This is because, once a certain degree of mastery is achieved, the user does not need to constantly stop and reflect on how to use it correctly. Instead their attentional resources are directed to the task at hand, and not to the technology they are interacting with.[21]

Before going any further, let us briefly indicate how transparency-in-use might be achieved. Several studies show that, following a certain amount of practice with a tool or technology, the representation of personal space in the brain is modified. This has been interpreted as the tool's incorporation into the *body schema*, that is, into what has been defined as the neural representation of the body's shape and posture (Gallagher 2005, p. 24). The central idea here is that when a tool is incorporated into the body schema, the brain represents it as part of the body.[22] And, roughly, the phenomenological indicator of this process of incorporation is the tool's phenomenological transparency.[23]

[19] Clowes (2020).

[20] Classic examples of this sort of transparency in the literature include Heidegger (1927) and Maurice Merleau-Ponty (1945). For more contemporary approaches, see Clark (2008), Heersmink (2015), Wheeler (2019) and Andrada (2020).

[21] It has usually been argued that transparency entails a lack of conscious thought or reflection for the artefact's proficient use. See Heersmink (2015), Andrada (2020).

[22] This is empirically supported by experiments, such as those performed by Maravita and Iriki (2004).

[23] Transparency has played an important role in the hypothesis of the extended mind, where some form of transparency-in-use is an indicator of mental or cognitive extension (Clark and Chalmers 1998; Clark 2008; Andrada 2020). Here we do not want to enter into the debate concerning the plausibility or otherwise of the extended mind thesis. Nevertheless, as will become clear, we do consider a certain degree of transparency-in-use to be central to a successful human–technology interaction.

This transparency-in-use has been applied to a wide variety of technologies, from rudimentary hammers (Heidegger 1927), to notebooks (Clark and Chalmers 1998), or to internet technologies (Clowes 2015; Wheeler 2019; Heersmink and Sutton 2020). However, although transparency-in-use has been discussed in the context of several twenty-first century technologies, its specific relationship to AI technologies has not been the subject of detailed scrutiny.

AI technologies are informational technologies, and achieving a certain degree of transparency-in-use might require the development of different types of skills in the user. First, when we are interacting with an informational device, a successful transparency-in-use entails developing skills such as easily interpreting the information that the device conveys. Second, the user needs to develop different procedural skills such that they can interact with the technology somewhat fluently. All this contributes to acquiring a transparency-in-use, which will depend on the particularities of the technology and the user's skills.

To get a better grip on the intricacies of how this notion of transparency relates to AI systems, it will be useful to employ Richard Heersmink's (2013) distinction between *procedural* and *representational transparency*: a distinction that Heersmink introduces in his taxonomy of dimensions of integration of informational technologies into an agent's cognitive system.[24] Heersmink (2013) refers to representational transparency as the ease with which a user can interpret the representational system or arrangement of an informational device. This concerns being able to read some sort of abstract pattern or interpret a representational system. Procedural transparency, on the other hand, concerns how effectively, fluidly and skilfully a device can be deployed in action. Both of these sorts of transparency are forms of transparency-in-use.[25]

It is worth noting at this point that transparency-in-use is taken to be a goal in design,[26] given that a good design might facilitate a fluent, easy interaction with a given technology. However, it is precisely this phenomenological transparency of certain technologies that has raised ethical worries. This is due to the fact that most AI technologies evidence high levels of what Clowes (2015) calls *practical incorporability*, i.e., the capacity to be easily and seamlessly incorporated in human action, while at the same time being highly reflectively opaque. In this respect, it has been argued that the "smart" technologies that we learn to treat as transparent might bias our cognitive processes in hidden ways,[27] or that our heavy reliance on them might jeopardize our selfhood,[28] or our intellectual autonomy.[29]

It is here that we reach a point of conflict between transparency-in-use and reflective transparency, as some degree of (individual or group-level) conscious permeation seems to be important for addressing some ethical concerns. The problem is that transparency-in-use is really a sort of *seeing through* of artefacts and technology, and not the *seeing-into* that we find in the cases described in the previous section. Their differences, however, should not make us think that we have to choose

[24] Heersmink (2015) refers to representational transparency as "informational transparency". In order to avoid confusion with our notion of information transparency, we have chosen to speak of "representational transparency".
[25] Note that this distinction holds for informational devices, but not for all cases of transparency-in-use.
[26] See for instance https://calmtech.com/ (last accessed: 2 August, 2021).
[27] See Wheeler (2019), for an account of the risks that transparency entails in some of our cognitive processes. See also Andrada (2021), for more on the connection between transparency and an agent's epistemic standing.
[28] See Clowes (2020).
[29] See Carter (2020).

between one and the other. In the following sections, we will see how both types of transparency are relevant and should be considered when discussing the desired properties of AI-involving systems.

4 Degrees of Transparency (In Practice)

Given our foregoing discussion, it may seem that the two broad types of transparency that we have discussed, i.e., transparency-in-use and reflective transparency, have little to do with each other. However, as we will now proceed to show, this is a mistake, as both are relevant for building more ethically adequate AI systems. To show this, we will focus on their relevance to building systems that promote, rather than undermine, human agency. But to see this, we must step back for a moment and consider some key features of transparency, before exploring agency within AI systems.

According to our account, transparency (both transparency-in-use and reflective transparency) is a relational category connecting user and technology. When an individual (or group of individuals) are interacting with a given AI system, the technology might be more or less transparent-in-use, and they might be more or less able to see into or reflect on some of its mechanisms, its processing, or its effects. This suggests that transparency is always relative to an individual or group of individuals. Depending on the properties of the technology and the skills acquired by an individual or group of individuals, a given technology might be more or less transparent. In this respect, it is worth noting that although AI systems might be fully transparent in principle, they will not be fully transparent in practice.[30] Think for instance of machine learning techniques. In most cases, not even the designers can fully understand the underlying processes. But this does not mean that they are not transparent in principle. Let us briefly explore how this affects the different varieties of transparency.

First of all, one might claim that an AI system is in principle fully reflectively transparent, which means that its mechanisms can be made visible and explained, and that its effects in a user could in principle be rendered intelligible (even if only to a mind superior to the human mind). However, what matters when discussing AI systems and their ethics is their transparency in practice. For instance, there are, in practice, questions of material transparency, e.g., concerning whether a company has a certain carbon footprint that they choose not to make available. Or there could be certain types of information transparency—such as what the weights of a given neural network mean, or how it carries out a computation—which might not be transparent to any individual, despite the fact that they are transparent in principle. What matters here is that the system is not equally opaque to all individuals, despite the fact that in principle a given system should be reflectively transparent. Transparency-in-use, on the other hand, is always transparent in practice; it is always relative to the skills of a user or group of users, that is, it is always agent-centred.

We may thus conclude that transparency in practice comes in various degrees. However, we now face the following challenge: how can we determine what type and degree of transparency is correct? As we will see, this can, and should, be determined against different backdrops, and agency is one illuminating possibility. Let us now proceed to explore agency within AI systems.

5 Exploring Agency within AI Systems

Reflective transparency and transparency-in-use both have an important relationship to the exercise of human agency in the context of the use of technologies, albeit concerning contrasting aspects of

[30] Thanks to an anonymous reviewer for bringing this to our attention.

human agency. Before continuing, let us clarify that we take "agency" to denote the exercise or manifestation of one's capacity to take actions, or "do things".

Reflective transparency (i.e., seeing into different aspects of the equipment and the mechanisms of the AI system with which human beings interact) relates to those aspects of human agency by which we are able to observe aspects of the workings of a given artefact, app or software system. One major reason to strive for the various forms of reflective transparency of AI systems is that they potentially offer us greater control over such systems, either at the individual or the group level. By gaining access to the mechanisms that produce the outcomes of AI systems, we have the potential to understand, challenge or indeed redesign such systems so that we can direct them to more desirable, conscious, ethical, or politically adequate ends. It is (arguably) only by some degree of "seeing into" the constituents of AI systems, and by being able to see the possibility of changing those constituents, that we can take control of AI systems.[31] Typically, such openness to cognition allows us to at least partially understand, deliberate on and thus (ideally) gain control of some process which restricts or shapes our ability to act. This, we believe, is uncontroversial, however, and crucially, a certain degree of reflective transparency will not always be enough. In some circumstances, simply being able to act through (transparent-in-use) technology is enough to carry out our purposes, which means that such reflective transparency will not always be needed or useful. This is why applying the notion of transparency-in-use to debates concerning AI ethics proves to be extremely useful.

Transparency-in-use highlights an agent's capacity to mediate some aspect of their action towards a given goal in a controlled, fluid and skilful manner. Transparency-in-use is required of technologies that mediate fluent human action, perhaps especially of the sort found in skilled mastery (Dreyfus and Dreyfus 1980). But there is no particular reason that technologies that incorporate or rely on AI systems should be any different, at least when we utilize them in skilful activities. Think, for instance, of the use of technologies such as CAD (computer aided design), or even word processing software. To write a piece of text, for example, it is important to not always be struggling with the interface. There is a need to project one's thought into what one is writing. An unresponsive, obtrusive or simply unfamiliar interface is often enough to impede work from being done, or a thought from being carried through to its conclusion. That is why some degree of transparency-in-use is necessary for an effective interaction with a technology.[32] It is true that for an effective action, the agent does not need to be unaware of *all* aspects of the technology, but the agent must be somewhat skilful, and this already entails some degree of transparency-in-use. In other words, being able to focus on the task at hand while using or relying on a given technology, however simple or complex it might be, is necessary for an effective interaction with an AI technology.[33] This reveals that transparency-in-use should be taken into account when thinking about AI ethics, given its connection with human agency.

We now reach the point where what is at stake is determining what the appropriate type of transparency is, in a given circumstance, and to do so, we need to dig deeper into the structure of

[31] This relates to the so-called *control problem* (Bostrom 2014; Russell 2019), which can be viewed as a problem of (collective) human agency gaining control over (in this case) a super-intelligent AI, in order to avoid an existential threat.

[32] Thanks to an anonymous reviewer for encouraging us to develop this point further.

[33] The link between transparency-in-use and agency also highlights the importance of accessible and inclusive technologies. See Andrada (2020), for more on the relationship between phenomenological transparency, technologies and diverse embodiments.

human agency. According to a widely accepted distinction, there are basic and strong forms of agency (Bratman 2000). Basic agency entails the capacity to act, while strong agency can be defined in terms of an agent's capacity for reflection, self-evaluation, self-regulation and organization with respect to projects over time (Bratman 2000).[34]

When thinking about interaction with technologies, simple agency entails the capacity to perform, while strong agency, by contrast, implies the ability to reflect on and reconstruct one's relationship with the technology. In the case of smart artefacts, i.e., artefacts that embody or incorporate elements of AI technology, transparency-in-use concerns the aspect of such artefacts that allows us to just—apparently thoughtlessly—use it towards some goal or purpose, or in the service of some craft or activity, whereas reflective transparency concerns the extent to which we can see into how the technology works, or how it affects our perceptual and cognitive abilities. The key point here is that, depending on different factors, we might want to enhance strong agency, thus requiring higher degrees of reflective transparency, or we might want higher degrees of transparency-in-use, to promote basic agency.

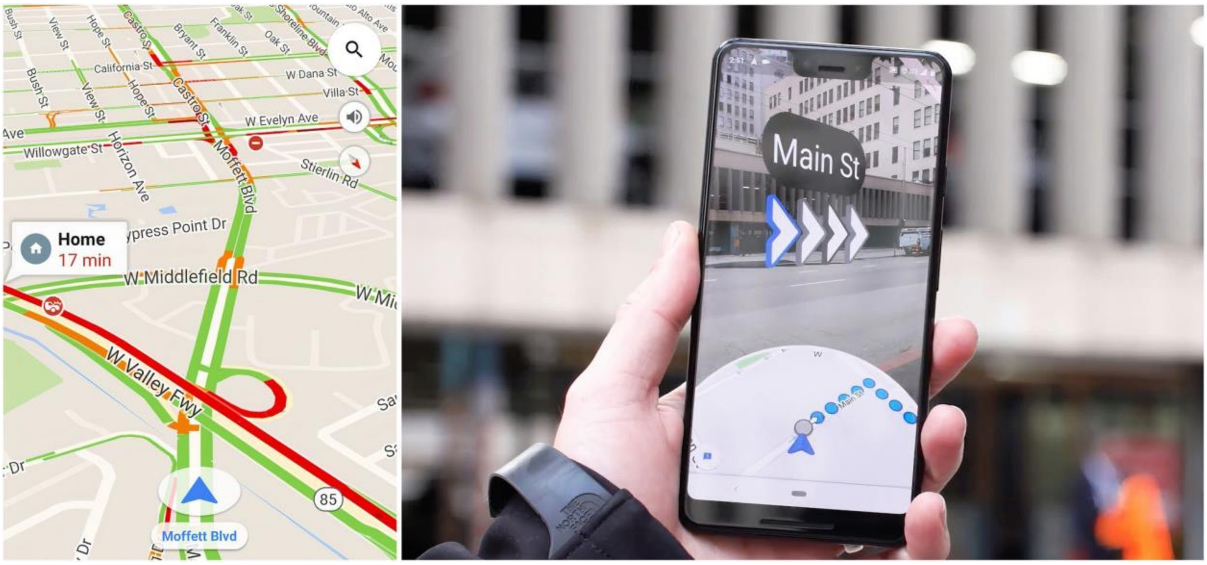

Figure 1. The Google Maps wayfaring system. The left image shows Google Maps in what is called "Driving Mode". The right image shows Google Maps in what is called "Augmented Reality Mode"

An example might be helpful here, and we will develop one in terms of the use of a wayfinding system such as Google Maps (see Figure 1). A good wayfinding system— indeed, any adequate one—is designed to be transparent-in-use at least some of the time, for instance when we are driving. A wayfinding system that obtrusively alerts the user to, e.g., changes in road conditions, or route changes, would likely be highly dangerous. A transparent-in-use software system, by contrast, can extend our abilities to act, for instance to drive on a road, when we effectively know where we want to go but not how to get there. As we emphasized before, using such systems requires substantial skills and adaptation to the software system.

[34] We wish to warn the reader that we are not saying that this distinction is correct. In fact, there might be good reasons to think that, even if we can make such a distinction for certain theoretical purposes, the relationship between such forms of agency would be much more dynamic and intertwined. Nevertheless, this distinction may help to clarify our proposed analysis.

However, wayfinding systems might also sometimes require us to know more about how the software itself works, so that we can configure it better, or indeed shape it to our own needs. Sometimes we might need to know why an algorithm is sending us via a certain route, or change how the algorithm allocates the route that is shown to us, or just understand how well an algorithm is performing. Being able to tailor such a system to our particular needs is a way of fine-tuning our abilities to perform actions and can also be viewed as a way of increasing our control over the system. Systems that allow such abilities to supervise, intervene and customize the informational resources that decide the system's actions can be said to be highly reflectively transparent.

At this point, it might be a good idea to clarify how algorithmic systems, including AI systems, can variously implement the properties of transparency-in-use and reflective transparency. A wayfinding system that exhibits enough transparency-in-use to promote safe and effective driving while on the road, may also deliver tools that allow sophisticated reflective transparency at times when the presentation of such information does not pose a danger to the driver (e.g., when planning a route). But it is also important to point out that not all systems are available to be regulated or customized in the same way. Facebook's Edgerank algorithm (see Bucher 2012) allows high degrees of transparency-in-use interactions with our friends and acquaintances (see Figure 2), to the point of practical invisibility. But, at the same time, Edgerank has very limited reflective transparency. That is, the mechanisms that determine which posts we see are hidden and allow us very limited ability to customize them.

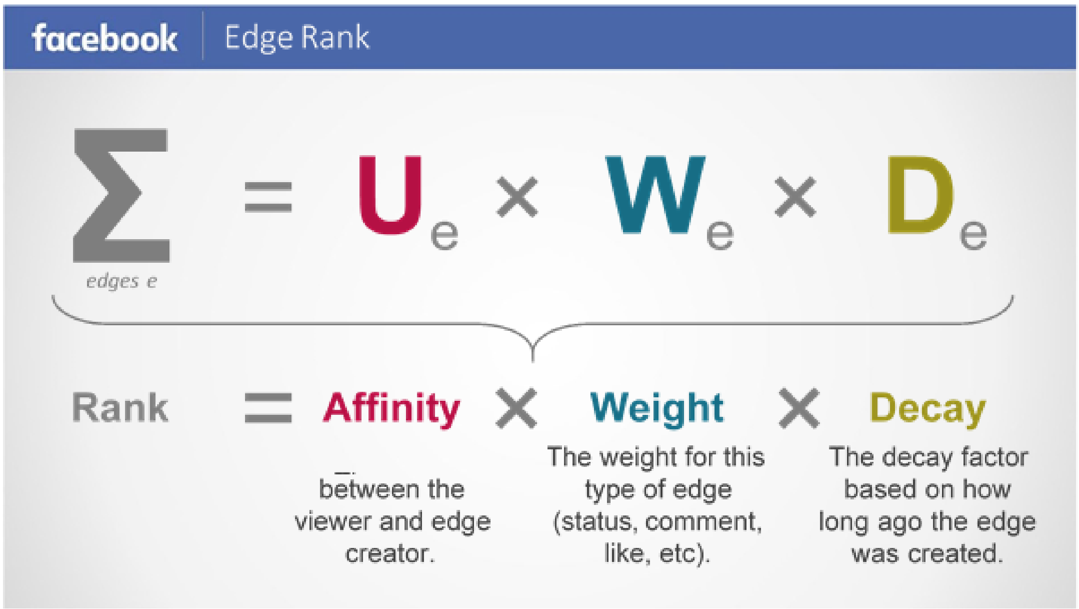
facebook
Edge Rank
$\sum_{\text{edges } e} = U_e \times W_e \times D_e$
Rank = Affinity × Weight × Decay
between the viewer and edge creator.
The weight for this type of edge (status, comment, like, etc).
The decay factor based on how long ago the edge was created.

Figure 2. Facebook's Edgerank algorithm.

We can view the interactions of the two properties as a sort of grid where an artefact–user relationship can be in one of four positions.[35] The interactions between transparency-in-use and reflective transparency are complex, but can be schematized as in the following table (Table 1).

[35] It is important to point out that the situation would be different if we wanted to determine the correct degree and type of transparency required for enhancing trust and trustworthiness in AI systems. From a practical perspective, theorists such as O'Neill (2020) have questioned the extent to which transparency is effective in supporting assessments of trustworthiness, and there are additional reasons to doubt the assumed relationship

	Basic Agency	**Strong Agency**
Transparency-in-use	Required for fluid interaction with artefacts and technology, and our abilities to "act through" them.	Can form the basis of the regulation of technology but can in some circumstances interfere with the reflective use of technology.
Reflective transparency	Too much, or the wrong sort of reflective transparency at the wrong time can inhibit basic agency.	Required for the ability to deliberate on and reshape our interactions with algorithmic technologies.

Table 1. Interactions between transparency-in-use and reflective transparency.

Technologies can exhibit a degree of both transparency-in-use and reflective transparency (albeit probably not high degrees of both simultaneously). They can also exhibit high degrees of transparency-in-use while exhibiting very low degrees of reflective transparency. For instance, Edgerank, from the vantagepoint of the average user of Facebook, might be a technology like this. Another possibility is that they might exhibit low degrees of transparency-in-use but high degrees of reflective transparency—some industrial decision support systems are likely of this type. Or they might exhibit low levels of both types of transparency. Systems that are both opaque-in-use and reflectively opaque may still be able to perform useful functions, but they are likely to be difficult to use and control, and especially difficult to understand or control when they misbehave. For reasons we will now go on to elaborate, they may also inhabit a dangerous ethical space.

6 Recommendations for Ethically Adequate AI Systems

We began this paper by noting that "transparency" refers, on the one hand, to different aspects regarding the visibility of underlying processes in AI systems (what we have called reflective transparency), and on the other, to the user's experience of interacting with AI systems (what we have called transparency-in-use). Importantly, we noted that there seems to be a tension between them: the former requires some degree of conscious access to certain features of the AI system, while the latter requires our conscious thought and attention to drift further away from the system. This is what we called "the problem of transparency". In the previous sections, we showed that both transparency-in-use and reflective transparency, and indeed their various sub-types, are important conceptual tools for clarifying and understanding various dimensions of our relationship to AI technology; and we have addressed their importance to human agency. We now want to conclude by briefly exploring how these factors have significant but generally unappreciated implications for the design and use of technologies that rely on AI.

Many AI systems, especially those that are highly personalized and targeted at individual users, are often highly transparent-in-use, and designed to be so, yet highly opaque with respect to reflective

between transparency and trust; that is, the idea that (reflective) transparency always plays a positive role in cultivating trust or supporting assessments of system trustworthiness (see also Nguyen 2021). On the other hand, users trust AI systems for engaging in various actions. They do not want to constantly check on well-functioning equipment, because that impedes their ability to act with it. That is why some degree of transparency-in-use seems to be necessary for trustworthiness. The crucial thing to bear in mind here is that the adequate type and degree of transparency required for promoting trust and trustworthiness in AI systems might turn out to be different from the level of transparency required for promoting agency. We hope to come back to this issue in future work, but this is already enough to show how applying our taxonomy to different normative frameworks can help to illuminate different dimensions of human–technology interactions and AI ethics.

transparency. For the reasons we have just described concerning agency, this can be problematic. However, our analysis has revealed that the opposite also holds: systems that are highly reflectively transparent might impair our capacity to interact with them. Where fluent action is required to mediate timely human actions, we often require our artefacts to be potentially highly transparent-in-use. Preserving, or at least affording, the capacity for fluid action and activity while using these technologies should not be forgotten.

The consequence of our analysis is that some degree of reflective opacity in AI systems is not all bad, or at least not always all bad. Think, for instance, of the wayfinding examples discussed above. Yet, there may be situations where transparency-in-use is not always a good thing, and some of these situations may be ones that are likely to show up around AI technologies. The problem, as noted above, is that as we come to rely on transparent-in-use technologies that incorporate (relatively) reflectively opaque elements, our actions may become ever more biased in ways that are hidden or impenetrable to us. Heavy reliance on cognitive systems that are by their nature reflectively opaque seems apt to undermine agents' autonomy. For instance, some forms of information transparency may be chosen by system designers, either to protect their competitive advantage, or else to bias a system's users in nefarious ways (Clowes 2020). Systems that are highly reflectively opaque tend to be those that inhibit users from making choices over how they act within such systems. Such systems move us into the dangerous terrain of the control problem (Russell 2019), where we face the question: how can we ensure that AI remains compatible with human interests?

We can see that exploring the different varieties of transparency has important ethical implications for the design and deployment of AI systems. Our recommendation is therefore that the AI systems with which we pervasively interact should afford, to some degree, both types of transparency. In other words, we should talk of designing systems that are open to certain degrees of reflective transparency and transparency-in-use, at least if promoting human agency is what is at stake. As we have just shown, while transparency-in-use is important for our cognitive economy and fluent interactions with technologies, reflective transparency is crucial for stronger forms of human agency involving planning, self-reflection and self-shaping.[36] Making technologies customizable in this way may turn out to be one of the fundamental ways of embedding AI in our societies.

Acknowledgements
Gloria Andrada: This work is funded by national funds through FCT—Fundação para a Ciência e a Tecnologia under the project UIDB/00183/2020. Robert W. Clowes: This work is funded by national funds through FCT—Fundação para a Ciência e a Tecnologia scientific stimulus grant FCT DL 57/2016/CP1453/CT0021. Paul Smart: This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) as part of the PETRAS National Centre of Excellence for IoT Systems Cybersecurity under Grant Number EP/S035362/1.

References

Andrada, G. (2020) Transparency and the Phenomenology of Extended Cognition. *LÍMITE Interdisciplinary Journal of Philosophy & Psychology*, 15(Article 20), 1–17.

Andrada, G. (2021) Mind the notebook. *Synthese*, 198, 4689–4708.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK.

[36] See Clowes (2019a, b), for examples of how the use of Fitbit and personal tracking systems is often a way of practising agency.

Bratman, M. E. (2000) Reflection, planning, and temporally extended agency. *The Philosophical Review*, 109(1), 35–61.

Bucher, T. (2012) Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*, 14(7), 1164–1180.

Carter, J. A. (2020) Intellectual autonomy, epistemic dependence and cognitive enhancement. *Synthese*, 197(7), 2937–2961.

Clark, A. (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, New York, New York, USA.

Clark, A., & Chalmers, D. (1998) The Extended Mind. *Analysis*, 58(1), 7–19.

Clowes, R. W. (2015) Thinking in the Cloud: The Cognitive Incorporation of Cloud-Based Technology. *Philosophy & Technology*, 28(2), 261–296.

Clowes, R. W. (2019a) Immaterial engagement: human agency and the cognitive ecology of the Internet. *Phenomenology and the Cognitive Sciences*, 18(1), 259–279.

Clowes, R. W. (2019b) Screen Reading and the Creation of New Cognitive Ecologies. *AI & Society*, 34, 705–720.

Clowes, R. W. (2020) The Internet Extended Person: Exoself Or Doppelganger*? LÍMITE Interdisciplinary Journal of Philosophy & Psychology*, 15(Article 22), 1–23.

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

Cristianini, N., & Scantamburlo, T. (2020) On social machines for algorithmic regulation. *AI & Society*, 35, 645–662.

de Fine Licht, K., & de Fine Licht, J. (2020) Artificial intelligence, transparency, and public decision-making. *AI & Society*, 35(4), 917–926.

Diakopoulos, N. (2020) Transparency. In M. D. Dubber, F. Pasquale & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 197–213). Oxford University Press, New York, New York, USA.

Dreyfus, S. E., & Dreyfus, H. L. (1980) *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*. Operations Research Center, University of California, Berkeley, California. (Ref: ORC-80-2)

Ferreira, F. G. D. C., Gandomi, A. H., & Cardoso, R. T. N. (2021) Artificial intelligence applied to stock market trading: A review. *IEEE Access*, 9, 30898–30917.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., & Rossi, F. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707.

Gallagher, S. (2005) *How the Body Shapes the Mind*. Oxford University Press, Oxford, UK.

Gillett, A. J., & Heersmink, R. (2019) How navigation systems transform epistemic virtues: Knowledge, issues and solutions. *Cognitive Systems Research*, 56, 36–49.

Heersmink, R. (2013) A Taxonomy of Cognitive Artifacts: Function, Information, and Categories. *Review of Philosophy and Psychology*, 4(3), 465–481.

Heersmink, R. (2015) Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, 14(3), 577–598.

Heersmink, R., & Sutton, J. (2020) Cognition and the Web: Extended, Transactive, or Scaffolded? *Erkenntnis*, 85, 139–164.

Heidegger, M. (1927) *Being and Time*. Basil Blackwell, Oxford, UK.

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019) *Ethics Guidelines for Trustworthy AI*. European Commission, Brussels, Belgium.

Lupton, D. (2016) Digital health technologies and digital data: New ways of monitoring, measuring and commodifying human bodies. In F. X. Olleros & M. Zhegu (Eds.), *Research Handbook on Digital Transformations*. Edward Elgar Publishing Ltd., Cheltenham, England, UK.

Maravita, A., & Iriki, A. (2004) Tools for the body (schema). *Trends in Cognitive Sciences*, 8(2), 79–86.

Merleau-Ponty, M. (1945) *Phenomenology of Perception*. Routledge Press, London, UK.

Müller, V. C. (2020) Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Stanford University, Stanford, California, USA.

Nguyen, C. T. (2021) Transparency is Surveillance. *Philosophy and Phenomenological Research*.

O'Neill, O. (2020) Questioning Trust. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 17–27). Routledge, New York, New York, USA.

Russell, S. J. (2019) *Human Compatible: AI and the Problem of Control*. Viking Press, New York, New York, USA.

Smart, P. R., Heersmink, R., & Clowes, R. W. (2017) The Cognitive Ecology of the Internet. In S. J. Cowley & F. Vallée-Tourangeau (Eds.), *Cognition Beyond the Brain: Computation, Interactivity and Human Artifice* (2nd ed., pp. 251–282). Springer International Publishing, Cham, Switzerland.

Turilli, M., & Floridi, L. (2009) The ethics of information transparency. *Ethics and Information Technology*, 11(2), 105–112.

Walmsley, J. (2020) Artificial intelligence and the value of transparency. *AI & Society*, 36(2), 585–595.

Wang, F.-Y. (2008) Toward a revolution in transportation operations: AI for complex systems. *IEEE Intelligent Systems*, 23(6), 8–13.

Weller, A. (2019) Transparency: Motivations and Challenges. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700, pp. 23–40). Springer, Cham, Switzerland.

Wheeler, M. (2019) The reappearing tool: transparency, smart technology, and the extended mind. *AI & Society*, 34(4), 857–866.

Zednik, C. (2021) Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34, 265–288.