Connecting ethics and epistemology of AI

Federica Russo, Eric Schliesser, Jean Wagemans
University of Amsterdam

Abstract

The need for fair and just AI is often related to the possibility of understanding AI itself, in other words, of turning an opaque box into a glass box, as inspectable as possible. Transparency and explainability, however, pertain to the technical domain and to philosophy of science, thus leaving the ethics and epistemology of AI largely disconnected. To remedy this, we propose an epistemology for glass box AI that explicitly considers how to incorporate values and other normative considerations at key stages of the *whole process* from design to implementation and use. To assess epistemological and ethical aspects of AI systems, we shift focus from trusting the output of such a system, to trusting the process that leads to such outcome. To do so, we build on 'Computational Reliabilism' and on Creel's account of transparency. Further, we draw on argumentation theory, specifically about how to model the handling, eliciting, and interrogation of the authority and trustworthiness of expert opinion in order to elucidate how the design process of AI systems can be tested critically. By combining these insights, we develop a procedure for assessing the reliability and transparency of algorithmic decision-making that functions as a tool for experts and non-experts to inquiring into relevant epistemological and ethical aspects of AI systems. We then consider normative questions such as how social consequences that harm intersectionally vulnerable populations can be modelled in the context of AI design and implementation, drawing on work on the literature on inductive risk in the philosophy of science to think them through. Our epistemology-cum-ethics is developed from the vantage point of the conditions for enabling ethical assessment to be built into the whole process of design, implementation, and use of an AI system, in which values (epistemic and non-epistemic) are explicitly considered at each stage and by every salient actor involved. This approach, we think, complements other valuable accounts that target post-hoc ethical assessment.

1. Epistemology *aut* Ethics?

Current debates in the epistemology and in the ethics of Artificial Intelligence (AI) focus on two largely disconnected problems:

[1.] Questions of transparency / opacity of AI, i.e. A.I. as a glass or opaque box [epistemology];
[2.] Questions of how to make AI ethically compliant, ensuring that algorithms are as fair as possible and as unbiased as possible [ethics].

We say 'largely' because attempts to connect these two problems exist but differ significantly from our entry point in the debate. Colaner (2022), for instance, discusses

the question whether there is an intrinsic (ethical) value in explainable AI (hereafter: **XAI**), and provides various arguments to answer in the positive.

In this paper, we aim to establish a direct connection between these two problems by appealing to a different argumentative strategy. In our view, two dimensions of the discussion intersect. One axis we call the 'epistemological—ethical dimension'. Another axis to consider concerns the expertise of the actors involved, when posing questions about the epistemology and/or the ethics of AI. We call this second axis the 'expert--non-expert dimension'. We aim at connecting these problems [1-2] by developing a framework for ethical *and* explainable AI, or an ethics-cum-epistemology, as we shall call it. And within this approach we aim to explain how expert and non-expert actors can legitimately and meaningfully inquire into the explainability or ethical compliance of the AI.

Before we turn to our proposal, it should be noted that some authors express sceptical attitudes towards the ethics of AI in general. Hagendorff (2020), for instance, persuasively argues that more often than not, ethical guidelines do not have real impact. He writes:

> "Science- or industry-led ethics guidelines, as well as other concepts of self-governance, may serve to pretend that accountability can be devolved from state authorities and democratic institutions upon the respective sectors of science or industry. Moreover, ethics can also simply serve the purpose of calming critical voices from the public, while simultaneously the criticized practices are maintained within the organization." (Hagendorff 2020, 100)

In fact, sometimes the ethics of AI risks repeating known existing problems within the policy work of economists. For example, when economists advocate for policies that involve in principle (so-called 'Kaldor' or 'Kaldor-Hicks') compensation to the policy's expected 'losers', but then leave actual compensation to the political process, which may lack incentives or political will to do so (Oxford Lexico n.d.). The analogy is especially noteworthy that, within the ethics of AI, when post facto mitigation by third parties as a policy is advocated, this is often combined with the explicit or implicit realization that third parties may lack expertise or political will to act on mitigation (see Zarsky (2016)). In general, many mitigation strategies are vulnerable to being hostage to the political process which may itself be captured by better financed vested interests.

Our entry point in the ethics of AI is very different. Without pretending to offer the magic bullet, we aim to offer an approach to ethics and epistemology to improve on the side of ethical compliance *from the design stage* and process and to offer ways to inquire about ethical compliance *from different levels of expertise*. We aim at sketching a framework to approach the process of design, implementation, and assessment of AI that attempts to simultaneously consider ethics and epistemology, and the expertise

of the actors that inquiry about these two. Thus, any time we talk about 'process', it is not merely the algorithmic process that we have in mind, but the 'whole process', from design to implementation and to use, which of course does include technical questions about algorithmic procedures.

Because we think ethics is *not* a cherry on a cake, relegated to a post hoc analysis, we start from epistemology and seek to identify relevant points of the process at which ethics must and should come in, and in this sense, we will speak of *internalizing values* already at the design stage of an AI system. In particular, we think that many foreseeable, undesirable social consequences can be internalized in the design process in ways that naturally extend precautionary and legal practices. We think the strategies we start developing here can be developed more fully and be taught and developed in computer science departments and design schools, internalized in corporate missions, and help create a culture of responsible AI.

The paper is organized as follows. In Section 2, we position our entry point into the rich debate on the ethics of AI. We also elucidate the terminology used within the field and explain the main concerns that have shaped the research questions and issues regarding the epistemology and ethics of AI. In Section 3, we turn to the epistemology of AI; we discuss 'Computational Reliabilism' developed by Durán (2018) and Durán and Formanek (2018) as well as Creel's (2020) approach to transparency, among others. We build on these approaches and use argumentation theory, and specifically its treatment of argumentation from expert opinion (Wagemans 2011b), to develop an epistemology for glass box AI. Next, in Section 4, we further articulate our position, explaining how to include values in the process of design and implementation as well as how non-experts can inquire into the ethical compliance of an AI system -- thus offering an epistemology-*cum*-ethics. In the conclusion, we provide further detail on how our approach is distinct and complements existing approaches to link ethics and epistemology of AI.

2. AI and its ethical challenges

Artificial Intelligence (and the philosophy thereof) has a long and established tradition in the respective fields of computer and cognitive science, and in philosophy of computing. The programme of understanding and reproducing human intelligence has undergone ups and downs since the seminal work of Turing, and it is undeniable that we are witnessing renewed interest in AI (see, e.g., (Crawford 2021; Floridi 2021)). It seems that, in this new wave of interest, projects, and applications, the question of what one can *do* with an AI seems to have entered central stage besides the already studied conceptual or theoretical questions. This, alongside some high-profile AI abuses that have received media attention, has contributed to shifting the whole discourse directly to questions about ethics and governance, which have been promptly recognized as fundamental by institutions such as the European Commission. In this context, the work of the 'High Level Group on Artificial Intelligence'

is both timely and relevant and an excellent example of the usefulness of interdisciplinary and intersectoral collaborations (AI HLEG 2019). This also constitutes the background of our contribution.

To set the stage, it will be useful to clarify the meaning and use of some key terms. There is no consensus about what AI is, but the definition proposed by the HLEG will be a useful starting point for our articulation:

> "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)." (AI HLEG 2018, 1)

In a definition like the one above it is not central to pin down exactly what intelligence is, but rather the fact that, whatever it is, an artificial intelligence is a piece of software (that it can, or cannot, be embedded in a hardware will not be relevant to our arguments later on).

Generally speaking, we can take a piece of software to be the whole set of instructions telling a computer what to do. More specifically, this set of instructions will be organized in an algorithm, a term that is often given different definitions, at times emphasising their mathematical basis, their implementation, or their procedural nature (Creel 2020; Mittelstadt et al. 2016; Primiero 2020). For our purposes, it is important to keep in mind that an algorithm, or an algorithmic procedure, is a piece of code that can be nested in other codes, and that is the product of a design by one or more agents (computer scientists, scholars with complementary expertise, or other algorithms), it is designed 'to do something', and the implementation of the code is as important as earlier phases of the design stage. Our interest, however, is not merely on the algorithmic procedure, but more broadly on the whole process, from design to use, which includes algorithms.

At the time of writing, there is abundant public discussion on AI, not just about its opportunities for scientific research or industry, but also about its potential pitfalls and misuse, and the normative framework needed to avoid them (Coeckelbergh 2020; Dignum 2020; Dubber, Pasquale, and Das 2020; Liao 2020; Stahl 2021; Vallor 2016; Vieweg 2021). One line of argument is that AI may reinforce racial and economic injustice, which requires ad hoc mitigation measures and many purportedly neutral algorithms turn out to be biased or promote biased outcomes. (see e.g. (Carr 2021)).

These discussions are right in putting ethics concerns at the very top of the agenda on AI. For instance, Zarsky (2016) identifies two problems: efficiency and fairness-based concerns. According to Zarsky, there are known problems with reaching the 'right' decisions for individuals, as it happens in the case of automated procedures for

credit assessment. The apparent paradox is that, despite algorithms may be (and are) wrong in individual cases, the algorithm may still be pretty efficient (in the sense of reasonably precise in making accurate predictions or outperforming human operators) at the aggregate level. The usual solution to improve on individual-level decisions is to increase transparency (for instance about data collection and analysis). But this is no panacea because asking for more transparency quite likely means more financial costs about disclosure, more search costs, and more opportunities for confusion. Zarsky's second concern is about fairness. He distinguishes three types of concerns: unfair transfers of wealth, unfair differential treatment of similar individuals, unfair harms to individual autonomy. His point is that, in cases of unfair treatment, increasing transparency and imposing disclosure-related solutions do not necessarily mitigate or prevent them or rectify any injustices/errors. Part of the problem seems to be that focusing exclusively on transparency obfuscates the potential role of other regulatory steps needed.

For other authors, privacy is also a key concern. Kearns and Roth (2020), for instance, argue that even if the majority of the data collected by the apps we use say data is anonymized, it can be proven that it does not take much effort to retrieve sensitive personal information (see, e.g., Matsakis (2019), Zuboff (2019)).

But these are not the only ethical worries. Mittlestadt et al. (2016) aim to provide a *map of ethical concerns*. For them, questions about implementation and execution are among the problematic aspects; their mapping "[...] include ethical issues arising from algorithms as mathematical constructs, implementations (technologies, programs) and configurations (applications)" (Mittelstadt et al. 2016, 2). We find their mapping particularly useful and illuminating about what we need to pay attention to when performing an ethical assessment of an algorithmic procedure. At the same time, to complement this line of work, we aim to shift focus: we are interested in *where*, in the whole process of design, implementation, use, assessment, do these issues arise, not just *what* kind of ethical issues arise. In other words, we take the mapping of Mittlestadt et al. (2016) to be a valuable tool to run an ethical assessment of an AI *after the fact*. But our concerns are already at the level of *design*: how can we anticipate the concerns identified in the map, *while* we research, design, evaluate, and implement an AI?

This is exactly the line taken by Kearns and Roth (2020). Their point is that the attention given to performance metrics veils any explicit consideration of social values, e.g. privacy and fairness (an argument of Zarsky (2016) too). Algorithms are instruments we use to achieve something, but unlike a hammer (which is an instrument too), they have a form of agency – something that philosophers of technology have long investigated, and even before the advent of digital technologies (see, e.g., Kroes and Verbeek (2014)). So, if we have to make these algorithms 'ethical', we need to act at the level of *design*. Generally speaking, Kearns and Roth argue that internalizing values in the process of algorithm design requires setting new goals, and especially

new constraints for the learning process. It is likely that if an algorithm has to include, by design, privacy or fairness, it will have to compromise on the 'usual' performance indicators such as precision or speed. But this, Kearns and Roth argue, may often be a price well worth paying. While Kearns and Roth have quite a lot to say insightfully about possible trade-offs, they do not provide a conceptualization of how to think about the internalization of values and social ends in the design process. In fact, at key junctures they leave such decisions to 'society' and 'policy-makers'. Our point is that some of these key junctures are also in the hands of those who design algorithms, and that is why the question of how to internalize values is so important.

Our contribution is very much in line with Kearns and Roth (2020), and complements it in that we inquire about the *process* of design, implementation and use, to identify key points at which critical questions about the epistemology and ethics of an AI system can be asked. If values are incorporated in the algorithm as Kearns and Roth suggest, we should be able to check the process, as we explain through section 3 and 4. Our argument is general in character, and complementary to the more specific analysis of Morley et al. (2020), that instead maps and documents various ways in which, according to the existing literature, AI can *in practice* be made ethically compliant. In section 3, specifically, we re-examine questions about the *reliability,* or the *trust*, epistemic agents can put in the AI system. This shift puts us straight into epistemological considerations, in which ethics is embedded or internalized. But questions about explainability and about ethical compliance can be asked in different ways, or at different levels of abstraction, by different actors, as we also explain.

3. The epistemology of glass box AI

One task of an epistemology of AI is to provide an account of the reliability and precision of the machine, or else the conditions under which we trust the results, or outcome, of an algorithmic procedure. In particular, it is often argued that epistemic trust is based on features such as transparency, accuracy, or explainability (see, e.g.,(Creel 2020; Durán 2018; Durán and Formanek 2018; Mittelstadt, Russell, and Wachter 2019; Ratti and Graves 2022)). We begin by presenting 'Computational Reliabilism' (Durán 2018; Durán and Formanek 2018); hereafter: **CR**) as a reaction to this received view on epistemic trust, based on the notions of transparency, accuracy, and explainability (Section 3.1). Then, partially motivated by Creel's (2020) account of transparency, we motivate a shift in focus from the reliability of the 'outcome' (i.e., the algorithm, the code, etc.) to the reliability of the whole process of design, implementation, and use (Section 3.2).

3.1 How can experts inquire into the reliability of an AI system?

3.1.1 Computational Reliabilism and the question of transparency

Computational Reliabilism (CR) is an approach for assessing the reliability of computational processes that primarily applies to computer simulations (the main focus of the work of Durán and Formanek (2018)) but can also be used for other types

of algorithmic procedures, including much of digital technologies used in e.g. medicine (2021). CR originates in Alvin Goldman's *process reliabilism*, which intends to cash out the idea that a cognitive agent S is justified in believing the results of a given process in case it holds a tendency to produce more true results than false results (Goldman 1979). To answer the question: "how to trust the results of a computational process?", CR adapts this idea to the specific needs of algorithmic procedures and simulations, resulting in the following definition:

> (CR) if *S*'s believing *p* at *t* results from *m*, then *S*'s belief in *p* at *t* is justified. where *S* is a cognitive agent, *p* is any truth-valued proposition related to the results of a computer simulation, *t* is any given time, and *m* is a reliable computer simulation. (Durán and Formanek 2018, 654)[1]

Before we get to our analysis of the fruitfulness of CR, it is worth noting that Durán and Formanek deviate from the strategy of asking for more transparency. For several authors, for instance Humphreys (2009), there will always be an element of opacity that remains because humans are too much outside of the process. Although full transparency can never be ensured, authors such as Newman (2016) have stressed the importance of sound practices, but for Durán and Formanek this is not good enough, because some parts of the algorithm will remain inaccessible at least in real time (say because it is too costly to access). And to authors such as Symons and Horner (2014), who warn that we cannot test all possible paths, Durán and Formanek rebut that instead of testing all paths, we can use *indicators* to trust the results, despite the inherent opacity of simulations or other algorithmic procedures. It is worth noting that Durán and Formanek are not the only one rejecting transparency as the way to ensure that an AI system is trustworthy. Ananny and Crawford (2018) reach the same conclusion, but based on different motivations and types of argument. Later, we explain why we think we need to consider transparency, even though it is not *the* solution to explainability or ethical compliance. To foreshadow our own position: the key to trust the output of an AI system is not transparency alone, we need instead cues from the process.

Let us then look at the 'indicators' used in CR to establish trust in the outcome. There are four of them:

(a) verification and validation methods
(b) robustness analysis
(c) a history of (un)successful implementations
(d) expert knowledge.

[1] Durán and Formanek embed CR into an epistemology of knowledge based on 'justified true belief' (JTB). In our view, we don't need to adopt JTB, and in any case the argument is orthogonal to our purposes. So we leave the discussion of what epistemology of knowledge should base CR or our preferred approach (to be developed in this section and in section 4) for future work.

The first two of these indicators cover internal, technical aspects of algorithmic procedures, while the last two address aspects of the context in which the procedures have been developed.

Regarding (a) verification and validation methods, Durán and Formanek adopt a rather standard approach, for instance that of Oberkampf et al. (2003). Simply put, verification is about the correctness of the model and validation is about whether the model yields accurate results when confronted with the 'real world'. Existing discussions in the literature concern, for instance, how best to adapt standard definitions from computer science to simulations or AI, or whether verification is more important than validation (or the other way around), a debate that is discussed by Durán and Formanek. For us, the take-home message is that, broadly speaking, from the perspective of CR, verification is about 'internal' aspects of modelling while validation is about 'external' aspects (a kind of empirical adequacy).

The second indicator, (b) robustness analysis, in a sense, extends the scope of validation methods to *different* (but sufficiently similar) models, rather than just one model. In computer simulation, as well as in other contexts such as econometric modelling, it makes a lot of sense to test for robustness, since models can be implemented in slightly different ways, even when applied to the same data set (see also Wimsatt (2007)). According to Durán and Formanek, "the core assumption in robustness analysis is that if a sufficiently heterogeneous set of models give rise to a property, then it is very likely that the real-world phenomenon also shows the same property" (2018, 15). In this sense, robustness is similar to "consilience", namely the convergence of evidence for scientific claims (see, e.g., (Wagemans 2016; Wimsatt 2007)).

The third indicator, (c) the implementation history, is based on the idea that we should look at the 'local' and specific histories of design and implementation of AI systems, which, according to Durán and Formanek, are largely cumulative. Durán and Formanek (2018, 17–18) say:

"[…] building techniques have their own life for 'they carry with them their own history of prior (un)successes and accomplishments, and, when properly used, they can bring to the table independent warrant for belief in the models they are used to build' (Winsberg 2003, 122). We include such history of (un)successful implementations as an important source for attributing reliability to computer simulations."

In the approach of Durán and Formanek, the fourth indicator, (d) expert knowledge, is normally used in combination with the third, and has to do with the expertise of the actors involved. Expert knowledge, in CR, is taken to be some kind of attribute of a group of experts, following the approach of, e.g., Collins and Evans (2009). Understood in this way, expert knowledge is key (i) to justify why scientists believe the results (i.e., the outputs of an algorithmic procedure), because they trust the

assumptions (made by *experts*), and (ii) in determining the robustness of a simulation and the (un)successful history.

There is a lot to learn from Computational Reliabilism, and we aim to build on CR to develop our epistemology for glass box AI. Let us clarify the main points of agreement and disagreement between us and CR proponents.

We broadly agree with Durán and Formanek that each indicator comes into 'degrees' and that none is decisive in establishing trust (but the more positive scores on each, the better). We disagree, however, that there is a hierarchy in these indicators, and especially that expert knowledge is weak because it could be "idiosyncratic in several ways". We think that this is an area where the exact, experimental, and computational sciences may have something to learn from the social sciences, notably about *reflexivity*. As it has been argued in social science methodology and in philosophy social science, the point is not to wipe away expert opinion as a source of bias, but to disclose it, precisely to model and handle bias and, as we will argue, to increase transparency *of the whole process* (Breuer 2003; Cardano 2009; Levy and Peart 2017; Russo Forthcoming; Subramani 2019); this the positive view we examine later in the section. To be sure, this is a general point about methods across natural and social science, and across algorithmic procedures, simulations, and other methods.

Our concern regarding CR relates to the number of stakeholders included in the conceptualization of reliabilism. We think it is important to give visibility to as many relevant actors as possible, but the definition of CR mentions only one, i.e. the cognitive agent assessing the process. But where are designers? And where are the quality control managers and users or the evaluator of the AI system? In the literature, some contributors emphasised the need to discuss different actors or stakeholders. For instance, according to Zednik (2021), there are different stakeholders affected by the opacity of an AI system, and his solution is to identify different levels of explanation needed for different stakeholders (drawing on literature on explanation from philosophy of science). This is somehow echoed by Langer et al. (2021), who make the point that different stakeholders will have different desiderata about explainable AI. In the rest of the paper, we'll try to take into consideration different actors and their stakes in the design and assessment of an AI system. First, by developing an analogy with the evaluation of arguments from expert opinion, in section 3.2 we explain how actors having different expertise can inquire into the epistemology or ethical compliance of an AI in different ways. Second, in section 4, we argue that values have to be internalized already at the stage of design and implementation, and in this way we aim to put *designers* and their interlocuters within firms and suppliers (etc.) into a vital position of responsibility. We confine however our discussion to the expertise of *human* actors, and do not consider, for reasons of space, the interactions of human actors *with the AI* system itself.

Finally, it is important to note that, although CR is set up as a form of control on the algorithmic process, the indicators are ultimately geared to provide a *content-related justification of the outcome*. Agreed, the third and fourth indicators can be understood as being about 'good practices', but (i) they are related to a lesser role and (ii) ultimately still contributing to establish trust *in the outcome*. From the perspective of Durán and Formanek, it is clear why transparency is not of immediate help, but we think that a different take on transparency can help us materialise this shift in focus from the outcome to the process, which we think CR begins but does not complete. In making this shift, we are also able to consider the actors involved as well as their expertise: designers, peer experts, the public, institutional stakeholders, and others. Our next step is to re-introduce transparency into the picture, building on the account of Creel (2020).

3.1.2 From the reliability of the outcome to the reliability of the process

As mentioned in the previous section, Durán and Formanek (2018) do not think that transparency helps with trusting the outcome of an algorithmic procedure. One of their concerns with transparency/opacity is that these notions are not well-defined, often vaguely referring to "accessibility and surveyability conditions on justification" (2018, 647). More importantly, accredited definitions of transparency / opacity seemingly refer to intrinsic properties of a process or system, leaving out entirely relevant actors. For instance, Humphrey's definition, quoted by Durán and Formanek (2018, 648) is as follows:

"A process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at all of the epistemically relevant elements of the process." (Humphreys 2009, 618)

Admittedly, definitions of transparency / opaqueness like the one above do not help much. In our view, the account proposed by Creel (2020) is instead a fundamental step in the right direction. Creel distinguishes three types of transparency, covering different aspects of the *process*, rather than the output. In this way, we can make specific inquiries about the process, at different levels, and depending on the actor(s) involved. Interestingly for our purposes, Creel frames the problem in terms of improving *knowledge* (of the algorithm), rather than establishing trust in the output. Why transparency is key will become fully clear in the next section, about argument assessment. But first let's present Creel's account.

The need for transparency is a controversial point in the literature, as we have just seen with Durán and Formanek, but is also clear from other contributions such as Lenhard and Winsberg (2010) or Humphreys (2009). Creel's take on it is that we *do* need transparency, and for two reasons. One reason hinges on practical arguments: transparency appears to be important for communication purposes to groups involved in the design and use of algorithmic procedures. Another reason has to do with

normative considerations, putting forward the idea that this interest in transparency *is* justified. And yet, these groups don't quite know how to philosophically unpack it. Creel also makes the point, but does not develop it further in the paper, that transparency is a form of accountability towards non-experts, e.g., the public; in section 3.3 we give some indications of how the public can engage with specialised algorithmic procedures in a meaningful way. In her paper, however, Creel focuses on questions of transparency for "skilled and knowledgeable creators and users of computational systems" (2020, 572).

The question of what transparency is cannot be given one monolithic answer, because transparency can be different things. Creel distinguishes three types of transparency:
1. Functional transparency;
2. Structural transparency;
3. Run transparency.

The *first* type of transparency helps us improve on "knowledge of the algorithmic functioning of the whole". Typically, this type of transparency is achieved when humans programme the algorithm and is clearly more difficult in cases of 'kludges', i.e., in case of nesting of modelling and algorithms, as it happens in climate modelling. The *second* type of transparency helps us improve on "knowledge of how the algorithm was realized in code". The problem to address is whether the same algorithm may be realized through different codes. As Creel puts it, the question is "not just to be able to read the code; it is to understand how the code as written brings about the result of the program." (2020, 575). Clearly, in an ideal situation, to achieve structural transparency we should decompose the algorithm line by line. In practice, this can be highly time consuming and the interrelations between different parts of the algorithm difficult to know, and in some cases prone to wrong use. Creel then concludes: "[...] although we know how the learning algorithm works and what formal guarantees (if any) we have about its performance, we do not know how the learned "algorithm" brings about the classification result. Thus, we lack functional transparency." (2020, 580). It is reasonable to say that this type of transparency is the most difficult to achieve. Finally, the *third* type of transparency helps us improve "knowledge of the program as it was actually run in a particular instance, including the hardware and input data used" (2020, 569). Any considerations about material aspects of the design process, of the software, machines, or data that have been used will be relevant to establish this type of transparency.

Recall, in Creel's approach, each type of transparency, individually, does not improve the *trustworthiness* as such but rather the *knowledge* of the algorithm. This is important for us because we are developing an *epistemology* for glass box AI. So the question is not just whether transparency makes the process trustworthy, but *how we can know* that it is trustworthy. Two remarks are in order. First, a relevant analogy here can be made with explanation. The literature on mechanistic explanation should be a role model here because the decomposition and reconstruction of the mechanism, the

identification of relevant entities and activities, and generally all these epistemic practices are part of what it means to explain, including explaining the falsehood of an explanandum (Glennan and Illari 2018). Following Creel's approach, we need transparenc*ies* to explain why/how knowledge of the algorithm is improved and this about different aspects of the AI system and/or at different levels of expertise of the actors. This brings us to a second remark: it is important to specify *who* wants to know about the process. This is where the first axis of the paper (epistemology--ethics) intersects with the second axis, namely expert--non-expert.

Now, we are in a position to explicate in detail and expand on the approach of Durán and Formanek here, that includes explicitly *one actor*, namely the cognitive agent 'S', in the definition of CR. The explanation of the four indicators implicitly refers to different cognitive agents engaging in different epistemic activities, but we want to make this aspect more visible and explicit. Also, CR seems to assume expert knowledge in the evaluation of the different indicators. But in this way, the approach is of use to peer experts, while non-experts will have little or no clue about how to decide whether to trust the process. Combining an adapted version of the approach of Durán and Formanek and of Creel, we aim to explain how questions about the epistemology of AI can be asked (and are answered) differently, depending on the level of expertise of the actors involved.

To do so, we draw an analogy with the evaluation of arguments from expert opinion: although non-experts can never assess the acceptability of an expert opinion directly because this would require expert knowledge, they can assess its acceptability indirectly, namely by asking so-called 'critical questions', a mechanism by which grounds for trust (or not) are revealed (Wagemans 2011b). In the next section, we turn the CR indicators and the three types of transparency into critical questions to assess arguments from expert opinion and we shall see that adapted critical questions can also be used differently by experts (e.g., the "skilled and knowledgeable creators and users of computational systems") and by non-experts (e.g., the general public) to inquiry about epistemic and ethical aspects of an AI system.

3.2 How can non-experts inquiry into the reliability of an AI system?

In this section, we explain how experts and non-experts can assess the reliability and transparency of AI assisted decision-making through asking so-called 'critical questions' associated with argumentation from expert opinion.

To begin with, we say that whenever there is expert-to-expert communication, we are in a situation of epistemic *symmetry*, for instance when a software engineer interacts with another software engineer, with comparable expertise. Otherwise, if one of the parties does *not* hold relevant expertise, for instance a patient interacting with a physician, we say that actors are in a situation of epistemic *a*symmetry (Snoeck

Henkemans and Wagemans 2012).[2] This is a very common situation in a great many communicative domains, and it also applies to the use of AI systems in domains such as medicine and diagnosis, or finance.[3]

We now provide a simplified description of the characteristics of this argument type and indicate how it is assessed. Based on this analysis, we then develop a procedure for assessing the reliability and transparency of algorithmic decision-making.

To establish or increase the acceptability of a certain claim or point of view, individuals may refer to various types of authority. One of these types is the 'epistemic' authority, usually denoting a scientific expert who is viewed as a specialist in a certain domain. An argument in which such a reference to authority is made is called an 'argument from expert opinion' (Goodwin 2011; Wagemans 2015; D. Walton and Koszowy 2017) and its general structure is the following:

Claim q
Reason q is said/endorsed by expert E

To assess the acceptability of such an argument from expert opinion, in other words, to determine whether to accept a claim based on the fact it is said/endorsed by expert E, one can ask specific 'critical questions' (Wagemans 2011b; D. N. Walton, Reed, and Macagno 2008)). A key move we make is that since 'trusting the output or process of an AI system' is like 'trusting an expert opinion', the argument can be assessed in a similar way. What is needed, however, is an adaptation or specification of the critical questions involved.

Because in cases of epistemic asymmetry the addressee is not able to assess the argument from expert opinion in a direct way, some scholars conclude it is always unreasonable or fallacious. They label such arguments as a fallacy, in particular, as an *argumentum ad verecundiam* (Goodwin 1998; Hinton 2015; Wagemans 2011b; 2015). What these scholars ignore, however, is that the reasonableness of arguments from expert opinion can also be determined in an *indirect* way. In general, arguments of any type can be assessed by determining (1) the acceptability of the reason given in support of q and (2) the solidity of the argument lever (i.e., the support relationship between the reason and the claim) (Wagemans 2020). In the case of the argument from expert opinion, these two points of assessment can be specified as follows:

[2] The division into experts/partial experts/non-experts is of course an idealization, and apart from the fact that there will be grades of shades in between the extremes, there are also 'hybrid' experts whose expertise partly overlaps with that of others. We also confine the discussion to *human* (non-)experts engaging in the assessment of epistemic and ethical aspects of AI systems. This means that we do *not* discuss the case of a human domain-expert (say, a physician) interacting with an AI system.
[3] Our distinction between epistemic symmetry and asymmetry of the actors involved and the use of critical questions to address questions about epistemology and ethics of AI ties well with the way Burrell (2016) treats different types of opacity, which partially resonates with the different types of transparency of Creel.

(1) The acceptability of “q is said/endorsed by expert E”
(2) The solidity of the relationship between “being said/endorsed by expert E” and “being acceptable”

For each of these two points of assessment, specific evaluation procedures apply. Regarding the first point of assessment, it should be noted that “q is said/endorsed by expert E” is a complex statement that is assessed in two parts. First, it should be checked whether q is really said/endorsed by expert E. It might well be the case that q was not asserted by E at all, or that the version quoted in the argument is somehow distorted or adapted to the strategic purposes of the arguer. This can be checked by looking at a source where the original statement is mentioned. Second, it should be checked whether E is really an expert in the relevant field. For sometimes, the expert quoted in the argument is not a real expert, for instance because it is just a celebrity or someone with expertise in a different domain than the one in which the specific claim is situated.

The second point of assessment pertains to the argument lever, i.e., the relationship between “being said/endorsed by expert E” and “being acceptable”. The reason for having this second point of assessment is that even if q was really said or endorsed by E and E is a real expert in the relevant field - in other words, even if the propositional content of the reason is acceptable - it doesn’t mean the reason renders the claim acceptable. To provide a full-fledged assessment of the argument, it should also be checked whether the claim is acceptable *based on* the reason. *This aspect of the assessment is related to our suggestion to shift from justifying the outcome to justifying the process*. In this case, such assessment would entail considering whether there are any other factors that made the expert say/endorse the claim, such as a personal interest or gain, whether the expert can provide reasons in support of the claim, and whether other experts agree with the one quoted in the argument -- aspects like this will be further discussed in section 4, as they are distinctively about axiology and deontology. In case of epistemic asymmetry, the burden of acceptability is shifted from the epistemic to the normative elements (axiological and deontological aspects), which are discussed further in section 4.

The following non-exhaustive list of Critical Questions (CQ) can be used to indirectly assess the acceptability of a claim that is supported by an argument from expert opinion (Wagemans 2011b). While CQ1 and CQ2 pertain to the content of the premise “q is said/endorsed by expert E”, CQ3, CQ4, and CQ5 can be used to assess the solidity of the lever “being said/endorsed by expert E is authoritative for being acceptable”.

(CQ1) Is q is really said/endorsed by expert E?
(CQ2) Is E really an expert in the relevant field?

These two questions aim to establish whether the supported claim corresponds to the claim endorsed by the expert and whether the latter has relevant and appropriate expertise, based on which a non-expert can trust their claims.

(CQ3) Does E have a personal interest in saying/endorsing q?

This question is to exclude that major problems intervene at the deontic level; we do not develop this further, although of course it is an important and pressing issue in many situations.

(CQ4) Is E able to provide reasons in support of q?
(CQ5) Do other experts agree with E?

These two questions are the most relevant to inquiry into the epistemology of XAI (indirectly). They do not tackle technical aspects directly, but indirectly try to establish whether what is told by an expert is to be trusted.

The answers to these critical questions are related to the outcome of the assessment in the following way. If one or more of the answers are negative, the argument from expert opinion is unreasonable and the claim unacceptable. As Goodwin (2011) has observed, all criteria for judging argumentation from expert opinion are 'external' in the sense that there is no possibility of verifying *directly* what the expert actually claims to know about the AI system itself. The truth or acceptability of opinion O can only be critically tested in an indirect way, namely by asking critical questions pertaining to the premise content and the argument lever, as explained above. *This characteristic can also be ascribed to algorithms, the working of which is sometimes even opaque for the people who have designed them*.

Let us know consider the working of an argument from expert opinion in the specific contexts of AI systems. In expert-to-expert communication, we are in a situation of epistemic *symmetry* and the acceptability of the answer does not so much hinge on the authority but on the technical details provided. In practice: expert A inquiries about explainability of AI system X, and expert B can reply by mentioning (aspects) of transparency and of CR as discussed in section 3.1. This strategy for expert-to-expert communication about the epistemology of AI works, unless we have reasons to doubt reliability or integrity of the actors involved -- but that is beyond the scope of the present discussion. However, the situation is very different if a non-expert or, as is also common, a partial expert, asks an expert about XAI. In this case, expert and non-expert are in a situation of epistemic *a*symmetry. For non-experts it is difficult, if not impossible, to determine in a direct way whether an expert opinion is acceptable or not. In such cases, trust in the process can be secured by inquiring with critical questions, and ultimately it will be ensured by the authority of the expert, or by some institutionalization of their expertise, reliability, or integrity. Thus, in cases of non-

expert – expert communication, there is an ineliminable normative component, already present in the epistemology of an AI system.

4 Epistemology *cum* ethics

4.1 From epistemic to axiological (a)symmetries

Let us recap the argument thus far. In section 3, we sketched the main lines of a glass-box epistemology for AI. We argued that such epistemology opens the door to ethics and, in particular, prepares the ground for *internalizing values* in the design and implementation process, which can then be subject to specific ethical inquiry and assessment.

We distinguished two scenarios. A first scenario is that of epistemic symmetry. Here, there is an 'expert-expert' inquiry into whether and to what extent one could trust the outcome of an AI system. According to our epistemology for glass-box AI, we trust the outcome because we trust the *process*; in an expert-expert exchange technical details are addressed directly, both for epistemic (e.g., explainability) and normative aspects (e.g., fairness), which is in line with the approach of Kearns and Roth (2020): we *can* introduce ethical compliance in the technical development of the algorithm.

A second scenario is that of epistemic *a*symmetry. Here, the inquiry is from non-experts to experts, and that is a very common situation: patient and physician, mortgage applicant and bank, are but eminent examples of epistemic asymmetry. We have seen that, in cases like this, an inquiry into epistemological aspects cannot be 'direct' but makes use of critical questions associated with the 'argument from expert opinion'. The question of trust ("we trust the outcome because we trust the process") then turns into a question of reliability of the expertise, which already introduces key axiological elements into the epistemology. When a question of ethical compliance is posed, these axiological aspects become even more prominent and institutionalization will be fundamental.

In this section, we further articulate our view on the axiological aspects of assessing outcomes of AI systems. We first explain how to include values in the process of design and implementation of AI systems (Section 4.2). We then introduce the idea that AI systems are not just value-laden, but also value-*promoting*, and that to properly take values into account we need a *holistic* approach to model validation, one that is broader than CR. We will illustrate and articulate this by offering a framework for incorporating attention to harms that affect intersectionally vulnerable populations into the design process. Finally, we address the question of how non-experts can inquire into the ethical compliance of an AI system (Section 4.3). We develop an account of axiological reliability that is complementary to epistemic reliability and address the issue of how institutionalization plays a role in guaranteeing axiological reliability.

4.2 Internalizing values and holistic model validation

This is a good moment to return to the work of Kearns and Roth (2020). They explain how values *can* be incorporated into an algorithmic procedure, because the code *can* reflect specific ethical principles and values, if and only if these can be axiomatized or formalized (it does not mean there is always only one way to do so). For instance, privacy of users whose data are processed in a given algorithm *can* be operationalized, and this may take more resources, for instance in terms of time, money, or energy. According to Kearns and Roth, integrating values to make algorithms fair and unbiased will lead to a system that is epistemically less efficient, but axiologically better. For Kearns and Roth, this is a *trade-off.*

We want to argue, instead, that making an AI system ethical should not be modelled as a trade-off. It should instead be a conscious and deliberate choice to internalize some values rather than others. It is in this sense, that we speak of AI systems as value-*promoting*, a term that we borrow from Russo (2021). The moment in which we decide to promote fairness and unbiasedness, we are not trading-off with efficiency, we are proactively internalizing and promoting these values rather than others. Bezuidenhout and Ratti (2021) talk about 'embedding' values into the process, which we take an approach very close to ours. In other words: we make a normative claim that designers, engineers, and any other stakeholders involved *ought to* explicitly consider the values (epistemic and non-epistemic) that play a role at each stage. That is, we can understand predictive accuracy as one of the competing goals/values. This is the idea of internalizing values in the process. As we have suggested above without some such internalization, the 'Ethics of AI' risks remaining a form of window-dressing, relegated to a post-hoc assessment. In addition, as we will argue, the value-ladenness of the algorithm is inevitable, so better be explicit about it. So where/how do the values come in?

In section 3, we saw that the critical questions associated with the argument from expert opinion can help us assess the *whole* process, and not just aspects of it in isolation. From now on, we shall use the term 'model validation' not in the restricted sense that is custom in computer science and that refers to the adequacy of the model with respect to empirical data. As is common in general philosophy of science and in social science methodology (see e.g., (Jiménez-Buedo and Russo 2021; Morgan and Grüne-Yanoff 2013; Russo Forthcoming)), we shall instead use 'model validation' in a broader sense, as to encompass the whole process, from beginning to end as we explain next.

With critical questions we can identify key stages in the whole process that need to be assessed. Note, and this is an extension with respect to CR: the subject of evaluation is not just the output or outcome of the AI system. In line with the epistemology of glass-box AI outlined in section 3, the relevant question is: How to trust the *whole* process? The 'whole process' is not reducible to the output or the algorithmic

procedure per se but refers to the whole process that begins with establishing the need / goal of an algorithm up to the evaluation of its use and outputs. In 'normal' scientific contexts, the 'whole process' involves the formulation of the research hypothesis, selection of background knowledge and literature, up to the interpretation of results, and discussion of possible use in policy (when relevant and applicable). Technical jargon varies across disciplinary contexts and in computer science we may rather talk about the initial process of design of the algorithm, which involves studying and considering the users' need as well as technical possibilities, constraints related to implementation, costs, and use of the algorithm, or other.

In our view, it is crucial, both for epistemological and ethical reasons, that the process under consideration starts with motivating the design and development of an AI system, and it includes technical, epistemological, and ethical considerations (as well as other resources) for its development, and use. In this way, we make an important shift towards a *procedure-based* justification of the outcome. We make sense of the emphasis given in CR to the 'trust in the outcome' because what is at stake is the commitment of agents towards the outcome. But this is precisely the reason to make this shift: we can trust an outcome if we can get a grip on the *procedure* that justifies it[4]. This is well in line with Creel's approach to transparency, because her three types of transparency capture different aspects of the whole process. Also, in her key contribution, Creel (2020) mentions other relevant aspects for consideration, even if they are not developed in that paper – so with this 'procedure-based' approach, we aim to complement and further build on Creel's work on transparency.

We make this shift to trusting the process rather than the outcome, with the consequence that trusting the process means that it leads to 'well-established knowledge' and 'intended use'. On the one hand, in epistemic terms, knowledge is well-established, when the epistemic choices and constraints made during the whole process support the outcome – critical questions about the technicalities of the model, for instance verification and validation in CS terms, are here key in this respect. On the other hand, we can check whether use was indeed the intended one if this was made explicit already at the beginning of the process of design. Notice, again, that this is a procedure-based justification of what is well-established and intended, not content-based justification, kind of 'after the fact'.

But there is more. What is 'intended' need not coincide with what is 'foreseeable' (see below). But it is precisely for this reason that we should plea for a procedure-based justification of the outcome, in which at least the following stages can be identified:

- process of design (engineering level);
- any process of control, e.g., computational reliabilism;

[4] Our arguments is broadly in line with Watson's (2020) emphasis of the importance of process over product for the interpretability of machine learning.

- process of design (intended use of a technology);
- process of control (any mechanism in place to ensure that the intended use is preserved).

Insisting on the importance of the intended use makes the approach very close to the very origins of AI, rooted in the contribution of Wiener (1950), who thought that cybernetics is a moral philosophy. It is intended to make good to humans, so there is an inherent moral dimension that arguably got lost.

To stress that a procedure-based justification must encompass not only the 'technical moments' of the process, but also considerations about design and use, we dub our approach *holistic*. In our terminology, a holistic approach to model validation is a procedure-based justification of the outcome, one that includes design, implementation, assessment, and use of an AI, or of any other techno-scientific object. In their presentation, Durán and Formanek alerted the reader that they were unable, at the time of writing, "to offer a measurement of the degree of reliability" (2018, 656). Under our account, however, such a measure may not even be needed. We think that having such a measure would not solve the problem of trust in the outcome of an AI system. Our holistic approach to model validation is ultimately an argument for a more qualitative-oriented assessment, in which we need to weigh the pros and cons at all stages, but that does not necessarily map into a final 'magic' number. These two ideas -- of internalizing values and of holistic model validation -- have at least the following four consequences.

First, model validation is to be done with *specific purposes* in mind. This means shifting jargon from the 'correct' model to the 'useful' model. This shift is far from innocent, because in this usefulness we can immediately embed, for instance, helping disadvantaged groups, or addressing other intersectoral vulnerabilities. Likewise, we can qualify alleged epistemic values; for instance, 'accuracy' is not an absolute epistemic value, but carries important axiological components, for instance accuracy *for whom*? This is no breaking news, as the argument that epistemic values carried some axiological contents was already made in the influential work of Douglas (2009).

Second, ethical assessment stands in a continuum with model validation, and they both begin very early at the design stage. We plead for an ex-ante ethical assessment, to be carried out in combination with all sorts of epistemological and methodological considerations, which should be carried out continuously through all stages, from design until implementation, use, and monitoring of use and maintenance of the system.

When one looks at the application of a reliable process in a social context, one is not just interested in its reliability for a specific task. One would also like to know what the effects are of failure. In particular, one would like to know something about the distribution of

possible or likely harms on different kinds of populations, especially if these populations have different kinds of vulnerabilities (and appetites for risk). So, for example, something may function reliably as designed with industry beating low failure rates; yet when it breaks, all too rarely, the artifact may still be especially dangerous for kids. Or, some safety gears work swimmingly on average male subjects, less so on average female subjects (e.g., some medicines interact badly with pre-existing conditions in subsets of the population). Now, in many cases the harms that follow from such *selective* or *asymmetric vulnerabilities* can be internalized in the design, implementation, and testing process (and often this is mandated legally or by in-house risk assessment or exploration with stakeholders).

How to characterize what counts as an asymmetric vulnerability is not so easy especially, because many of the ethically or politically most salient harms may only become asymmetric due to causally intersectional effects (Bright, Malinsky, and Thompson 2016). In addition, some asymmetric harms may be due to the fact that a truthful *P reinforces* or entrenches a socially bad status quo *Q*. For many purposes one may wish to distinguish among such selective vulnerabilities, but here we lump them together as an especially important set of unfair outcomes (Mittelstadt et al. 2016). So, now we can enhance the Durán and Formanek's (2018) framework as follows:

> (ECR) if *S*'s believing *p* at *t* results from *m*, then *S*'s belief in *p* at *t* is justified. where *S* is a cognitive agent, *p* is any truth-valued proposition related to the results of an AI, *t* is any given time, and *m* is a reliable algorithmic mediation without generating asymmetric harms to vulnerable populations.

Here 'reliable' already presupposes an ordinary use of the reliable algorithm in an assigned task. One thing that follows from this is that in order to generate an ECR, its sources must also be made to seek out and track asymmetric vulnerabilities. While this clearly makes initial research and development (R&D) more expensive, it may also reduce litigation costs and social harms (including withdrawal of the product) downstream.

Third, algorithmic mediation may generate both unintended and unforeseeable outcomes. Here, too, there are many subtleties. Some unintended consequences may just be a matter of negligence. And these can be simply assimilated to (ECR). Morally, legally, and politically one may be held accountable for those if there are harms in use.

Other consequences may be unforeseeable in *detail*, or their tokens unknown, even though the outcome *pattern* (or outcome type) may be quite predictable after a while. For example, algorithmic mediation has made financial markets move at much higher speeds and has also increased the likelihood of mini and maxi flash crashes (Draus and van Achter 2012). The first was entirely predictable (and desired), but the (evolution of) exact speed(s) and volume of market transactions may have been unknowable in advance. And that it would generate new kinds of financial transactions was also known, even if the exact strategies were not. By contrast, it's possible that the likelihood of flash crashes was initially unexpected. But by now any given mini-crash may be surprising or unpredictable, but that they occur is foreseeable and so they become a 'new normal' (Kirilenko et al. 2017).

That is to say, unforeseeable tokens can occur in foreseeable outcome patterns/types. If an outcome pattern has possible tokens with asymmetric vulnerabilities, these patterns should, all things being equal, be avoided and ought to be internalized in ECR. (Of course, sometimes one can compensate for downside risks, etc.) So, we propose the following modification to our framework:

> (ECR) if *S*'s believing *p* at *t* results from *m*, then *S*s belief in *p* at *t* is justified. where *S* is a cognitive agent, *p* is any truth-valued proposition related to the results of an AI, *t* is any given time, and *m* is a reliable algorithmic mediation without **(intentionally)** generating foreseeable asymmetric harm patterns to vulnerable populations.

Obviously, this leaves the prevention, accountability, and remedy of some unforeseeable asymmetric harm patterns outside (ECR), so we do not view this as the last word.

Fourth, a crucial feature of algorithmic mediation is (as Mittelstadt et al. (2016) note) that it can affect how our social reality is conceptualized, and becomes actionable in ways that are utterly unexpected (including a reinforcement of a bad status quo). So, algorithmic mediation can generate consequences that are not just unintended, but also unforeseeable in principle because they are *transformative* (Mittelstadt, Russell, and Wachter 2019) also use this terminology going back to Floridi (2016)). Here, too, the fact that an algorithmic meditation is transformative may be intended and foreseeable. And it is possible that some of the higher order outcome-patterns including the asymmetric vulnerabilities can be predicted. And that is assimilable to (ECR). The glass-box epistemology sketched in section 3 should help, we think, in ensuring the *possibility* of inspecting the system at any time, and to allow for both expert--expert and non-expert--expert queries about epistemological and ethical aspects alike. Nevertheless, we also think that there is an important role of institutionalization in all this, as we explain next in the section.

4.3 Axiological authority, and the role of institutionalization

As we explained in Section 3.2, critical questions associated with the argument from expert opinion are meant to facilitate the non-expert--expert exchange about epistemological aspects of an AI system. Regarding this exchange, it is important to note that having an answer to the critical questions does not take away the epistemic asymmetry but leaves it intact. The asymmetry can never be levelled out: there is no way in which transparency of the process *alone* can answer the question. It can only be handled, namely by targeting axiological rather than epistemic aspects.

But the sheer fact that non-experts can ask critical questions to evaluate AI systems is not enough. In addition, given their status as non-experts, they should be facilitated to do so. In other words, what is needed is an institution that guarantees the solidity of the process both from an epistemological and ethico-political perspective. In fact, even if we cannot check the process, we can meaningfully ask (and expect) that institutionalization makes criteria and motivation explicit and transparent, and that this form of institutionalization also safeguards them – for an explanation of how various types of 'institutional safeguards' can enable non-experts to find answers to critical

questions related to expert opinion in the medical domain, see (Snoeck Henkemans and Wagemans 2012) . On these aspects, the asymmetry between expert non-expert remains, and we can handle it by asking critical questions to the relevant *institutional authority* or by making use of specific rights that help us to find answers to these questions in a different way.

Introducing authorities may seem like a cop out. We had promised to internalize values into the design process, but now we are introducing institutions beyond it. That is a feature but not a bug of our approach. It also fits the practice of engineering more broadly: nearly all fields of engineers have professional associations, codes of ethics, certification authorities (including for ongoing training and recertification), and public/private institutions that design and promote quality metrics and standards (see, e.g., (DeMartino 2011; Barry and Herkert 2014).

We have distinguished between two routes for assessing the trustworthiness of AI outcomes. The first route is when there is a situation of epistemic symmetry in the sense that the assessor has access to *the process* through which the outcome has been generated. Here, experts can inquire directly into several aspects of the process (including CR indicators and Creel's (2020) types of transparency). This route assumes expert knowledge and is thus only viable by experts. The second route is when there is a situation of epistemic asymmetry, i.e., when a non-expert evaluates an AI generated opinion. As we have seen, in this situation the evaluator can make use of the critical questions associated with the argument from expert opinion, which address two aspects (1) the correctness of the interpretation of the opinion and (2) the trustworthiness *of the source* of the opinion. To enable non-experts to carry out this assessment, several other conditions should be fulfilled. We will call these conditions 'institutional safeguards', which can be seen as an 'institutionalized anticipation of the critical questions pertaining to argumentation from expert opinion (Snoeck Henkemans and Wagemans 2012).

An example of such a safeguard in the domain of medical communication is when the critical question whether the person is a real expert can be checked by consulting the official list of doctors (in the Netherlands, this list is called the BIG register). Within the same domain, the critical question whether other experts agree with the opinion referred to in the argument is institutionally anticipated by granting the patient the right to ask another doctor, the so-called right to a second opinion. Ideally, and this is a measure for the degree of institutionalization, all the critical questions are institutionalized so as to correct for the epistemic asymmetry. Or, to give another example, as a non-expert, one does not hold epistemic symmetry with respect to the designers of the mortgage algorithm. But one can inquire the relevant authority (i.e., bank, or bank officers) about the underlying values used in the design process. If 'ethical banking' has any meaning, this is what should be included.

Our approach to internalize values, to perform *ex-ante* ethical evaluation, and to inquiry into relevant axiological authorities well complements recent approaches to ethics-based auditing, which are however *ex-post*. For instance, our approach is very much in line and complements that of Mökander and Floridi (2021). They propose the main lines for an ethics-based auditing. We agree with Mökander and Floridi that ethics is not about the result but about the process, and that the process of monitoring must be continuous. We find particularly valuable their roadmap to guide an ethics-based auditing. They list several constraints, and at different stages of the process. Notably, they distinguish between conceptual, technical, economic, and social, and organizational and institutional constraints. Auditing mechanisms clearly vary depending on which level is tackled. Our approach is complementary to Mökander and Floridi because, while they focus on *auditing*, we are interested in the perspective of the designer, and so how, from the side of the developer, we can follow a process that *internalizes* values. Similarly, the guidelines of the High-level Group of the European Commission push in the direction of more compliance, which at times is difficult because of a mismatch between the ethical principles we wish to promote and the legal bases that would enforce them (Pupillo et al. 2021).

5 Discussion and conclusion

We started this paper by noticing that debates about the epistemology and ethics of AI are largely disconnected. We offered an overview of the current approaches to ethical aspects of AI, and, as we observed, these approaches focus on developing criteria for determining the trustworthiness of the outcome of algorithms and in doing that, they can be characterized as 'post hoc'. Moreover, they assume that the assessor is a high-level expert who understands the workings of algorithms, thus making the ethics of AI heavily dependent on the epistemology of AI.

In this paper, we have set out to complement ethical approaches to AI by developing a normative framework that establishes a connection between the ethical and epistemological aspects of AI. The framework has two main characteristics contrasting with existing approaches. First of all, it focuses on the *whole process* of design, implementation, and use of AI systems. Second, it enables experts and non-experts alike to act as an assessor of such processes. The framework combines insights from argumentation theory and holistic model validation and brings together epistemological and axiological aspects of assessing AI-assisted decision-making. For this reason, it can be characterised as an *epistemology-cum-ethics*.

We don't claim originality in identifying the missing links between ethics and epistemology. Yet, our approach is significantly different from some available accounts, and so it complements them in important ways. Within epistemology, there is a trend to take a more explicit stance about ethics and to contribute to the ethics of science discussion. Bezuidenhout and Ratti (2021), for instance, focus on *teaching* data ethics; our approach is broader because it focuses on research and development

(R&D). Our approach is distinct from Bezuidenhout and Ratti also because they explicitly adopt a microethics approach, based on virtue theory, while our approach is not committed to a microethics approach nor to a virtue theoretical approach.

Furthermore, we believe our approach can work across both the micro and macro levels. A general approach to model validation, in which epistemic and non-epistemic values are internalized should be able to explain, in each specific case, both the micro- and macro-dimension not just of ethics, but also of epistemology. Also, our approach is about how modelling practices, including AI, can be value-*promoting*, through an epistemology-cum-ethics approach. Our idea of value-promoting is not necessarily based on virtue ethics (see good discussion in Bezuidenhout and Ratti (2021), in the context of data ethics). We think of this as an advantage: part of the problem is the pretention, in scientific camps of the value-neutrality of scientific methods, algorithms, or anything of the like. They are *never* neutral, and the point of epistemology-cum-ethics is not to plea for promoting some moral virtues (also debatable across cultures), but to raise awareness that some of these tools will always, even implicitly, promote *some* values. Thus, this needs to be addressed, before a discussion about *which* values ought to be promoted. This feature of our approach marks an important extension with respect to 'value-sensitive design' too, because in this approach is confined to the design of technical artefacts, and how values become to be embodied in such artefacts (Friedman and Hendry 2019; Nair 2018; van de Poel 2020). While clearly in line with value-sensitive design, our approach is broader in scope.

With respect to approaches in the epistemology of science and technology, our approach is broader than existing ones also in another sense: much of discussion about data science and ethics (see again Bezuidenhout and Ratti (2021), references therein, and especially Floridi and Cowls (2019)), seem to address the question of ethics in digital techno-scientific environments as a special case. In our view, however, AI and digital technologies are not special with respect to 'normal' scientific contexts. Rather, they should be seen as a case in point. For this reason, we embedded the question of trust in the outcome of AI systems in a broader framework of argumentation theory and holistic model validation, which also applies to 'normal' techno-scientific contexts.

Moreover, although the topic of our research is similar to that of the field of 'critical technical practice', our method has an epistemological and *normative* thrust rather than a socio-cultural one. We share with the literature within this field the idea that planning includes, beyond technical aspects, a vernacular aspect; for us, this vernacular aspect is connected to argumentation theory and holistic model validation. But unlike that literature we believe we can offer a framework in which practices can be normatively *improved*, and not 'just' assessed based on socio-cultural considerations.

Finally, our approach entails a further extension of the application of indirect assessment methods for arguments from expert opinion in situations of epistemic asymmetry. While such methods have been applied for this purpose to, for instance, politicians' references to economic expertise (Wagemans 2015), medical expert opinion in doctor-patient communication and its institutional guarantees (Snoeck Henkemans and Wagemans 2012), the application of insights about the assessment of argument from expert opinion to AI has been alluded to (Wagemans 2011a) but never worked out in full detail.

Acknowledgments

The research for this paper was conducted by the authors as part of a seed money grant of the University of Amsterdam, for the project 'Towards an Epistemological and Ethical XAI'. We received very valuable feedback by Katie Creel, Juan M. Durán, Gregory Wheeler during the project meetings. We presented an earlier version of this paper at the Workshop 'Bias and Discrimination in Algorithmic Decision Making' organized at the University of Hannover, and as part of the seminar series of the Platform for Ethics and Politics of Technology (PEPT) at the University of Amsterdam. We are grateful for all the comments and questions received during these events. We received useful suggestions also from Aybüke Özgün and Emanuele Ratti, who read an earlier version of the manuscript. Any error remains of course ours.

References

AI HLEG. 2018. 'A Definition of AI: Main Capabilities and Scientific Disciplines'. Brussels: European Commission. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf.

———. 2019. 'Ethics Guidelines for Trustworthy AI'. Brussels: European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Ananny, Mike, and Kate Crawford. 2018. 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability'. *New Media & Society* 20 (3): 973–89. https://doi.org/10.1177/1461444816676645.

Barry, Brock E., and Joseph R. Herkert. 2014. 'Engineering Ethics'. In *Cambridge Handbook of Engineering Education Research*, edited by Aditya Johri and Barbara M. Olds, 673–92. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139013451.041.

Bezuidenhout, Louise, and Emanuele Ratti. 2021. 'What Does It Mean to Embed Ethics in Data Science? An Integrative Approach Based on Microethics and Virtues'. *AI & SOCIETY* 36 (3): 939–53. https://doi.org/10.1007/s00146-020-01112-w.

Breuer, Franz. 2003. 'Subjectivity and Reflexivity in the Social Sciences: Epistemic Windows and Methodical Consequences'. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 4 (2). https://doi.org/10.17169/fqs-4.2.698.

Bright, Liam Kofi, Daniel Malinsky, and Morgan Thompson. 2016. 'Causally Interpreting Intersectionality Theory'. *Philosophy of Science* 83 (1): 60–81. https://doi.org/10.1086/684173.

Cardano, Mario. 2009. *Ethnography and Reflexivity. Notes on the Construction of Objectivity in Ethnographic Research*. Vol. 1. Torino: Dipartimento di scienze sociali Università degli studi di Torino.

Carr, Kareem @@kareem_carr. 2021. ‘FOUR Things to Know about Race and Gender Bias in Algorithms’, 27 March 2021. https://twitter.com/kareem_carr/status/1375828049720135691.
Coeckelbergh, Mark. 2020. *AI Ethics*. The MIT Press Essential Knowledge Series. Cambridge, MA: The MIT Press.
Colaner, Nathan. 2022. ‘Is Explainable Artificial Intelligence Intrinsically Valuable?’ *AI & SOCIETY* 37 (1): 231–38. https://doi.org/10.1007/s00146-021-01184-2.
Collins, Harry, and Robert Evans. 2009. *Rethinking Expertise*. Paperback edition. Chicago London: University of Chicago Press.
Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.
Creel, Kathleen A. 2020. ‘Transparency in Complex Computational Systems’. *Philosophy of Science* 87 (4): 568–89. https://doi.org/10.1086/709729.
DeMartino, George. 2011. *The Economist’s Oath: On the Need for and Content of Professional Economic Ethics*. Oxford ; New York: Oxford University Press.
Dignum, Virginia. 2020. *Responsible Artificial Intelligence: How to Develop and Use Ai in a Responsible Way.* S.l.: SPRINGER.
Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, Pa: University of Pittsburgh Press.
Draus, Sarah, and Mark van Achter. 2012. ‘Circuit Breakers and Market Runs’. CSEF Working Papers 313. Centre for Studies in Economics and Finance (CSEF). University of Naples. https://ideas.repec.org/p/sef/csefwp/313.html.
Dubber, Markus Dirk, Frank Pasquale, and Sunit Das, eds. 2020. *The Oxford Handbook of Ethics of AI*. Oxford Handbooks Series. New York, NY: Oxford University Press.
Durán, Juan M. 2018. *Computer Simulations in Science and Engineering*. New York, NY: Springer Berlin Heidelberg.
Durán, Juan M., and Nico Formanek. 2018. ‘Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism’. *Minds and Machines* 28 (4): 645–66. https://doi.org/10.1007/s11023-018-9481-6.
Floridi, Luciano. 2016. *The 4th Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press.
———, ed. 2021. ‘Ethics, Governance, and Policies in Artificial Intelligence’.
Floridi, Luciano, and Josh Cowls. 2019. ‘A Unified Framework of Five Principles for AI in Society’. *Harvard Data Science Review*, June. https://doi.org/10.1162/99608f92.8cd550d1.
Friedman, Batya, and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA, USA: MIT Press.
Glennan, Stuart, and Phyllis Illari, eds. 2018. *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. Routledge. https://www.routledge.com/The-Routledge-Handbook-of-Mechanisms-and-Mechanical-Philosophy/Glennan-Illari/p/book/9780367573416.
Goodwin, Jean. 1998. ‘Forms of Authority and the Real Ad Verecundiam’. *Argumentation* 12 (2): 267–80. https://doi.org/10.1023/A:1007756117287.
———. 2011. ‘Accounting for the Appeal to the Authority of Experts’. *Argumentation* 25 (3): 285–96. https://doi.org/10.1007/s10503-011-9219-6.
Hagendorff, Thilo. 2020. ‘The Ethics of AI Ethics: An Evaluation of Guidelines’. *Minds and Machines* 30 (1): 99–120. https://doi.org/10.1007/s11023-020-09517-8.
Hinton, Martin David. 2015. ‘Mizrahi and Seidel: Experts in Confusion.’ *Informal Logic* 35 (4): 539. https://doi.org/10.22329/il.v35i4.4386.
Humphreys, Paul. 2009. ‘The Philosophical Novelty of Computer Simulation Methods’. *Synthese* 169 (3): 615–26. https://doi.org/10.1007/s11229-008-9435-2.
Jiménez-Buedo, María, and Federica Russo. 2021. ‘Experimental Practices and Objectivity in the Social Sciences: Re-Embedding Construct Validity in the Internal–External Validity Distinction’. *Synthese*, June, 1–31. https://doi.org/10.1007/s11229-021-03215-3.

Kearns, Michael, and Aaron Roth. 2020. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. New York: Oxford University Press.
Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. 'The Flash Crash: High-Frequency Trading in an Electronic Market: The Flash Crash'. *The Journal of Finance* 72 (3): 967–98. https://doi.org/10.1111/jofi.12498.
Kroes, Peter, and Peter-Paul Verbeek, eds. 2014. *The Moral Status of Technical Artefacts*. Philosophy of Engineering and Technology, v. 17. Dordrecht ; New York: Springer.
Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. 'What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research'. *Artificial Intelligence* 296 (July): 103473. https://doi.org/10.1016/j.artint.2021.103473.
Lenhard, Johannes, and Eric Winsberg. 2010. 'Holism, Entrenchment, and the Future of Climate Model Pluralism'. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41 (3): 253–62. https://doi.org/10.1016/j.shpsb.2010.07.001.
Levy, David M, and Sandra J Peart. 2017. *Escape from Democracy: The Role of Experts and the Public in Economic Policy*. https://www.cambridge.org/core/books/escape-from-democracy/D56EB10CECD0CAC0CCDF6B3F54344C5D.
Liao, S. Matthew, ed. 2020. *Ethics of Artificial Intelligence*. New York, NY, United States of America: Oxford University Publication.
Matsakis, Louise. 2019. 'The WIRED Guide to Your Personal Data (and Who Is Using It)'. *Wired*, 15 February 2019. https://www.wired.com/story/wired-guide-personal-data-collection/.
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. 'The Ethics of Algorithms: Mapping the Debate'. *Big Data & Society* 3 (2): 205395171667967. https://doi.org/10.1177/2053951716679679.
Mittelstadt, Brent Daniel, Chris Russell, and Sandra Wachter. 2019. 'Explaining Explanations in AI'. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–88. Atlanta GA USA: ACM. https://doi.org/10.1145/3287560.3287574.
Mökander, Jakob, and Luciano Floridi. 2021. 'Ethics-Based Auditing to Develop Trustworthy AI'. *Minds and Machines* 31 (2): 323–27. https://doi.org/10.1007/s11023-021-09557-8.
Morgan, Mary S., and Till Grüne-Yanoff. 2013. 'Modeling Practices in the Social and Human Sciences. An Interdisciplinary Exchange'. *Perspectives on Science* 21 (2): 143–56.
Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices'. *Science and Engineering Ethics* 26 (4): 2141–68. https://doi.org/10.1007/s11948-019-00165-5.
Nair, Navneet. 2018. 'What Is Value-Centered Design?' *UX Collective* (blog). 2018. https://uxdesign.cc/what-is-value-centered-design-a9c5fbf2641.
Newman, Julian. 2016. 'Epistemic Opacity, Confirmation Holism and Technical Debt: Computer Simulation in the Light of Empirical Software Engineering'. In *History and Philosophy of Computing*, edited by Fabio Gadducci and Mirko Tavosanis, 487:256–72. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47286-7_18.
Oberkampf, William, Charles Hirsch, and Timothy Trucano. 2003. 'Verification, Validation, and Predictive Capability in Computational Engineering and Physics.' SAND2003-3769, 918370. https://doi.org/10.2172/918370.
Oxford Lexico. n.d. 'Compensation Principle'. In *Oxford Reference*. https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095628821.
Poel, Ibo van de. 2020. 'Embedding Values in Artificial Intelligence (AI) Systems'. *Minds and Machines* 30: 385–409. https://doi.org/10.1007/s11023-020-09537-4.

Primiero, Giuseppe. 2020. *On the Foundations of Computing*. First edition. New York: Oxford University Press.
Pupillo, Lorenzo, Stefano Fantin, Afonso Ferreira, Carolina Polito, and Centre for European Policy Studies. 2021. *Artificial Intelligence and Cybersecurity Technology, Governance and Policy Challenges: Final Report of a CEPS Task Force*. https://www.ceps.eu/download/publication/?id=33262&pdf=CEPS-TFR-Artificial-Intelligence-and-Cybersecurity.pdf.
Ratti, Emanuele, and Mark Graves. 2022. 'Explainable Machine Learning Practices: Opening Another Black Box for Reliable Medical AI'. *AI and Ethics*, February. https://doi.org/10.1007/s43681-022-00141-z.
Russo, Federica. 2021. 'Value-Promoting Concepts in the Health Sciences and Public Health'. *Philosophical News*, Special issue 'Ethics, Health Data and Bio-Citizenship', 22 (10): 135–48.
———. Forthcoming. *Techno-Scientific Practices: An Informational Approach*. Rowman and Littlefield International.
Snoeck Henkemans, A. F., and J. H. M. Wagemans. 2012. 'The Reasonableness of Argumentation from Expert Opinion in Medical Discussions: Institutional Safeguards for the Quality of Shared Decision Making'. In *Iowa State University Summer Symposium on Science Communication*, 12247422. Iowa State University, Digital Press. https://doi.org/10.31274/sciencecommunication-180809-83.
Stahl, Bernd Carsten. 2021. *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-69978-9.
Subramani, Supriya. 2019. 'Practising Reflexivity: Ethics, Methodology and Theory Construction'. *Methodological Innovations*, May. https://doi.org/10.1177/2059799119863276.
Symons, John, and Jack Horner. 2014. 'Software Intensive Science'. *Philosophy & Technology* 27 (3): 461–77. https://doi.org/10.1007/s13347-014-0163-x.
Vallor, Shannon. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York, NY: Oxford University Press.
Vieweg, Stefan, ed. 2021. *AI for the Good: Artificial Intelligence and Ethics*. Management for Professionals. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-66913-3.
Wagemans, Jean H. M. 2011a. 'Argumenteren Met Behulp van Juridische Expertsystemen. Commentaar Op Mireille Hildebrandt, 'Oordeelsvorming Door Mens En Machine: Heuristieken, Algoritmes En Legitimatie'. In *Gewogen Oordelen: Essays over Argumentatie En Recht: Bijdragen Aan Het Zesde Symposium Juridische Argumentatie 24 Juni 2011*, 357–60. Erasmus University Rotterdam: Boom Juridische uitgevers. https://hdl.handle.net/11245/1.462093.
———. 2011b. 'The Assessment of Argumentation from Expert Opinion'. *Argumentation* 25 (3): 329–39. https://doi.org/10.1007/s10503-011-9225-8.
———. 2015. 'Argumentation from Expert Opinion in the 2011 U.S. Debt Ceiling Debate'. In *Disturbing Argument: Selected Works from the 18th NCA/AFA Alta Conference on Argumentation*, edited by Catherine Helen Palczewski, 49–56. London ; New York: Routledge, Taylor & Francis Group.
———. 2016. 'Criteria for Deciding What Is The@ Best Scientific Explanation'. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, edited by Dima Mohammed and Marcin Lewiński, II:43–54. London: College Publications.
———. 2020. 'Why Missing Premises Can Be Missed: Evaluating Arguments by Determining Their Lever'. In *Proceedings of OSSA 12: Evidence, Persuasion & Diversity*, edited by J. Cook. OSSA Conference Archive. https://scholar.uwindsor.ca/ossaarchive/OSSA12/Saturday/1.

Walton, Douglas, and Marcin Koszowy. 2017. ‘Arguments from Authority and Expert Opinion in Computational Argumentation Systems’. *AI & SOCIETY* 32 (4): 483–96. https://doi.org/10.1007/s00146-016-0666-3.
Walton, Douglas N, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*.
Watson, David. 2020. ‘Conceptual Challenges for Interpretable Machine Learning’. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3668444.
Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, Mass: Harvard University Press.
Winsberg, Eric. 2003. ‘Simulated Experiments: Methodology for a Virtual World’. *Philosophy of Science* 70 (1): 105–25. https://doi.org/10.1086/367872.
Zarsky, Tal. 2016. ‘The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making’. *Science, Technology, & Human Values* 41 (1): 118–32. https://doi.org/10.1177/0162243915605575.
Zednik, Carlos. 2021. ‘Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence’. *Philosophy & Technology* 34 (2): 265–88. https://doi.org/10.1007/s13347-019-00382-7.
Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Paperback edition. London: Profile Books.