Scientia*et*Fides 14(2)/2026 ISSN 2300-7648 (print) / ISSN 2353-5636 (online)

Received: May 15, 2025. Accepted: January 23, 2026 DOI: http://dx.doi.org/10.12775/SetF.2026.018

MICHAŁ GŁOWALA
Uniwersity of Wrocław
mm.glowala@gmail.com
ORCID: 0000-0003-3687-1716

AI Machines vs the Non-Processing Minds of Classical Theism. Transparency, Trust, and the Deep Black-box Problem

Abstract. In the paper I contrast what I take to be a universal feature of all the AI research – namely focusing on processing information or transitions from one state of a machine to another – with the role assigned to non-processing minds in the classical theism framework. I briefly discuss the latter in epistemological context – the issue of inferential internalism and some problems of transparency, opacity, and trust. In this I show that the contrast between AI and the classical theism outlook is highly relevant for the cluster of issues relating to transparency or opacity, reliabilism, and trust, which is discussed today as the black-box problem. In particular, I identify in this context what I take to be the most serious aspect of the black-box problem; I reject the idea that the problem is a superficial one; and I claim that it poses basic limits on the AI research and cannot be solved along the lines of Explainable Artificial Intelligence.

Keywords: artificial intelligence, classical theism, unchangeable mind, inferential internalism, religious faith, black-box problem

Contribution. The paper claims that the universal and highly relevant feature of otherwise utterly different AI machines is that they all process information, that is, make transitions from one state of a machine to another. This feature is contrasted

with the role played in the classical theism by non-processing cognitive activity, in the context of both the very concept of religious faith (as a sort of trust) and the traditional concepts of human learning and teaching. The fact that AI is basically unable to emulate non-processing cognitive activity shows both the depth of the black-box problem and its irresolvability within the AI research.

Use of AI. The AI tools were not used in the preparation of the manuscript.

Introduction

In spite of deep diversity of strands and attitudes within the AI research, there seems to be one basic feature they all share; they all focus on *transitions* from one state of a machine to another, in other words: on *processing* information. This is clear, to start with, in a simple example of a finite Turing machine (see e.g. van Leeuven and Wiedermann 2001, 1036). We know today numerous kinds of machines that depart in various ways from that simple case; for example, hypercomputers are supposed not to be bound by the limits of computability (see e.g. Syropoulos 2008); neural networks have deep learning possibilities; many solutions involve parallel processing (see e.g. Rosta 2000, 501–534). Moreover, it is possible to study both *abstract* machines of a given kind and their *physical* implementations, e.g. quantum computers. But, to repeat: what those utterly different approaches do have in common is focusing on various forms and aspects of *transitions* from one state of an (abstract or physical) machine to another one; computational and cognitive tasks in general are analysed in terms of *processing*; and what is crucial is the *speed* of the processing, depending on algorithms, on excessive use of parallel processing, or on the architecture of physical computer systems. In other words, the AI research is focused on and limited to designing more and more sophisticated ways of processing information, moving from one state of a machine to another one, in more and more rapid ways.

Now there is a striking – although perhaps too rarely noted – contrast between that ubiquitous feature of the AI landscape and the landscape of classical theism. Two points are particularly relevant here. On the one

hand, the latter landscape is marked with the *absolutely unchangeable* and *non-processing* divine mind. On the other hand, the claim that there is such an absolutely unchangeable mind is a consequence of the claim that even human thinking cannot consist *solely* in *processing*; clearly there are many transitions in the human mind (including both inferring new conclusions and updating or correcting the premises, changing one's mind). Within the classical theism framework, the perfection of a mind does not consist solely in its (more or less sophisticated and more or less rapid) way of processing information, but rather in the scope of its activity that does *not* consist in processing information.

In this paper I show that this contrast between processing and non-processing minds is immediately related to some issues of transparency and opacity of other minds, and the very idea of trust. In this way, I show that it is strictly relevant to the so-called black box problem with AI. The problem is, roughly speaking, that we often lack understanding of the algorithms that link the input data with the outputs, and the lack of understanding of the link seems relevant to the *trust* we could reasonably place in various forms of AI. The problem seems particularly important where the lack of transparency seems deep (for example in the deep learning contexts) and where the matter is particularly serious (for example in legal or medical contexts). Some authors have recently claimed that the problem is not well formulated, consists in an unfortunate conflation of various issues and is henceforth illusory or exaggerated (Brożek et al. 2024); some authors have proposed normative constraints for explainable artificial intelligence (XAI) (Zednik 2021). In this paper I attempt to (i) identify, in the context sketched above, the core of the black-box problem; (ii) show that the *trust* problem is far more serious that it may seem and (iii) show that it cannot be solved along the lines suggested in XAI, but shows some inevitable limits of the AI research.

I proceed in the following way. I start with a brief presentation the idea of non-processing cognitive activities within the framework of the classical theism (1), starting with some general background (1.1) and then proceeding to some issues of inferential internalism (1.2) and the issues of transparency and trust (1.3) involved in traditional pictures of human teaching/learning activity (1.3.1) and the very religious faith as

a sort of trust (1.3.2). Then I claim that the AI research is basically unable to analyse or emulate the non-processing cognitive activity of the kind described in 1; in other words, I show some general features of AI that seem striking in the context of the non-processing cognitive performances (2). Finally I discuss in this setting the black-box problem in some detail (3). In Conclusion, I offer a summary of results.

1. Non-processing minds in the classical theism: some basic ideas and contexts

Open theists claim that the very idea of the absolutely unchangeable mind of God is quite foreign to religious practices and hard to square with the conception of a personal God either in philosophy or in the Biblical tradition (call it the general open theism objection; for a brief overview see e.g. Davies 1993, 143–147). Below I sketch some reason to reject the general open theism objection (1.1) and I briefly discuss the relevance of non-processing minds to some crucial issues of epistemology, namely inferential internalism in the sense of Boghossian (1.2) as well as to issues of transparency, opacity and truth (1.3).

1.1. A basic response to the general open theism objection

To see that the general open theism objection is basically flawed consider a subject uttering the sequence of words 'I will read a book or I will write a letter'. Now in a quite natural way those words might be taken to express either a *change of mind* of the subject or, by contrast, a *single resolution* restricting the spectrum of planned activities to the two. This shows that a sequence of utterances (or actions) may be seen as a manifestation of either a sequence of thoughts or rather a single unchanging thought. It seems clear, moreover, that the scope of a single unchanging thought encompassing a sequence of words (or actions) may be lesser or greater. This means that minds can be more or less unchanging as regards various sequences of utterances or actions that manifest their states. In a similar vein, a series of chess movements might manifest either a single

comprehensive strategy of a skilled player or rather a series of unsteady tactical ideas of a beginner. Such considerations pave the way for the idea of a mind that does not change while comprehending very much, and, in the limiting case, comprehending absolutely everything.

In the aristotelian and neo-platonic traditions we thus find a contrast between two kinds of cognitive activity: *dianoia* or *ratiocinatio* which consists basically in transitions from one cognitive state to another; and *nous* or *intellectus* which consists in a simple grasp free from any sort of processing. God can be conceived as a pure *nous* or *intellectus*, enjoying the eternal non-processing cognitive activity. Human minds, by contrast, are typically busy with *dianoia* or *ratiocinatio*, involving transitions between various cognitive states; but they do partake, from time to time, in some forms of non-processing cognitive activity.

I do not dwell here on the details of this idea or numerous difficulties of various sorts it necessarily involves. My point here is that it is quite natural from a commonsense point of view to construct a hierarchy of more or less unchanging minds producing sequences of utterances or actions. What I discuss in some detail below, is the relevance of these commonsense ideas for some more sophisticated ideas of epistemology, namely the issues of inferential internalism (1.2) and the issues of transparency and truth (1.3).

1.2. Epistemology of non-processing minds: inferential internalism

The main claim I want to argue for in this section is that some weak form of internalism is true and that this weak form of internalism speaks in favor of the claim that humans do enact some non-processing cognitive activities.

Roughly speaking, the inferential internalism problem is the following: suppose that someone performs a *warrant-transferring* deductive inference and accepts some conclusion. Does it involve the knowledge on the part of the subject that the premises he appeals to do in fact *entail* the conclusion? An inference might be said to be blind if it happens to exemplify a valid logical scheme, but the subject is not aware of the validity of the scheme (or perhaps even unaware of the scheme

itself). So, in other words: are there *blind* and still *warrant-transfering* inferences?

The issue is widely discussed in contemporary epistemology (see, for example, Boghossian 2001, Boghossian 2003, and Boghossian 2014). It used to be the standard issue of the scholastic philosophy of logic as the question whether an inference that produces *knowledge* (as opposed to a mere opinion that happens to be true) involves the judgement on the part of the subject concerning the logical validity of the step made (*iudicium de bonitate illationis*); as regards the enormously rich late scholastic literature on the subject, see, for example, Polizzius 1675, 502–507; Maurus 1876, 614–617; Pázmány 1894, 46–51.

There are two main opposite answers to the question. The reliabilists (or externalists), both modern and scholastic, argue that there are really warrant-transferring and still *blind* inferences; or even that most of our important warrant-transferring inferences are blind. The internalists, by contrast, claim that the very logical form of the inference as well as its logical validity must be transparent to the subject if the inference is to be warrant-transferring or producing knowledge. The point is that there are severe problems relating to the form of logical insight that might play the role required by the internalists; in particular, it seems that the application of a general premise that any substitution of *modus ponens* is valid may seem to involve *modus ponens*, and in this way internalism may seem to involve some form of vicious circularity (see e.g. Boghossian 2001 and Boghossian 2003, 232–236).

Now it is here that scholastic internalist theories of inference appeal in a very interesting way to the idea of non-processing cognitive acts. They namely claim that an inference that would consist just in moving from an acceptance of a set of promises to the acceptance of the (validly inferred) conclusion would be indeed blind. It must be accompanied (or followed) by another act in which both the premises and the conclusion (that is: all the steps of the reasoning) must be grasped simultaneously – the former *as entailing* the latter; and this single indivisible act of grasping both the premises and the conclusion must also comprehend the entailment relation between them, or, in other words, the logical structure of the whole path as well as its logical validity (see e.g. Auriol 1953, 263–267; Maurus

1876, 215–216). Now this final act may be seen as an act of *understanding the proof as the proof*. And whereas the very moving from the acceptance of the premises to the (perhaps causally dependent on it) acceptance of the conclusion is a kind of processing, the final act comprises somehow all the steps of the former journey in a single view. Those acts may be compared to the series of views from the journey and a single view on the whole route respectively. This form of internalism is quite compatible with the view that, for some specific reason, the latter unifying view is accessible to our minds only *via* the former sequence of views; that, in particular, we are able to *understand* a proof only *via making* or *conducting* the proof and travelling along its steps. This is basically what makes our minds basically processing minds. But the point is that inference and the whole processing would be incomplete without the final unifying grasp.

Within a broadly aristotelian-neoplatonic view the ability to perform *processing* in our thought is called *ratio*, whereas the ability to have the final unifying grasp is called *intellectus*; human minds, as opposed to the divine, are basically processing minds, and perhaps we are not able to *understand* proofs and arguments without *conducting* them (without actual transitions from one step of the proof to another); in this respect we are *rational* and not *intellectual*. But to *conduct* a proof, to *proceed* in the right sort of way, is not yet to *understand* it; the proper role of the processing is to pave the way for the final unifying grasp; in this respect we have to be *intellectual*, and not just *rational* beings.

I do not claim that this traditional internalist picture of inference is free from mysteries or serious problems. I do claim, however, that it offers an interesting solution to the main problem of internalism sketched above; this solution may be juxtaposed with Boghossian's attitude towards the way we grasp the content of logical constants (see e.g. Boghossian 2003, 239ff).

Moreover, I do not claim or presume here that some strong form of inferential internalism is universally true. The point, however, is that it is striking that from time to time we actually do have this sort of *understanding proofs* that is comparable to seeing the whole route at once (as opposed to having a sequence of images during the journey). This sort of grasp seems crucial for understanding *reductio ad absurdum* proofs,

and, for other reasons, for understanding inductive proofs in arithmetic. Moreover, it seems to me that it is basically this sort of *understanding proofs* that explains our constant interest in *new* proofs of already proven theorems both in mathematics (for interesting examples see Dawson 2015) and in the scholastic tradition in philosophy. New proofs make it possible to see old theorems in new ways, in a new light or from a new perspective. And I think this sort of understanding proofs is indeed crucial for many cognitive attempts. Finally, I should claim that having this sort of *understanding* of a proof makes its result particularly valuable from an epistemic point of view.

This amounts to claiming that humans do succeed to perform – from time to time – some proofs that are not blind in the sense discussed above; I think that an overview of some activities in mathematics and related areas justifies the claim. Moreover, I claim that the idea of non-processing cognitive activity (or the unifying grasp) of the sort sketched above offers a very good key to the explanation of such cognitive performances.

There are clearly various degrees in this sort of unifying grasp, corresponding to the scope of things or proof steps that are comprehended in one view and the complexity of the logical structure. For example, the unifying grasp of Cantor's diagonal lemma in the standard way is quite demanding, but the unifying grasp of other uses of diagonalization, for example in the proof of Gödel's incompleteness theorem, involves the comprehension of much more and is much more demanding. In this sense one unifying grasp can comprehend more and be more powerful than another. What is ascribed to the divine mind in the classical theism is the maximum power and maximum comprehension. It grasps at once *all* the truths as well as *all* the logical connections between them. Moreover, unlike humans, God does not attain this sort of grasp *via* some sort of processing, and can understand a proof by grasping its logical structure without *conducting* that proof: he grasps all the structures of all valid proofs at once.

I am not claiming of course that the idea of maximum unifying grasp is free from difficulties; my point is just that in the epistemological context of inference internalism it makes sense to introduce the ordering of more

and less powerful unifying grasps, and thereby to pave the way for the idea of maximum unifying grasp.

1.3. Epistemology of non-processing minds: transparency, learning, and religious faith

There are two further specific contexts to which the above ideas of unifying grasp and weak internalism prove immediately relevant: the classical idea of human teaching and learning and the classical idea of religious faith (the latter being in some respects opposite to the former). In both contexts we encounter some problems of transparency, opacity, and truth. I present them in turn.

1.3.1. The classical picture of learning: transparency and trust

Since machine learning (and deep learning in particular) are among the most prominent topics in contemporary AI research, it might be interesting to sketch a classical account of learning involving the ideas of unifying grasp and the weak inferential internalism presented above.

The picture of learning to be found in Aquinas and other scholastic writers (see e.g. Quinn 2001) is obviously not applicable to many situations that might aptly be called learning (for example, language learning); yet I think it grasps well what is essential for some core situations we know from good courses of mathematics. In such a situation, the teacher conducts a proof (for example, the famous proof of Cantor's diagonal lemma) and understands it in the sense discussed in 1.2 above (has a sort of unifying grasp of all the steps of the proof, their logical structure and, last but not least, of the validity of the scheme). The teacher shows the pupil how to conduct the proof. The details of the procedure depend on the understanding of the proof on the part of the teacher. In this way, the teacher makes the pupil proceed in some specific way and conduct the proof for himself. The pupil conducts the proof – becomes engaged in some kind of processing, moving from one thought to another – and finally arrives at this sort of unifying grasp of the conducted proof that the teacher enjoys at the very outset. In particular, the pupil grasps

both the truth of the premises and the truth of the lemma, as well as the relevant logical structure of the proof and the validity of the scheme in one indivisible grasp. The teacher is causally responsible for the fact that the pupil succeeds in performing the relevant sort of unifying grasp. Without the final unifying grasp the conducting of proof and the learning would be futile and the pupil would fail to learn the content of the diagonal lemma. On the other hand, without the initial unifying grasp of the proof, the teacher would not be a teacher; he would not have played the role of one who *teaches* and would not be causally responsible for the knowledge of the pupil. It is basically possible to invent, conduct and understand a proof without this sort of causal role of any teacher; but we know that it is very difficult for humans to go on without teachers of this sort, especially as far as more sophisticated truths and more sophisticated proofs of them are concerned.

To repeat: admittedly almost everything that counts as language learning is not learning of this sort; and perhaps much that goes on even during math lessons does not qualify as teaching in this sense. Still it seems to me striking that much of what might count as good math education involves just this kind of teaching that is described above in the classical picture of Aquinas; it is a kind of teaching involving arguments which are really insightful and illuminating, just because they stem from the unifying grasp in the mind of the teacher.

Another crucial point is the way in which *transparency* is involved in the above classical picture of teaching. On one hand, the whole proof procedure is transparent to the teacher; on the other hand, the proof procedure performed by the teacher is transparent to the pupil – or perhaps rather it *becomes* transparent to the pupil just as he succeeds in his own understanding of the proof, in his own unifying grasp of all the steps of the proof, their logical structure and the logical validity of the scheme.

Finally, the role of transparency in this classical picture of learning is immediately relevant to trust issues in the learning situations. The teacher must be trusted at the very outset, and the reasons for the trust do not become transparent for the pupil until the pupil produces the relevant sort of the unifying grasp, the understanding of the proof. Yet

when the pupil succeeds to learn and to understand the proof, the reasons for the trust become evident for the pupil himself.

1.3.2. The classical picture of religious faith: opacity and trust

The idea of religious *faith* stems from some considerations of issues of trust and transparency. Religious *faith* (that we know, for example, from the story of Abraham) may be defined as believing God – believing what God says or reveals – for a very special reason: just because *as God* he cannot be mistaken, change his mind or lie (for some introductory remarks concerning the classical concept of faith see e.g. Siebert 2015). There is a room for religious faith when one assumes that what is said to him comes from God and on this assumption decides to believe it – just because of God's attributes (perhaps it should be added that it is a *real* case of religious faith when the assumption that what is said comes from God is true; I do not discuss here the details of how one is to decide whether it is really so). In this way religious faith is a kind of *trust* in God[8]. Now – *pace* the intuitions of open theists discussed in 1.1 – it is just the changelessness of the divine mind that is mostly relevant for the kind of trust specific to religious faith. It is essential for religious faith to assume that what is said comes in some way from the *changeless* God. In particular, the divine changelessness is relevant to faith in two basic ways. On the one hand, God is to be trusted just in virtue of his cognitive perfection: of the comprehensive grasp of every truth he enjoys; it is basically impossible that he has not yet arrived to some conclusion or has not yet taken into consideration some circumstance. On the other hand he is to be trusted just because he *cannot* change his mind; for example,

[8] One of the anonymous reviewers has pointed out that religious faith cannot be reduced to what is propositional; I do not want to exclude non-propositional aspects of faith. I do assume, however, that propositional contents expressed for example in the articles of the creed are basic aspects of religious faith. In particular I should claim that it is just some propositional content that might serve as a reason or justification of some religious activity or non-propositional religious attitude (as, for example, loving God or administering Sacraments). On the other hand, there are many sorts of relationship with God that are not propositional (visio beatifica would be the most obvious example); still, the point is that visio beatifica goes far beyond faith.

as Geach points out, God cannot fail to keep promises just because he cannot change his mind (Geach 1977).

The negative aspect of religious faith is that one who trusts in God and accepts what is said just because it is assumed to come from the changeless mind of God does not rely on any proof of the truth of what is said; no proof to be understood (in the way discussed in 1.2) is offered to him. One who believes God does not arrive at this sort of unifying grasp that God enjoys; the divine omniscience is not transmitted to the mind of a believer. Yet it is still true that the details of what is said by God do depend on the unifying understanding in the divine mind. Moreover, a believer can assume that they do depend on the divine unifying grasp and he can make it the basic reason for his belief in what is said. A believer might have some rough conception of what the divine cognitive perfection is – some rough idea of a universal unifying grasp – without having any specific insight into its content.

So, whereas in the teaching situation described in 1.3.1 the activity of the teacher's mind becomes transparent to the pupil just in the successful process of teaching, in the faith situation the divine mind remains non-transparent to the believer. The object of religious faith, as opposed to the object of learning, remains opaque for the believer. Yet it is crucial for the very notion of religious faith – for believing God – that the reason to believe in what is not transparent for the believer is just the assumption that the object of faith is transparent in itself, or that the performance of the divine mind is transparent to God in the fullest possible way. In this way the trust in God specific to religious faith is somehow grounded in the assumption of total transparency of the performance of the divine mind for God himself – which, for some basic reasons, remains inaccessible for us. It is this inaccessibility, rather than lack of transparency in general, that sets apart religious faith from the learning situation described in 1.3.1.

2. AI research *vs* non-processing minds: a general claim

My main claim concerning the AI research is that it is so exclusively focused on *processing* – on *transitions* from one state of a machine to another one – that any unifying grasp of the sort discussed in 1.1 or 1.2 seems to be basically outside its scope; it cannot either analyse or emulate it. It seems that research progress may well pave the way for quite new and incredibly sophisticated or rapid sorts of processing, but not for an analysis or emulation of the unifying grasp of all the steps in such an incredibly sophisticated procedure and their relationship. It may be within the scope of AI approach to make some machine conduct immensely sophisticated algorithms even beyond the limits of computability (see Syropoulos 2008). But the point is that it is basically beyond the limits of this sort of research to endow such a supercomputer with a unifying grasp of each of the steps of the relevant supercomputational procedure at once. The very idea of a successful cognitive activity that does not consist in any sort of processing at all – either within or beyond the limits of computability, either involving or not involving any sort of parallel processing or deep learning – seems quite foreign to this sort of research starting from simple Turing machine causes and proceeding towards more and more sophisticated and rapid sorts of processing, within, for example, more and more sophisticated neural networks.

Given the universality of my claim and the striking or even overwhelming richness of different approaches in the AI research, my claim may well seem very controversial. It seems to me, however, that it is surprisingly difficult, and perhaps impossible, to find any good counterexample. In other way, focusing on transitions from one machine state to another one seems to be a constant feature of the AI research program, a feature of the initial Turing machine idea just inherited by all the different strands in the AI research. To put it in other words, it seems to me that the basic tenet of the AI research in general is to reconstruct any cognitive activity in terms of *transitions* from one machine state to another one, and the AI research programmes as such are basically unable

to analyse the forms of cognitive activity that do not consist in processing information or in transitions from one cognitive state to another.

Now I was trying to show in 1.1–1.3 that the idea of non-processing minds and non-processing cognitive activity is intrinsically relevant to accounts of understanding proofs and grasping the validity of their logical schemes; in other words, some forms of weak inferential internalism seem to involve the idea of non-processing cognitive activity and non-processing minds. From this point of view the basic inability of the AI programme to analyse non-processing cognitive activities has serious systematic consequences. I think three of them are particularly important.

(i) It seems that AI research systematically tends towards some form of inferential or computational reliabilism or externalism, that is, to the claim that inability to see the validity the way we proceed from input data to the outcome is basically irrelevant from a cognitive point of view (see for example Durán forthcoming). Now I think that strong computational reliabilism is very debatable: not because computations someone makes without any grasp of the validity of the strategy or understanding of computation algorithms are void of cognitive value; but because just *sometimes*, in some *core* contexts, what we can reasonable require is just the understanding and the justification of the computation algorithm. To use a simple example: it is not a serious problem that most of the school children solve the quadratic equations using the algorithm they do not understand; but it is still vital that in some contexts one who employs the algorithm for quadratic equations grasps its validity of the sort I was discussing in 1.2.

(ii) A related tendency is to be observed in algorithm epistemology; some approaches obviously tend to focus on pretty weak senses of 'understanding an algorithm'. So, for example, von Eschenbach in his discussion of cases in which we lack the understanding of algorithms seems to use interchangeably 'the understanding of an algorithm' and 'the understanding of how the algorithm reaches its result' (von Eschenbach 2021). The point is that there are many sorts of understanding-how that might be quite valuable in many contexts, but are still quite far from the sort of understanding I tried to discuss in 1.2. To return to the quadratic equations example, someone might well spell out the algorithm and

repeat correctly all the relevant steps of a calculation, but still lack an understanding of why it is just this algorithm that may be employed to solve the equations. Again, the problem is not that everyone doing maths at school must have this sort of understanding; the problem is that it is vital that some people in some core contexts do know why we should calculate just like this.

(iii) In 1.2 I have argued for the *weak* inferential internalism: the view that *sometimes* we do succeed to have the described sort understanding a proof in a unifying grasp of all its steps, their logical interrelations and the validity of the logical scheme; and that such performances do have great epistemic value. Now it seems to me that because of the exclusive focus on *processing information*, the AI research tends to *systematically overlook* cases in which humans do succeed in understanding proofs; and cases in which humans do succeed in teaching and learning in the sense described in 1.3.1, that is in the transmission of the understanding of a proof; and, henceforth, to overlook some important examples of transparency of cognitive performances.

3. AI and the deep black-box problem

The black-box problem appears in circumstances in which the connection between the input and the output of some AI device remains hidden to us; we do not know how the relevant algorithm works. This is a deep problem particularly in the case of neural networks and deep learning, since their architecture may well seem strange and opaque from the point of view of logical structures that we are consciously aware of or consciously use. Moreover, the problem is deep in obviously serious or grave contexts, like criminal law or medical cases (see for example von Eschenbach 2021). Finally, it seems that the lack of transparency should have substantial impact on the degree to which we may rationally trust such AI results.

Brożek et al. (2024) claim that the black-box problem (discussed in particular in the context of the AI use within some legal procedures) is an unfortunate conflation of four distinct kinds of issues: the opacity problem, the strangeness problem, the unpredictability problem, and the

justification problem; they also claim that the very black-box problem is therefore superficial.

I agree that there is a complex of distinct issues in the black-box problem and in what follows I discuss the the first of the four problems singled out in Brożek et al. 2024, namely the opacity problem. As far as the problem of opacity problem is concerned, I defend – on the basis of some points established in 1.1–1.3 – the following three claims: (i) the *deep* opacity problem (or the *deep* black-box problem) should be distinguished from other sorts of the problem; (ii) the main problem with the contemporary debate on the opacity problem is that it systematically overlooks and ignores some cases that are crucial for understanding transparency and opacity; (iii) the opacity problem cannot be solved along the lines of various XAI projects.

3.1. Opacity and the deep black-box problem

According to (Brożek et al. 2024), the opacity problem in the legal contexts "does not seem to be a genuine issue" (Brożek et al. 2024, 430). The reason is that the real patterns of *human* decision making (in the minds of the judges) are in fact no less obscure and opaque so that human mind "seem to be black boxes *par excellence*" (*ibidem*, 430). This claim is quite close to a more general idea that our opacity objections and concerns are deeply biased so that we tend to overlook the fact that human minds are black boxes; or we have illusions of transparency and understanding of other human minds (see e.g. Bonezzi, Ostinelli and Melzner 2022).

Now it is clearly true that in some sense the performances of the minds of our neighbours (and of our own minds) often remain for us an inscrutable mystery. Moreover, it is obvious that there are many cases of illusion of transparency and understanding, including illusions of understanding of someone else's or our own cognitive performance. No doubt illusions of understanding in cases when one falsely thinks one understands well what is the reason for someone's claims are extremely interesting in epistemology.

On the other hand, however, those claims seem to me just to ignore two basic facts related to situations of learning discussed in 1.3.1 above.

The first fact is that from time to time we do manage to learn something from a good teacher in the way described above, so that we arrive at this sort of unifying grasp that our teacher was enjoying at the outset. To say that his mind is a black box for us would be to ignore the extent to which we understand him in such situations. The second fact is that good teachers of the kind discussed in 1.3.1 are basically able to tell whether on a given occasion we do understand the relevant proof or not, and whether we succeed in understanding the proof or not. In other words, they are able to inspect the minds of their pupils in this way, and to say that our minds are black boxes would be to ignore their ability to inspect them, even if our minds are still black boxes to ourselves. I do not pretend to know how exactly good teachers do inspect our minds that are black boxes for us; yet it seems clear to me that there are good teachers that are basically able to do so. Hopefully such cases are not quite foreign to the juridical contexts and that the story of the wisdom of Solomon tells something about the job of the human judges today, too. For one of the points of the Solomon story seems to be that judges should, and sometimes are able to, invent as exciting, illuminating and insightful arguments as great mathematicians sometimes manage to invent.

What is clearly relevant to squaring the two insights sketched above is a distinction of various sorts of opacity. Burrell (2016,1) offers a threefold division: (i) opacity as a sort of "intentional corporate or state secrecy", (ii) opacity as "technical illiteracy", and (iii) opacity arising "from the characteristics of machine learning algorithms and the scale required to apply them successfully". Now all these three kinds of opacity seem relevant in some contexts; yet it seems to me that there is another crucial kind of opacity relevant for the black-box problem in the AI contexts. This sort of opacity is revealed in the contexts discussed in 1.2–1.3 above. It is a sort of opacity consisting precisely in the fact that a machine – as opposed to very good human teachers, to hard-working pupils and to God – necessarily lacks the relevant sort of the unifying grasp of the whole procedure; it does not understand the proof or the algorithm.

The lack of the unifying grasp in the processing machine constitutes a particularly strong case of opacity, and what I call a case of a *deep* black-box problem. What goes inside the minds of good teachers is not

transparent to us until we learn what they already know; what is in the mind of God is basically not transparent to us; yet both good teachers and God enjoy unifying grasps of their cognitive performances so that their own performances are fully transparent to them. The performance of a machine, by contrast, is not just opaque to humans who use or even design it. The point is that the performance of the machine is necessarily opaque to the machine itself, too, just because it lacks the sort of unifying grasp described in 1.2; it does not understand the proof or the algorithm. That is what I call a deep black-box problem. We cannot trust the machine performance in the way we trust good teachers even before we learn something from them. And we cannot trust it in the way we trust in a divine authority, assuming that the performances of the divine mind are completely transparent to God, although they remain opaque to us.

I think it is obvious that the *deep* opacity or black-box problem is basically deeper than what we find in other versions of opacity problem. And if my main claim in section 2 is true, *every* AI machine performance is *necessarily* opaque for the very machine, so the deep opacity problem arises for any AI machine.

3.2. Opacity and the overlooking of transparency cases

The strong internalist expectation that any valuable cognitive performance should be absolutely transparent to the agent that performs it is clearly not reasonable. Those who claim that the black-box problem is exaggerated may well appeal just to the rejection of strong internalism. The point is, however, that some weak form of internalism seems to be true. In other words, we cannot reasonably require that every cognitive performance is absolutely transparent to one who performs it; but we cannot ignore or overlook the fact that *sometimes* we really manage to *understand proofs* we conduct: our proof procedures do become really transparent to us. Moreover, *sometimes* we really manage to *understand proofs* conducted by others: the performances of other minds become transparent to us. Finally, sometimes it makes sense to assume that someone's cognitive performance is completely transparent to the agent, although it remains (temporarily or not) opaque to us. In other words, we have a set of

transparency cases in math classes, in legal or medical disputes, and in many other contexts; we just cannot ignore or overlook the transparency cases if we want to understand human cognitive performances in those areas. The classical doctrine of non-processing minds outlined in section 1 offers some guidance for finding such cases and for their analysis.

The problem with the AI research is that, as I was trying to show, it seems systematically blind to this set of transparency cases. And those who claim that the black-box problem is exaggerated seem to turn a blind eye to them.

3.3. Opacity and the XAI strategies

An alleged solution to the black-box problem is some set of strategies developed within Explainable Artificial Intelligence (for its basic normative tenets see e.g. Zednik 2021). There are, of course, many interesting strategies of explaining algorithms of proofs; good teachers manage to apply them even in cases of pupils whose minds are black-boxes in a very strong sense. It seems to me, however, that the ability to offer good explanations of algorithms or proofs is rooted in the understanding the sense of the arguments in a pretty strong sense, that is in something like the unifying grasp of the whole proof sketched in 1.2. For example, it takes a very advanced understanding of proofs in some branch of geometry to offer scuh an explanation of some geometrical issue that is accessible for children. Now if my main claim from the section 2 (that the unifying grasp is basically foreign to machines) is right, and that what we have in the AI cases is a *deep* black-box problem, then it seems that the expected results of such strategies must be very modest: it seems that we cannot solve the deep black-box problem in this way.

Conclusion

To sum up: I have claimed that a common feature of a wide variety of AI projects is that they focus exclusively on *processing* information; and that, by contrast, *non-processing* minds and *non-processing* cognitive

performances play a prominent role in the classical theist framework, related to the issues of inference internalism (understanding arguments, proofs and algorithms) and a cluster of issues of transparency, trust and faith. In 1.2 and 1.3 I tried to offer a set of specific examples of epistemological issues related immediately to the idea of non-processing minds in the contexts of teaching and religious faith. I reject stronger forms of inferential internalism, yet I claim that *sometimes* we do manage to *understand proofs* or *algorithms*, and the classical framework makes us sensitive to such cases.

In this context I claimed in section 2 that the AI research, in spite of its richness and internal differentiation, focuses exclusively on processing information, and thus is basically unable to analyseor emulate cases involving *understanding* proofs or algorithms. This is a problem even if one rejects strong inferential internalism, because, as a matter of fact, there are numerous cases when humans do *understand* proofs or algorithms and their cognitive performances are really *transparent* to them.

In section 3, I applied those points to the black-box or opacity problem in AI. I defined the *deep* black-box problem as arising in situations in which a cognitive performance is opaque to the very agent of that performance and I claimed that it is ubiquitous in the AI. I argued that it is a grave problem just because there are crucial cases of understanding and transparency of proofs, arguments and algorithms both in mathematical practice and in legal or medical contexts. Finally, I argued that XAI is not able to offer a solution for the deep black-box problem.

References

Auriol, Peter. 1952. *Scriptum Super Primum Sententiarum. Prologue – Distinction I*, ed. By E.M. Buytaert OFM. St. Bonaventure, N.Y.: The Franciscan Institute.

Boghossian, Paul. 2001. "Inference and Insight." *Philosophy and Phenomenological Research*, 63: 633–40.

Boghossian, Paul. 2003. "Blind Reasoning." *Proceedings of the Aristotelian Society, Supplementary Volumes* 77: 225–48.

Boghossian, Paul. 2014. "What is inference?" *Philosophical Studies* 169: 1–18.

Bonezzi, Andrea, Ostinelli, Massimiliano, Melzner, Johann. 2022. "The Human Black-Box: The Illusion of Understanding Human Better Than Algorithmic Decision-Making." *Journal of Experimental Psychology*, https://doi.org/10.1037/xge0001181.

Brożek, Bartosz, Furman, Michał, Jakubiec, Marek, Kucharzyk, Bartłomiej. 2024. „The black box problem revisited. Real and imaginary challenges for automated legal decision making." *Artificial Intelligence and Law* 32: 427–40.

Burrell, Jena. 2016. "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data&Society*, https://doi.org/10.1177/2053951715622512.

Davies, Brian. 1993. *An Introduction to the Philosophy of Religion*. Oxford: Oxford University Press.

Dawson, John W., Jr. 2010. *Why Prove it Again? Alternative Proofs in Mathematical Practice*. New York: Birkhäuser.

Durán, Juan Manuel, Jongsma Karin Rolanda. 2021. "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI." *Journal of Medical Ethics* 47: 329–35.

Durán, Juan Manuel, forthcoming. "Beyond transparency: computational reliabilism as an externalist epistemology of algorithms." Forthcoming in *Philosophy of Science for Machine learning: Core Issues and New Perspectives*, edited by Juan M. Durán and Giorgia Pozzi, available at https://philpapers.org/archive/DURMLJ.pdf.

Eschenbach, Warren J. von. 2021. "Transparency and the Black Box Problem: Why We Do Not Trust AI." *Philosophy & Technology* 34: 1607–22.

Geach, Peter Thomas. 1977. "Can God Fail to Keep Promises?" *Philosophy* 52: 93–95.

Maurus, Sylvestrus. 1876. *Quaestiones Philosophicae*, t. 1: *Summulae et quaestiones prooemiales logicae*, Paris: Ch. Taranne.

Pázmány, Petrus. 1894. *Dialectica*, ed. S. Bognár. Budapest: Typis Regiae Scientiarum Universitatis.

Polizzius, Iosephus. 1675. *Disputationes philosophicae. Tomus Primus De Logica*. Palermo.

Quinn, Patrick. 2001. "Aquinas's Views on Teaching." *New Blackfriars* 82: 108–20.

Roosta, Seyed H. 2000. *Parallel Processing and Parallel Algorithms*. New York: Springer.

Siebert, Matthew Kent. 2015. "Aquinas on Believing God." *Proceedings of the Aristotelian Catholic Philosophical Association* 89: 97–107.

Syropoulos, Apostolos. 2008. *Hypercomputation. Computing Beyond the Church-Turing Barrier*. New York: Springer.

Van Leeuven Jan, Wiedermann Jiří, "The Turing Machine Paradigm in Contemporary Computing." In *Mathematics Unlimited – 2001 and Beyond*, edited by Björn Engquist and Wilfried Schmid, 1139–56. New York: Springer.
Zednik, Carlos. 2021. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34: 265–88.