AI and Ethics
https://doi.org/10.1007/s43681-022-00152-w

ORIGINAL RESEARCH

Should explainability be a fifth ethical principle in AI ethics?

João Figueiredo Nobre Brito Cortese[1,4,5] · Fabio Gagliardi Cozman[2,6] · Marcos Paulo Lucca-Silveira[1,3] · Adriano Figueiredo Bechara[1]

Received: 17 November 2021 / Accepted: 2 March 2022

Abstract
It has been recently claimed that explainability should be added as a fifth principle to AI ethics, supplementing the four principles that are usually accepted in Bioethics: Autonomy, Beneficence, Nonmaleficence and Justice. We propose here that with regard to AI, on the one hand explainability is indeed a new dimension of ethical concern that should be paid attention to, while on the other hand, explainability in itself should not necessarily be considered an ethical "principle". We think of explainability rather (i) as an epistemic requirement for taking into account ethical principles, but not as an ethical principle in itself; (ii) as an ethical demand that can be derived from ethical principles. We do agree that explainability is a key demand in AI Ethics, with practical importance for stakeholders to take into account; but we argue that it should not be considered as a fifth ethical principle, to maintain a philosophical consistency in the organization of AI ethical principles.

Keywords Explainability · Explicability · Interpretability · Explainable AI · AI Ethics

1 Introduction

The need for "explainability"[1] in Artificial Intelligence (AI) ethics has been emphasised by several recent works and declarations on AI ethics [8]. But what is the role of explainability in AI Ethics: should it be considered an "ethical principle" [10]?

In this paper, we wish to discuss some notions of Explainability in AI and address the relationship of Explainability to the four principles of Bioethics (Respect for Autonomy, Beneficence, Nonmaleficence and Justice, as proposed by Beauchamp and Childress [2]) in the context of AI Ethics. Floridi et al. [10] revisited 6 declarations on AI Ethics, which together present 47 principles.[2] According to Floridi and co-authors, these ethical recommendations could be embraced by five principles for AI Ethics, based on the four listed principles of the Principlist bioethical approach, plus a new one: a *fifth principle* of "Explicability".[3] Although *we do claim that explainability is desirable for AI ethics* in several situations, we will argue that it is not appropriate to consider it as a "fifth principle".

Some authors have already made claims against the understanding of explainability as an ethical principle.

✉ João Figueiredo Nobre Brito Cortese
joao.cortese@usp.br

1 Núcleo de Bioética, Fundação José Luiz Egydio Setúbal, Av. Angélica, 2.071, Consolação, São Paulo, SP 01228-200, Brazil

2 Center for Artificial Intelligence, C4AI at Universidade de São Paulo, Av. Prof. Lúcio Martins Rodrigues, 370, Butantã, São Paulo, SP 05508-020, Brazil

3 São Paulo School of Economics, FGV, Rua Itapeva, 474, São Paulo 01332-000, Brazil

4 Departamento de Fisiologia, Instituto de Biociências, Universidade de São Paulo, R. do Matão, 321, Butantã, São Paulo, SP 05508-090, Brazil

5 Ibmec-SP, Alameda Santos, 2356, Jardim Paulista, São Paulo, SP 01418-901, Brazil

6 Department of Mechatronics and Mechanical Systems, Escola Politécnica, Universidade de São Paulo, Av. Professor Mello Moraes, 2231, Butantã, São Paulo, SP 05508-030, Brazil

[1] We will adopt the term "explainability"; the terminology of the concept is contested, as we will see below. We will use other terms only when commenting on other authors' ideas, in which case we preserve the terms used by them.

[2] Using the term "intelligibility", the House of the Lords report from 2018, for instance, already suggests "five hierarching principles", that one could associate in some way to the four principles of bioethics and to explainability [18].

[3] Floridi et al. [10], use the term "explicability", whereas the more common term in the domain is "explainability". On "explicability" as a fifth principle, see also [11].

Published online: 28 March 2022

To begin, it is not a consensus at all that AI Ethics should be organised by principles. B. Mittelstadt proposed that "Principles alone cannot guarantee ethical AI" [24]. From this perspective, principles should not be at the core of ethical theories in AI. Accordingly, there would be no motivation to debate whether or not explainability is a new principle.

Furthermore, even if one accepts principles in AI ethics, there are reasons to discuss whether and how an approach by principles should be adopted, and what are its limitations [39], as well as there is significant literature on which principles should be adopted in Bioethics itself [17]. However, this will not be addressed in this paper. It is thus "conditionally" that we analyse the relation between explainability and the four other principles of Principlism, as other foundations for AI Ethics are surely possible.

Regarding specifically explainability as a principle, some arguments have been presented against it. S. Robbins [30] argued that a "principle" of explainability is "misdirected": one of the arguments for this would be that AI applications with low risk would not require explainability—but which applications can be easily characterised as being of low risk? Another important contribution, by Ursin et al. [38], argues that "explicability" is not a fifth principle in medical AI ethics, because the four principles of bioethics already encompass it. We later come back to the work by Ursin et al. as we deepen their analysis from a conceptual point of view.

The discussion presented in this article is centred on the *role* of explainability in an ethical framework based on principles: in short, even if explainability is a *desideratum* in AI, it is not necessary that it should be considered for that reason as a fifth ethical principle. We will show that the claim for explainability can be considered (i) as derived from the principles already established in the four principles framework; and (ii) as an epistemic requirement for ethical principles.

2 The many concepts around explainability

Concerns about the explainability of artificial intelligences are not a new phenomenon. In 1973, G. A. Gorry wrote, regarding a system for clinical decision making:

> "If experts are to use and improve the programme directly, then it must be able to explain the reasons for its actions. Furthermore, this explanation must be in terms the physicians can understand". [12]

A variety of explanatory schemes for expert systems have been investigated, often by explicitly presenting to the user the rules employed in reasoning chains; the reader can consult a 1988 review paper on experts systems where the authors declare: "One of the defining criteria of expert systems is their ability to 'explain' their operation" [4]. Some explanations generated at the time by expert systems described how a decision was reached; others tried to articulate why the system asked for some information. Since then, there has been steady interest in explanations associated with formalisms for knowledge representation and reasoning. For instance, there has been a decades-long effort to define the kinds of explanations one can extract from statistical models known as Bayesian networks: sometimes the purpose is to explain the reasons why a Bayesian network is structured in a particular way, while in other cases the goal is to explain the probabilistic calculations that generate a result [20]. No unified theory of "explicability" or "explainability" can be found in that literature.

Today the debate around explainability is hotter than ever in AI circles. This is likely a consequence of AI extraordinary progress in the last decade or so, a progress that has been grounded on pattern extrapolation based on ever-increasing data sources, along with ever-increasing computing power. The pragmatic success of AI, and more precisely of machine learning, has raised a few flags: *prima facie*, society cannot tolerate artificial intelligences that discriminate, that are prone to catastrophic mistakes and to accept evil—risks that are in some cases related to a lack of understanding or interpretability. Such artificial intelligences are said to rely on "black boxes", that is, complex functions, learned from huge datasets, that are too opaque and inscrutable. Hence the call for fairness, transparency, accountability, interpretability, intelligibility. The expression "Explainable AI", originally coined in the context of a DARPA program [7], now appears in all major conferences and venues that deal with AI. Yet the precise definition of explainability and its associated concepts still eludes consensus.

Floridi et al. [10] adopt the term "explicability" and take it to have an "intelligibility" dimension and an "accountability" dimension. Other authors have taken intelligibility to be a general concept that embraces the broad notions around explainability [18]. Most of the literature instead adopts the term "explainability". Often "explainability" and "interpretability" are used interchangeably as synonyms, for instance, Miller [22] does so explicitly. Miller presents an often cited definition for interpretability: "the degree to which an observer can understand the cause of a decision" ([22], p. 8; one must take causation in a weak sense in such a definition). Thus he takes interpretability to be a matter of degree; we might even conceive of an approach that quantifies interpretability explicitly, not as a binary question, but rather as a continuous quantity—this may be useful, for instance, to distinguish situations where a low interpretability level matches a system with corresponding low potential for damage. Such a path may be fruitful in future work; we do not deal with such possibilities in this paper as

the question as to whether explainability is a principle or not does not seem to rely on it having a binary aspect or not.[4] It should also be clear that explainability has a potential aspect: not everything must be explained, but one should be able to provide an explanation if this is requested.[5]

An influential piece by Lipton [21] addresses "transparency" concerning how AI models work. By transparency Lipton means, roughly, that the user can grasp how the model works, perhaps just looking at its code, at its parts, its algorithms, perhaps even extracting a causal mechanism. We might refer to such a notion as "intrinsic explainability", one of the two interpretability techniques categories proposed by Lipton; the other is "post-hoc explainability". In the latter, one is concerned with additional information of interest about the black box.

In addition, we should differentiate between explanations and justifications: quoting again Miller, a justification "explains why a decision is good but does not necessarily aim to give an explanation of the actual decision-making process" [22]. So we might imagine a system whose decisions are discriminatory and that offers persuasive justifications for its actions, perhaps by resorting to all sorts of rhetorical devices, while never coming up with honest descriptions of its internal (discriminatory) processes. Even though justification can be of major importance in ethical discussion, it is outside our scope in this paper, and it also raises problems concerning the public embracing of principles in general, such that it is not specific to explainability.

However, the same doubts regarding the honesty of justifications can be raised in connection with explanations. Might a system make a decision and then provide an explanation that selectively describes elements of the decision-making process, so that its explanations always resonate with users without ever clarifying what is really going on? This is clearly possible; indeed if the designers of artificial intelligences are strongly pressed to always provide explanations due to ethical or even legal constraints, we can imagine that at least some might produce explanations that always corroborate decisions. Now, if the system itself is opaque, how is one to question such sanctioning explanations? But even if explainability may lead to explanations that are satisfactory to the user but not faithful to system behaviour, one might argue that the pursuit of explainability and transparency would reduce the risk of deception and encourage honest explanations.

The current literature on Explainable AI investigates a number of approaches to attain interpretability. Of course, one can always build a simple system. Alternatively, one may have a complex system, perhaps one based on large neural networks, coupled with a tool that indicates which features of the input, or perhaps with elements of the system, mostly affect a decision. For instance, the tool GradCAM highlights the parts of an image that, if changed, led to the largest change in the output produced by a neural network [34]. Another possibility is to build an interpreter that learns to explain decisions by examining the inputs and outputs of the opaque system; this is sometimes referred to as an agnostic strategy [28]. Here the particular elements of the underlying complex system do not even matter; all that matters is its input–output behaviour. Such a scheme might be prone to explanations that simply move towards a foregone conclusion.

Most existing techniques in the literature focus on *local* explanations; that is, the goal is to explain a particular decision and not the global behaviour of an opaque system. While a local explanation may be relevant to an end user interested in a particular decision that affects her, a global explanation may be useful for instance to an auditing body interested in the behaviour of a system for automated credit analysis. One should suspect that a global explanation is usually directed to a sophisticated user, say by offering a graph with statistical metrics on the relative impact of observations.

Indeed, it seems that most current techniques in Explainable AI are geared to data scientists rather than to the general public. Many techniques only make sense if the underlying structures, say neural networks, are already well understood, and even the agnostic techniques produce results that can only be understood given some experience. It seems fair to say that, as the field evolves, there will be at some point a set of tools for the expert user, and a different set of tools for the general public. Some of these tools will be part of black box systems (so as to turn them into "white-boxes"), while other tools will be used to obtain insight into black boxes (so that the combination of a black box and external tool becomes understandable). In any case, explanations must be generated in distinct ways for expert users and for the general public.

From a broader perspective, it is known that explanations are dependent on the receiving user; for example, the explanation for a medical diagnostic varies if the target is a patient, a doctor, or a hospital auditing authority[22].[6] In any case, the difference between techniques that aim at expert

[4] We thank a reviewer who suggested this possible research path to us.

[5] This is similar to accountability of a company in some aspects: it should not declare and explain everything it does, but in a particular case (for instance being judged for a crime), it should be capable of showing relevant information and justifying its decisions.

[6] Besides, one should keep in mind that several people are involved in the process of the development of an algorithm. As Coeckelbergh [6] says, AI is a problem of "many hands", in the sense that "many people are involved in technological action", which makes responsibility attribution difficult in this case—we could think that to know to whom address explainability and how to do is also made harder by this "many hands" aspect of AI.

users such as data scientists and techniques that aim at the general public is worthy of emphasis as they usually have different functions and consequences, with different kinds of ethical demands.

3 Explainability as a fifth principle for AI ethics

Despite the little consensus one can find around explainability and all interrelated concepts, Floridi et al. [10] propose "explicability" as a fifth principle of AI Ethics. The principle of "explicability", according to them,

> "complements the other four principles: for AI to be beneficent and nonmaleficent, we must be able to understand the good or harm it is actually doing to society, and in which ways; for AI to promote and not constrain human autonomy, our 'decision about who should decide' must be informed by knowledge of how AI would act instead of us; and for AI to be just, we must ensure that the technology—or, more accurately, the people and organisations developing and deploying it—are held accountable in the event of a negative outcome, which would require in turn some understanding of why this outcome arose".

We agree with the authors to the extent that the four "canonical" principles of Bioethics, when applied in the AI ethics field, already involve "explicability". Indeed, explainability is an ethical *demand* that is specific to AI—at least as to the specific way in which one generally talks about "explicability" in AI. But if explicability *always* comes together with other principles, as this passage seems to imply (we "*must*" "understand", "be informed" or "ensure"), it is conditioned to them, and not a principle in itself. Indeed, the evaluation of whether "explicability" should be a principle concerns the question as to whether it has a value that is not conditioned to the other principles—we will come back to this.

In April 2019, the European Commission's High-Level Expert Group on Artificial Intelligence presented a document under the title "Ethics Guidelines for Trustworthy AI" [15], in which, following Floridi et al. [10],[7] the existence of four ethical principles for AI ethics is proposed: "respect for human autonomy", "prevention of harm", "fairness" and "explicability". We can clearly see a relation with the four bioethical principles, the principle of beneficence not appearing here.[8] About "explicability" as a principle, they write:

> "Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions—to the extent possible—explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate" [15]

Several questions appear here. Firstly, to whom the explanation of AI should be addressed? The layperson? The programmer? An user that is an expert in a domain (e.g. a doctor?). As we said above, generally strategies of explainability are directed mainly to people involved in the design of algorithms, whereas public demands such as that of the High-Level Expert Group on Artificial Intelligence have a tendency to demand explanations for all stakeholders. How to unite these two extremes is a task still to be performed. Furthermore, what is the relation between "transparency" and "explicability" in the passage quoted? This is not clear either. While one can understand that both are desirable for AI to be ethical, how to conceive the relationship between them[9]? This is another question that could be elucidated in future works, both in technical AI literature as well as in

[7] An explicit quotation is made at page 11 of the document [15].

[8] The document does not absolutely ignore beneficence, since it proposes that the implementation of a Trustworthy AI "entails seeking to maximise the benefits of AI systems while at the same time preventing and minimising their risks" [15]. One should recall that at the Belmont Report [1] one could find the *three* principles of "Respect for Persons", "Beneficence" and "Justice". Beauchamp and Childress [2] defend that we should consider two different principles, Beneficence and Nonmaleficence, having a total of four principles.

[9] One could think that if a system is transparent, it does not, for that very same reason, need to be explained—if this is the case, simply saying that both transparency and explainability are desirable is not enough.

AI ethics.[10] In any case, it is clear from this passage that for the High-Level Expert Group on Artificial Intelligence "explicability" is desirable—to which we agree—and is a principle—to which we disagree.

Ursin et al. [38] concluded that, in the case of medical AI (more precisely radiology), "as long as the properties of explicability are covered by at least one of the four principles of biomedical ethics, explicability may not have to be recognized as a free-standing principle" [38, p. 151]. We agree with the conclusions of their work: in the case of AI, explainability is not a fifth principle side by side with the four principles of bioethics, and at the same time the search for explicability is important, even if it is not a principle. However, the present work has an important methodological difference from theirs: the discussion made in [38] is based on a survey on the reasons alleged by AI ethical declarations for radiology; our path will be rather to consider *a conceptual analysis of principles*. Moreover, we extend the claim that explainability is not a principle for AI ethics in general (and not only for medical ethics in the case of AI).

If we come back to the core of Beauchamp and Childress' Principlism [2], a moral principle is defined as being a general norm or group of norms that are used to guide and evaluate conduct. In this sense, respect for autonomy would be, according to them, "a norm of respecting and supporting autonomous decisions"; nonmaleficence, "a norm of avoiding the causation of harm"; beneficence, "a group of norms pertaining to relieving, lessening, or preventing harm and providing benefits and balancing benefits against risks and costs"; justice, "a cluster of norms for fairly distributing benefits, risks, and costs".

But which norm would express the supposed explainability principle? Of course, one could say that the recommendation to "develop algorithms in the most explainable possible way" is a guiding directive for programmers. In this sense, there would seem to be no problem to consider explainability as a principle. Nevertheless, what is the point in asking for explainability? To make these algorithms accountable by making clear how responsibilization should be dealt with. But what are we responsible for? Not doing harm, being just, doing good: these norms or groups of norms are something desirable to guide or evaluate conduct, but explainability seems to be rather a requirement for, or a derivation from, these general norms, and not a general norm in itself (as we will argue, one could a priori think about a non-explainable model that does not violate any ethical requirement). Furthermore, from an Aristotelian point of view, these principles seem generally to be more "final ends" than explainability [5]. In other words: we think that explainability *is important* to AI ethics—but even though it seems to be *desirable* in this domain, we cannot logically conclude that explainability is an "ethical principle".

"A principle is a fundamental standard of conduct from which many other moral standards and judgements draw support for their defence and standing" [3]. If this is so, could one say that explainability is a standard of conduct from which other moral standards are derived? Or rather explainability could perhaps be derived *from* the other four principles? Therefore, explainability alone does not directly or per se affirm moral obligations that must *always* be acted upon.

Floridi et al. [10] propose that the "explicability" is a principle "both in the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question: 'who is responsible for the way it works?')". We agree with them that, regarding the ethical relevance of explainability, both these senses should be considered: talking about explainability is precisely talking about accountability derived from epistemological aspects.

But we do not think that this necessarily makes explainability an ethical principle, for two reasons: (i) explainability should not always be looked per se, but rather it can be demanded as a requirement for ethical principles according to the circumstances;[11] (ii) moreover, it could be understood as morally relevant, but *derived from* ethical principles. In both cases, consequently, explainability seems to be an element that, although having ethical implications, is not for that reason an ethical principle.

In addition, it should be noted that significant previous work has gone in this direction. Tsamados et al. [36] recognize that "transparency" (a term, as we saw, closely related to explainability) is rather a requirement than an end in itself. Floridi [9] addresses transparency in the context of an "infraethics". Floridi and Turilli [37], finally, write that transparency is not an "ethical principle in itself but a pro-ethical condition for enabling or impairing other ethical practices or principles", and, following this, a recent work [13] accepts explainability as "a pro-ethical condition for enabling or impairing judgments of *beneficence*, *nonmaleficence*, *justice*, and *autonomy*". To say thus that explainability is a kind of *requirement* looked for is to agree with those proposals.

Coeckelbergh [5], discussing moral agency from an Aristotelian point of view, affirms that two conditions are needed for responsabilisation: a "control condition" and an "epistemic condition". Thus, attributing responsibility for an

[10] One should not think that for a principle to be adopted the nature of its concept should be clear. This is not our argument for explainability not being an ethical principle: as we will see, the problem lies rather in the ethical demand for explainability, that we see as originating (when it is the case) from the other principles.

[11] We will address below the fact that, in some cases, explainability can be valued per se—but we will show that even though it is not for that reason a principle.

action requires the agent both to be free to accomplish the action or not (control condition), and to know what is the quality of the action that is made (epistemic condition). That is to say that one cannot be blamed for something which she or he is forced to do, or without knowing that a harm is concerned by this action. More broadly, the epistemic condition is conceived as concerning "whether the agent's epistemic or cognitive state was such that she can properly be held accountable for the action and its consequences", such that it is equivalent to asking "was this person aware of what she was doing (of its consequences, moral significance, etc.)?" [32].[12]

Explainability can be understood as a kind of *requirement* for ethical principles in AI. It is related to the notion of epistemic condition, but as this last concept is rather a condition for the responsibility of an agent, we adopt the more broad concept of an epistemic requirement, which is a condition, regarding knowledge, to the fulfilment of AI ethical principles. Indeed, in certain cases, one must understand a system to know if it could do harm. However, this requirement is necessary from an *epistemic* point of view, as it helps the other principles (in which sense we could agree with the choice of the verb "complements" by Floridi et al. [10]); but even if there is a certain kind of dependence here, and an ethical demand derived from this, explainability does not become a principle in itself for that reason (in which sense we disagree with these authors).[13] Finally, we do not claim that *all* occurrences of principles in AI ethics claim for explainability as a requirement, such that it is not unconditional.

With respect to explainability as *derived* from principles, one should note that principles are general, and to make them come into practice, so as to affect particular cases, requires *interpretation*. This is brought by several philosophical traditions, but in Bioethics one could think broadly on the work of H. Richardson [29], who understands interpretation as a modification of the content of a norm. So, principles such as nonmaleficence or justice may imply a demand for explainability, but only as a result of their interpretation in concrete cases—and, for that reason, explainability is not a principle because, by definition, a principle is whatever is primary.

We can thus see how difficult it is to conceive explainability as a principle. An ethical theory could take explainability in consideration, but explainability should perform different roles other than that of a principle.

Yet what is its precise relation to the other principles? In what follows, we will try to understand how explainability is related to ethical principles inspired from Bioethics' Principlism. We will not present here a definition of explainability—as we said, there is no consensual definition for it even at a technical level, and neither at an ethical one. Nevertheless, we can think that the supposed trade-off between explainability and other ethical principles can be analysed from the presentation of the concept by Floridi et al. [10] if one wants to keep in mind a reference, but we think our arguments will go beyond that, covering also other similar conceptualizations of explainability.

4 Explainability and its relationship with the four principles

Let us take a look at how explainability could be related here to autonomy, nonmaleficence and justice. Finally, we will consider separately the relationship of explainability and beneficence, given the trade-offs in which supposedly explainability would be involved as an ethical principle opposed to beneficence. In all cases, we will see that there is no need for a fifth ethical principle.

4.1 Explainability and autonomy

According to Floridi et al. [10], "autonomy" in AI Ethics is more complex than in Bioethics. In AI, one could find what they call " 'meta-autonomy', or a 'decide-to delegate' model: humans should always retain the power to decide which decisions to take" [10]. A great deal could be said here regarding this principle. Let us just say that for these authors, the principle of autonomy in AI Ethics would be associated with the possibility to decide whether to delegate or not a decision to a machine.

In Bioethics, one deals with "respect for autonomy" as a principle regarding the autonomous choices of persons, in particular "to examine patients', subjects', and surrogates' decision making in health care and research" [2]. Beauchamp and Childress present an important distinction regarding this principle: "as a *negative* obligation, the principle requires that autonomous actions not be subjected to controlling constraints by others. As a *positive* obligation, the principle requires both respectful disclosures of information

[12] Coeckelbergh does not say that we can attribute responsibility to an AI as a moral agent, but rather that "only humans can be responsible agents" [5]. However, he claims that the aristotelian "philosophical analyses of ignorance"—in particular, to know the technology you are using—"can guide discussions about knowledge problems with AI" [5]. In this sense, one can say that having an explainable algorithm can *contribute* to giving the agent an epistemic condition.

[13] Herzog [14] claims that, for Floridi et al. [10], "the principle of explicability is introduced as enabling the other principles of 'beneficence', 'nonmaleficence', 'autonomy' and 'justice', rather than being a primary principle". We cannot agree with this reading of the passage: if it is not a principle, how could the authors write in the continuation of the article about the "addition" of a "principle" concerning explicability? Herzog does not explain how one should differentiate between a principle and a "primary principle".

and other actions that foster autonomous decision making" [2]. One could think that regarding the application of this principle to AI Ethics, its *positive* aspect would be related to making available the relevant information to ensure understanding to the person that decides to use an AI system or not.

Both in research ethics and in clinical bioethics, one important concept related to autonomy is informed consent. It is a complex notion, with several different understandings; but even if its definition is not clear, one should note that the Belmont Report associates autonomy[14] to informed consent—that is to say, we understand that consent can only be autonomous if someone is informed regarding the relevant aspects about what is involved. In the same way, one could say that the explainability of an algorithm is a *required feature* to evaluate autonomously whether to agree with its use or not. One could think here of the notion of a "right to explanation" of the user, regarding some information relevant for a decision.

This does not mean that the puzzle of the role of information is solved, since knowing what is relevant for informed consent is a problem in itself. Amongst the elements needed for informed consent, Beauchamp and Childress present "information elements", which are "disclosure (of material information)", "recommendation (of a plan)", and "understanding" of the two previous.

> "No general consensus exists about the nature and level of understanding needed for an informed consent, but an analysis sufficient for our purposes is that persons understand if they have acquired pertinent information and have relevant beliefs about the nature and consequences of their actions". [2]

Although dissent can exist about what constitutes an adequate degree of understanding for informed consent, could the "intelligibility" part of the alleged "explicability principle" proposed by Floridi et al. be included here? What they label "explainability principle" is actually the epistemological dimension *presupposed* by the principle of Respect for Autonomy when the last is applied in the AI field, such that explainability is not a principle, but rather a requirement.

With respect to AI, one could think for instance of the role of recommendation systems for users. A classical form of ethical dilemma can occur between Beneficence and Respect for Autonomy: should we opt for a recommendation system which has "better" output in general, despite violating individual autonomy, or to preserve autonomy and to have worse general results? More specifically, recommendation systems are subject to problems related to diversity and serendipity [26], concerns that are related to a person's autonomy (the fact of adding "more of the same" as an "ossification" being opposed to the possibility of the introduction of "novelties").

Of course, there is an important difference between what is claimed for explainability in AI and the informative component of understanding in the principle of Respect for Autonomy. While in the latter the demanded understanding concerns the research subject/patient, and the explanation should enable him or her to make a decision, in AI decisions the situation is more complex. In AI, the question seems to be whether *anyone* can understand what is happening. What is of interest here is to consider the *explainability in itself*: could *someone* understand the reasons why the algorithm made this recommendation? This brings us back to the problem of to whom the explanation of explainability is addressed: to the programmer? To the specialised user (e.g. the doctor)? To the layperson (e.g. the patient)? In correspondence, whose autonomy is considered? We do not claim to solve this problem as related the definition of explainability here.[15]

In any case, talking about a respect for autonomy in some way already presupposes some kind of "intelligibility", and thus of "explainability". Explainability is thus not necessarily a new "principle", even if desirable for autonomy.[16]

4.2 Explainability and justice

As Floridi et al. note [10], the principle of justice in Bioethics is typically discussed in connection with the distribution of resources—that is, the area of distributive justice that discusses the adequate distribution of benefits and burdens among people with divergent claims [31]. These problems find parallel situations on AI ethics.

In this section, we intend to analyse the moral requirements of justice and the conceptual connection of these with explainability. Beauchamp and Childress [2] recognize that

[14] The Belmont Report deals with "respect for persons", which was later associated with an "respect for autonomy" principle [1].

[15] Regarding the fact that explainability should be presented to stakeholders with different backgrounds that interact with the system at different moments, one could think for instance of an approach to explainability that tries to present a "minimal" explainability, comprehending demands that appear for all users; another solution would be to differentiate explanations according to the stakeholders (Herzog [14], for instance, claims that "we should not only be interested in the developing party as the responsible entity, but also in the commissioning, deploying and, ultimately, the utilizing parties.", thinking in a conceptualization of "explicability" that encompasses several stakeholders). However, this question will not be further developed here.

[16] We should note that, despite adopting a different theoretical approach, Mirbabaie and colleagues [23] propose a framework in which "explainability/explicability", by means of the concept of "transparency", appears under the Autonomy principle (the four principles of Bioethics being assumed).

no single principle is capable of addressing all problems of justice: there is more than one principle of justice. Several terms—such as fairness, merit, capabilities, well-being, and opportunity—are used by philosophers to materialise the moral demands associated with justice [27, 35]. However, as Beauchamp and Childress [2] claim, all theories of justice accept a common minimum requirement, traditionally attributed to Aristotle: equals must be treated equally, and unequals must be treated unequally. This formal princíple of justice, which does not provide a material criterion for determining what or who are "equals", is the most basic standard for evaluating issues of distributive justice. We argue that even the least demanding principle of justice presented in Principlist ethical approach—the formal principle of justice—requires explainability as a condition for its fulfilment. Requirements that come from principles of justice entail that AI is *constrained* by demands for explainability, so that algorithmic decision making would be just, unbiased and respecting the rights of people, especially those in disadvantaged groups. In other words, the most basic demand for justice, according to which we must treat equals equally, when applied to the field of AI, requires explicability.

Thus, ethical requirements related to explainability seem to be a central topic of other demands of justice. To know whether an AI decision, output or recommendation does not violate a principle of justice, we need to "explain" it. As Robbins notes, "the opacity of the algorithm prevents us from knowing whether it is unethically biased" [26]. Thus, the use of black box algorithms may very likely be restricted in many fields because it is not possible to verify whether they violate a basic principle of justice.

Biases can come from several sources: from the selection of the data for the training of the algorithm, from the data itself, from the programmer, even from society values; but it can also come from the algorithms themselves, and in this sense explainability is a highly relevant aspect [6]. To know whether an algorithm respects this formal principle of justice, some degree of explainability is required. However, if it can be affirmed that the use of biased databases concerns a particular problem of justice in AI, this is also the case when the algorithm itself is biased, and in particular it could be difficult to identify the emergence of bias in an algorithm without explainability. Thus, the lack of explicability is potentially associated with violations of justice. Consequently, to talk about justice in AI we need explainability, since some degree of explainability becomes a minimum necessary requirement for evaluations of justice in AI. Thus, explicability could be understood not as an unconditional ethical principle, but as an epistemic requirement demanded by all principles of justice, as it is a requirement of the formal principle of justice, the common minimal requirement to all theories of justice.

4.3 Explainability and nonmaleficence

Floridi et al. [10] present nonmaleficence ("do no harm") in AI Ethics as for instance preventing the violation of people's privacy, and more generally as preventing both accidental ("overuse") and deliberate ("misuse") harms.

How is this principle understood in Bioethics?

> "The principle of nonmaleficence obligates us to abstain from causing harm to others. In medical ethics this principle has often been treated as effectively identical to the celebrated maxim *Primum non nocere*: 'Above all [or first] do no harm' ". [2]

But what about exposing someone to the *risk* of harm?[17] When dealing with uncertainty, the question becomes more complex, and one can consider invoking a precautionary measure to justify not using a technology. On the other hand, a potential benefit can exist in case one decides to use the new technology. As uncertainty pervades decision making, we will never be sure of not doing harm. How then to address the risks? Knowing *reasonably* how a system works and what are the odds of it making harm: knowing its possible unintended consequences.

> "Usually programmers and users know what they want to do with the AI. More precisely, they know what they want the AI to do for them. They know the goal, the intended consequences; Aristotle would say the end. However, users of AI are not necessarily aware of the non-intended consequences and moral significance of what they do". [5]

But knowing the risks of possible unintended consequences couldn't be directly related to the possibility to, in some measure, explaining the work of a system? Here, once more, we think that, in the cases it is required, explainability is not an extra principle, but it can be seen as a requirement in these cases to well pursue the principle of Nonmaleficence.

4.4 Explainability and beneficence

We think that the relationship between explainability and beneficence should be considered as a different case than those of the other three principles.

AI applications can be developed with the aim of augmenting predictive power. If one is concerned with medical diagnosis, for instance, a good AI algorithm should be the most accurate possible.[18] But what if a better algorithm is

[17] "Obligations of nonmaleficence include not only obligations not to inflict harms, but also obligations not to impose *risks* of harm." [2]

[18] Here, we assume accuracy as a good proxy of performance.

a "black box" one—should one adopt it instead of another with lower prediction power, but more "transparent" in its operation? As Lipton writes, "the short-term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving health care" [21]. Explainability as total "transparency" can thus be at odds with other demands for an artificial intelligence, such as doing better prediction with an accurate algorithm that would however be "opaque".

For instance, Nguyen [25] presents a criticism towards the aim of the ideal of an universal public transparency, which he names *epistemic intrusion*, arguing that the "drive to transparency forces experts to explain their reasoning to non-experts". But, says Nguyen, by definition the reasons of the experts are not accessible to non-experts, such that "the demand for transparency can pressure experts to act only in those ways for which they can offer public justification". For Nguyen, it would be intrusive to always ask for absolute transparency, which then becomes a form of surveillance. Transparency (as surveillance) has a kind of cost, and it should then be reserved to cases where there is a higher risk of corruption or *bias*. In Nguyen words, "Transparency is not an unreserved good". We could perhaps think that the idea that transparency should be evaluated regarding corruption or *bias* is analogous to say that explainability is more morally relevant if higher risks of harm or injustice are involved in the circumstances of the case. Both arguments imply that explainability (transparency, in the case of Nguyen) should not be an ethical principle.

In addition, a trade-off may exist between accuracy and explainability [33]. But what is the status of this trade-off? Is it a technical balancing or an ethical dilemma? Better accuracy can typically be considered under a *beneficence* directive: in the case at stake here, one looks for a better algorithm. But what about explainability? We agree that it is desirable, for reasons that will be developed below, but as we will show it is not an ethical principle for AI.[19]

Typical ethical dilemmas concerning explainability would then concern a tension between beneficence (the good that could be done with the use of the accuracy of the model predictions) and the lack of explainability. This dilemma would apparently suggest that explainability should be an ethical principle.

However, as we argued, this is not the case, since for each of the principles of autonomy, justice and nonmaleficence, we can find explainability as *a requirement for*, or *derived from* them. That is to say that in the case of a supposed ethical dilemma between the principle of beneficence and the alleged principle of explainability, the latter can be shown to be not a principle in itself, as it would be a dilemma between beneficence and another one of the four principles of bioethics (explainability being required by or derived from this last one). From an ethical point of view, explainability does not take the place of a principle in a dilemma because there are already principles that consider what is at stake regarding explainability.

But what about the relationship between explainability and beneficence when they are not on the opposite sides of a dilemma?

Situations could also exist where the lack of explainability would impair the beneficence principle—one could think of a situation where the lack of explainability is associated with the impossibility of doing a good, for instance if a procedure is not understood by the physician or by the health professional. In this case, the lack of confidence could prevent the treatment from being used, and then the accomplishment of the beneficence principle would be less probable.

More broadly, under the Beneficence Principle, Floridi et al. [10] find the ends of Promoting Well-Being, Preserving Dignity, and Sustaining the Planet, which are listed by the AI Ethics declarations analysed by them. One could ask, however, what justifies to identify Beneficence with these values? In what follows, then, we will rather deal with the Principle of Beneficence according to Beauchamp and Childress [2]: "a statement of a general moral obligation to act for the benefit of others".

An hypothetical algorithm may not be explainable, but it may not yield significant impact on nonmaleficence, justice and respect of autonomy, while it may do good (beneficence). Or let us say rather that in some cases the possibility of beneficence could be evidently more important than possible harms (think for instance about a system used for traffic light optimization) or than the infringement of other principles. In these situations, would the lack of explainability be an ethical reason to deny its use? This question is a hard one, which is difficult to reply to. It seems to show that even if explainability is desirable, it is not always demanded, and thus is not a principle, but rather a requirement that serves for principles.[20]

But problems come even before this, as it is not that simple to identify with certainty the cases in which there is an absence of harm, injustice, or autonomy violation. Robbins makes the important point that we should determine *which decisions* require explanations in AI, since some of them are

[19] One could say that in general it is better to understand a technology than not—of course, this is not absolute, as better understanding it typically causes lack of some other aspect. These are the trade-off situations we will analyse.

[20] Ethical principles are not absolute, in the sense that in the case of a conflict between principles one can be postponed for the sake of another. There is always a demand for a principle—what is sometimes designed as *prima facie* principles: the principle should be considered in itself, except in the case of a more urgent demand.

harmful, even without explanation. "A terrible chess move may result in the loss of the chess game, but life, limb, reputation, and property are not at stake. An AI making decisions in other contexts, such as medical diagnosis and judicial sentencing, could cause real harm" [30]. We agree with this: the lack of explainability, in the case that there is no harm, injustice, or autonomy infraction, seems to be no problem. However, we think that we *cannot* evaluate the risk of violating these principles without considering explainability, as we tried to show.[21]

One could think for instance in the prioritisation by an algorithm of images of high-risk cases such that radiologists can analyse them [16]: this could have beneficial effects for the ones being prioritised, and there would supposedly be little harm to the next ones, as they are not left out of the treatment, but just wait more due to the urgency of the supposed higher risk cases. But isn't it extremely hard to know for sure that there is no bias regarding some group in this prioritisation? It is not sufficiently clear that the delay in the offer of a health care service for some people does not eventually constitute a harm. But sometimes, it is inevitable that this occurs, and for that reason the setting of criteria to avoid injustice in these allocations is important. If this is not so, there would be some risk of violating here the principles of Nonmaleficence and Justice, which, as we claimed, suppose explainability.

Finally, we should call attention to another category besides that of "requirement", which was used by us to show the relationship between explainability and the four principles of bioethics. Ursin et al. [38] highlight the distinction between an *instrumental* valuation of explainability and an *intrinsic* one. To look *instrumentally* for explainability goes together with what we presented here from an ethical point of view: one could try to respect explainability *in order to* do no harm, or *in order to* be just. On the other hand, we agree with them that, *in some situations*, explainability could be valued *intrinsically*, or as an end in itself.[22] According to the authors, this would be the case when explainability appears as an *epistemic* goal—what we acknowledge as a possibility, but that doesn't make explainability an ethical principle. Moreover, the authors indicate also the possibility of an *ethical* intrinsical valuation of "explicability", writing that explicability can appear, for example, "as an element facilitating informed consent, explicability also has an intrinsic value. Patients may value intrinsically that procedures were followed correctly independently of the outcome". However, also in this case, even if explainability has an ethical intrinsic value, it is not for that reason an ethical *principle*—as the authors say, in this case "explicability relates to the principles of justice and respect for autonomy".

From our point of view, a further specification can be made here. Something that is valued in itself can indeed be subordinated to higher values.[23] As we said above, trust can be developed by the fact that an AI system is explainable—but this can well be understood under the principle of beneficence. On the one hand, if trust would improve the outcome, there would be an instrumental valuation. On the other hand, even if trust is valued for itself (and not implying a better outcome), it could be subordinated to a principle as beneficence (let us think, for instance, in a doctor, who has the duty to inform his patient, but at first place to heal the patient).

Explainability is thus not necessary as a principle regarding the Principle of Beneficence.

As we saw in this Sect. 3, it becomes clear that explainability is not an ethical principle by itself, but an epistemic requirement that becomes ethically relevant only when demanded by ethical principles.

5 Conclusions

Explainability *is important* in AI ethics, but we assert that it is not a fifth ethical "principle". In an AI ethics guided by principles, Explainability does not have the same ethical status as Respect for Autonomy, Beneficence, Nonmaleficence and Justice.

Ethical decision making should not restrict its range of morally acceptable algorithms because of explainability, unless this claim is based on one of the four principles of bioethics; hence the demand for explainability is an indirect one.

The distinctions we made in this paper are not just a conceptual refinement, with no practical impact. Explainability *is required*, from an ethical perspective, in several situations in AI Ethics. We *do* think that the use of AI algorithms *do present new aspects* of explainability questions with ethical implications. Nevertheless, we think that even if this is

[21] The European Commission's High-Level Expert Group on Artificial Intelligence [15] declares that the degree of explicability that is needed is dependent on the severity of consequences of it, to which they comment: "for example, little ethical concern may flow from inaccurate shopping recommendations generated by an AI system, in contrast to AI systems that evaluate whether an individual convicted of a criminal offence should be released on parole". One could think, however, that, even in this case, to evaluate whether a risk of infringement of a principle occurs is not so simple.

[22] It is not clear whether every end in ethics is necessarily intrinsically valued—cf. Korsgaard [19], which claims for two different distinctions, between means and ends and between intrinsic values and extrinsic values. These problems should be addressed in future work regarding AI ethics.

[23] As we said, there is an open question regarding if one should talk about something that is "valued in itself", "intrinsically good" or an "end" for the present question, and we leave it for future works.

so, explainability cannot be considered as a 'fifth principle" in AI Ethics, for the reasons outlined above, namely that explainability (i) can be considered as an epistemic requirement for ethical principles; and (ii) it can be derived from other ethical principles. Moreover, we think that clarifying what are the aspects involved in the debate about the definition of explainability can improve the ethical debate related to it. This would allow one to ask for explainability when necessary, but using an adequate ethical terminology.

Acknowledgements We thank André Levi Zanon, Cláudio Amorim, Lucas Petroni, Marcelo Santis, Marcos Menon José, Paulo Pirozelli and Tiago Lubiana for reading earlier versions of this article. We also thank an anonymous reviewer for some valuable suggestions. We thank Fapesp, CAPES and CNPq agencies, as well as the Fundação José Luiz Egydio Setúbal, for their support through research grants (the second author is partially supported by CNPq grant 312180/2018-7 and FAPESP grant 2019/07665-4).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. The Belmont Report: Ethical Guidelines for the Protection of Human Subjects. Washington: DHEW Publications (OS) 78-0012 (1978)
2. Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics, 8th edn. Oxford University Press (2019)
3. Beauchamp, T. L., & DeGrazia, D.: Principles and principlism. In: G. Khushf (ed), Handbook of Bioethics, Springer, Dordrecht, pp. 55–74 (2004)
4. Buchanan, B.G., Smith, R.G.: Fundamentals of expert systems. Ann. Rev. Comput. Sci. **3**, 23–58 (1988)
5. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci. Eng. Ethics **26**(4), 2051–2068 (2020)
6. Coeckelbergh, M.: AI Ethics. The MIT Press, Cambridge (2020)
7. DARPA: Explainable Artificial Intelligence (XAI). Broad Agency Announcement DARPA-BAA-16-53 (2016)
8. Fjeld, J., et al.: Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society (2020)
9. Floridi, L.: Infraethics – on the conditions of possibility of morality. Philos. Technol. **30**(4), 391–394 (2017). https://doi.org/10.1007/s13347-017-0291-1
10. Floridi, L., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**(4), 689–707 (2018)
11. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. Harv. Data Sci. Rev. **1**(1), 1–15 (2019)
12. Gorry, G.A.: Computer-assisted clinical decision making. Methods Inf. Med. **12**, 45–51 (1973)
13. Hermann, E., Hermann, G.: Artificial intelligence in research and development for sustainability: the centrality of explicability and research data management. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00114-8
14. Herzog, C. On the risk of confusing interpretability with explicability. *AI and Ethics*: 1–7 (2021).
15. High-Level Expert Group on Artificial Intelligence: "Ethics Guidelines for Trustworthy AI". Document made public on 8 April 2019 (2019). https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai
16. Jha, S.: Value of triage by artificial intelligence. Acad. Radiol. **27**(1), 153–155 (2020)
17. Kemp, P., Dahl Rendtorff, J.: The Barcelona declaration. Towards an integrated approach to basic ethical principles. Synth. Philos. **23**(2), 239–251 (2008)
18. House of the Lords: AI in the UK: ready, willing and able? HL Paper 100 (2018).
19. Korsgaard, C.M.: Two distinctions in goodness. Philos. Rev. **92**(2), 169–195 (1983)
20. Lacave, C., Diez, F.J.: A review of explanation methods for Bayesian networks. Knowl. Eng. Rev. **17**(2), 107–127 (2002)
21. Lipton, Z.C.: The mythos of model interpretability. ACM Queue **16**(3), 1–27 (2018)
22. Miller, T.: Explanation in artificial intelligence. Artif. Intell. **267**, 1–38 (2019)
23. Mirbabaie, M., et al.: Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research. AI Soc. (2021)
24. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
25. Nguyen, C. T. Transparency is Surveillance. *Philos Phenomenol Res*. (2021). https://doi.org/10.1111/phpr.12823
26. Rana, A., Bridge, D.: Explanations that are Intrinsic to Recommendations. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, 187–195 (2018)
27 Rawls, J.: A Theory of Justice, 2 revised Belknap Press, Cambridge (1999)
28. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, 1135–1144 (2016)
29. Richardson, H.S.: Specifying, balancing, and interpreting bioethical principles. J. Med. Philos. Forum Bioethics Philos. Med. **25**(3), 285–307 (2000)
30. Robbins, S.: A misdirected principle with a catch: explicability for AI. Mind. Mach. **29**(4), 495–514 (2019)
31. Roemer, J.E.: Theories of Distributive Justice. Harvard University Press, Cambridge (1996)
32. Rudy-Hiller, F.: The epistemic condition for moral responsibility, The Stanford Encyclopedia of Philosophy (Fall 2018 Edition). In: Zalta, E.N. (ed.). https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/. (2018). Accessed 22 Mar 2022
33. Sarkar, S., et al.: Accuracy and interpretability trade-offs in machine learning applied to safer gambling. In: CEUR Workshop Proceedings, vol. **1773**, 1–9 (2016)
34. Selvaraju, R.R., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, 618–626 (2017)
35. Sen, A.: Commodities and Capabilities. North-Holland, Amsterdam (1985)
36 Tsamados, A., et al.: The ethics of algorithms: key problems and solutions. AI Soc. **37**, 215–230 (2022)
37. Turilli, M., Floridi, L.: The ethics of information transparency. Ethics Inf. Technol. **11**(2), 105–112 (2009). https://doi.org/10.1007/s10676-009-9187-9
38. Ursin, F., Timmermann, C., Steger, F.: Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? Bioethics **36**(2), 143–153 (2022)

39. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 195–200 (2019)