The Man Behind the Curtain: Appropriating Fairness in AI

Minds and Machines (2024) 34:7
https://doi.org/10.1007/s11023-024-09669-x

The Man Behind the Curtain: Appropriating Fairness in AI

Marcin Korecki[1] · Guillaume Köstner[2] · Emanuele Martinelli[3] · Cesare Carissimo[1]

Received: 31 May 2023 / Accepted: 21 February 2024

Abstract
Our goal in this paper is to establish a set of criteria for understanding the meaning and sources of attributing (un)fairness to AI algorithms. To do so, we first establish that (un)fairness, like other normative notions, can be understood in a proper primary sense and in secondary senses derived by analogy. We argue that AI algorithms cannot be said to be (un)fair in the proper sense due to a set of criteria related to normativity and agency. However, we demonstrate how and why AI algorithms can be qualified as (un)fair by analogy and explore the sources of this (un)fairness and the associated problems of responsibility assignment. We conclude that more user-driven AI approaches could alleviate some of these difficulties.

Keywords AI fariness · AI normativity · Responsibility in AI

1 Introduction

As artificial intelligence (AI) research achieves more and more of its aims, some previously thought unreachable, the increased integration and enmeshment of AI and the social fabric becomes a new reality. While AI promises new heights of humanity's development and perhaps a way of transcending some of the boundaries

✉ Marcin Korecki
marcin.korecki@gess.ethz.ch

Guillaume Köstner
guillaume.kostner@unil.ch

Emanuele Martinelli
emanuele.martinelli@uzh.ch

Cesare Carissimo
cesare.carissimo@gess.ethz.ch

[1] Computational Social Science, ETH Zurich, Stampfenbachstrasse 48, 8006 Zurich, Switzerland

[2] Philosophie des sciences, UNIL, Quartier Dorigny, 1015 Lausanne, Vaud, Switzerland

[3] Philosophisches Seminar, UZH, Zollikerstrasse 117, 8008 Zurich, Switzerland

Published online: 25 April 2024

of our material struggle it is also, especially in its interactions with the individual and the social, tremendously dangerous and not well understood. One of the main issues standing in the way of a wider introduction of AI into different social spheres is that of fairness. As AI solutions are proposed for driving, traffic signal-control, writing support, image recognition and much more, it is of utmost importance to make sure these AIs, that would be deployed for such sensitive applications, would follow the rules and ensure fair treatment to all its users. However, to make sure the AI is fair one must first understand what it means to refer to AI as fair. Indeed, AI is often considered from a futuristic perspective as an entity imbued with intentionality and even subjectivity and through that it is referred to as fair or unfair. This normative status, we argue, is unfounded and mistakenly applied in the context of modern day AIs.

In this work we present a strong philosophical foundation for what a given entity needs to be considered fair or unfair in the true sense. We then apply these conditions to modern AIs and conclude that they are not normative beings and so cannot be considered fair or unfair. They can nevertheless be referred to as '(un)fair' by analogy. Basing ourselves in that analogous understanding we ask 'What is the source and nature of unfairness in AI algorithms?'. Further, we point to the risks of misunderstanding AI as normative, which might allow for a misappropriation of responsibility for any biases that the AI algorithms might exhibit. Since we argue that at present AI algorithms are completely lacking in certain key properties, the accountability lies with other parties (designers, owners or users). Nevertheless, hiding behind the AI allows many kinds of manipulations, intended or unintended. 'I just did what the AI told me to do.' or 'We just followed the model.' are the type of explanations that might become increasingly common over time. In other worlds, the AI might be used as a scapegoat, while the true source of the unfairness or failure lies elsewhere. We propose a classification of source of potential 'unfairness' in AI algorithms that can be used to conceptualize and better understand where the responsibility for the eventual 'unfairness' lies. We give examples of edge cases where assigning responsibility poses significant difficulties.

Finally, we conclude that the key aspect to consider at this stage of AI development is who decides which values are being followed by the algorithms. As long as the AI's objectives are chosen by a select few any change to existing power structures is unlikely (and thus fairness is unlikely to be prioritised). A potential solution is allowing users to select the rewards of the AI directly. We also point out, based on the sources of 'unfairness' that we identify that *an unfair society is bound to create unfair AIs.* Thus the problem of AI 'unfairness' is related to the overall problem of unfairness as it has been present in all our modern societies.

2 In What Sense Can AI Algorithms Be Called (Un)Fair?

When we attribute responsibility for actions we deem fair or unfair, both legally and morally speaking, it seems to be the case that only individual natural persons can be endowed with responsibility. However, this may raise legitimate objections. In everyday language as well as in ethics, legal theory and political sciences, we say,

in a perfectly acceptable way, that things which are not people, public institutions or economic systems for example, can be fair or unfair. For instance, the conditions of the American job market in the steel industry may be said to be 'unfair' not in the sense that the social system is itself morally or legally responsible for, say, the systematic exploitation of the workers or widespread exclusion of women from the labor force. There is a sense in which the industry is unfair as it brings about unfair consequences that unjustly breach people's rights or dignity.

Regardless of a precise definition of which rights or principles tell what is fair from what is unfair in the abstract, a first point to establish is that, while all responsible beings are capable of producing fair or unfair actions, not all things that can be described as fair or unfair need to be responsible. In this section, we examine the various ways in which (un)fairness can be predicated of a thing, as well as the ways in which these different uses are related to the notion of responsibility. This will allow us to then clearly define the type of (un)fairness that can be attributed to AI algorithms. Let us define two different senses in which we may attribute normativity and, by inclusion, (un)fairness to something:

- Proper normativity or proper (un)fairness: x is (un)fair as it produces (un)fair actions for which x is responsible.
- Analogical normativity or analogical (un)fairness: x is (un)fair as it participates in one way or another in (un)fair events for which x cannot be responsible (other beings, either dependent or independent of x for their existence, are responsible).

2.1 Two Brands of Normativity

We readily admit that things such as judicial, economic or political systems, institutions, and companies can be qualified as fair or unfair. According to us, we simply need to be aware that we are then predicating fairness analogically, as with the predicate 'healthy' applied to animals, drug, urine and food in the example given by Aristotle (1998).[1] Strictly speaking, only the animal can be healthy, and the drug is then called 'healthy' insofar as it produces health in the animal, urine is called 'healthy' insofar as it is the sign of a healthy animal, and food is called 'healthy' insofar as it contributes to the health of the animal. Similarly, a judicial system can perfectly well be called unfair, but only insofar as it reflects unfair actions and beliefs of certain people. The analogical extension of the concept's meaning can even reach the stage of mere metaphor, for example when we talk about the toss of a fair coin. In this case, the predicate 'fair' is also not unrelated to its first usage in reference to people: we speak of a fair coin because the result of the toss is analogous to the result of an

[1] Here, we will use the notion of analogy in the sense given to this concept by medieval Scholastic philosophy. Aristotle only named 'analogy' the analogy of proportions, that is, an identity of relationships between four terms (A/B = C/D). In the medieval period, analogy was also used to refer to the so-called 'analogy of attribution', which is the derivation of the meaning of a notion by reference to a first term. Aristotle's example of health corresponds to such an analogy of attribution in which the health of the animal is the first term the meaning of which is extended to designate other things. An analogy of attribution can involve an analogy of proportion, but not necessarily.

impartial person's random choice between two items. Although admissible, the use of the predicate 'fair' is so far removed from its original meaning that it is indeed a metaphor,[2] and there is no sense in making a normative judgment about the coin, unlike what is possible in the case of an unfair economic system or the responsibility of a juridical person.

On the other hand, many voices in the literature about the philosophy of AI adopt a different approach. Whereas social institutions are conventional in nature and do not enjoy the same sense of individuality as individual human beings, they may be more easily excused from concepts like responsibility or accountability. When assessing the responsibility of corporations for crimes, for instance, moral responsibility is more often than not applied to the managers, CEO, workers, or anyway a set of individual members of the corporation (List & Pettit, 2011). Perhaps because AI devices retain a stronger sense of individuality, researchers are more willing to approach the (un)fairness of AI systems by considering how much agency, autonomy, and responsibility we should attribute to 'AI agents'.

One important example of this popular approach is the discussion surrounding the 'responsibility gap' yielded by AI systems, sometimes called the 'problem of many hands' (De Poehl, 2015; De Sio, 2021). Put simply, the issue is the following. Suppose an AI application tasked to automatically assess the creditworthiness of the prospective customer of a financial institution is found to systematically penalize people of color for no real reason. Perhaps, the bias has been hidden in the data set on which the AI application was trained to begin with, and does not represent in any way a real pattern of lower creditworthiness of the people of color requesting access right now. Suppose the behavior of the system is completely unexpected and does not reasonably come from negligence on how it is used. Who is to blame for the *unfainess* of these decisions? The developers? The company trusting the system? The users? The AI system itself?

There have been different reactions to this problem. To treat them with the necessary attention would be way beyond the scope of this paper. However, a couple of instances will facilitate the point we want to make. One possible approach is to explore the extension of agency and responsibility to AI agents themselves, in order to hold them accountable for their own autonomous decisions (Laukyte, 2017; List, 2021). Another solution would be to pretend that AI systems could be treated as agents, and treat them *as if* they were the active cause of unfairness under certain circumstances (Tollon, 2023)—for instance, in a similar way to what we do when we treat big corporations as 'legal persons', e.g. in the case of limited-liability companies. Again, De Poehl (2015) suggest to deconstruct the problem by capturing the blame for unfair decisions coming from AI systems within the overall 'socio-technical system' that made the construction and the deployment of the AI in such a way possible in the first place.

[2] Even if we consider the expression 'fair coin' not only as a colloquial term but also take into account its mathematical definition (a sequence of independent Bernoulli trials with a probability of 1/2 of success on each trial), the application of the term 'fair coin' to this mathematical definition is, once again, a matter of metaphorical usage.

Table 1 A taxonomy of matter as functional or non-functional, automatic or allomatic, designed or not designed. AI algorithms have a function, are automatic and designed

No Function	Function		
		Automatic	Allomatic
mere heap of matter (e.g. pile of sand, rock)	Design	machines (e.g. AI)	tools (e.g. hammer)
	No Design	organs (e.g. heart)	animal architecture (e.g. ant hill)

In our opinion, all these solutions have one fundamental limit. They more or less explicitly treat AI systems, due to their relative autonomy in action, as potential *loci* of agency, responsibility, and (un)fairness. They thus proceed to either include AI agents in their understanding of the world or to see whether and how we can fictitiously ascribe agency to AI or otherwise make sense of why AI seems to be an agent. These all presuppose that, for some reason, we should predicate proper normativity to AI applications or particularly autonomous artifacts, rather than treating them with the same analogical normativity we ascribe to social systems (Table 1).

In the following, we will explore this intuition. We will now introduce some useful conceptual tools to decide whether this popular tendency to attribute proper normativity even in potentiality to AI systems makes sense, or if we should move to an analogical understanding of normativity. After spelling out the conditions for proper normativity in Sects. 3 and 4, we will argue in Sect. 5 that AI systems should be deemed fair or unfair only in analogical sense. The rest of the paper is devoted to pave the way for a treatment of the unfairness of AI along these lines.

2.2 From Agency Comes Responsibility

It becomes immediately apparent that normativity in the proper sense requires the ability to be held responsible for one's own actions. We will develop the conditions for proper normativity in the following section. However, we are going to anticipate that a core precondition for responsibility is *agency*, i.e. the capacity to act. Let us then deliver relevant definitions that will serve for the rest of the paper.

Following Arp and Smith (2008), we say that x's function is a disposition of x that defines its nature, and whose realization is forced by sufficiently favorable conditions in the environment. In other words, an object necessarily tends to realize its function, while it just contingently tends to realize other dispositions. For instance, consider your PC charger. The PC charger is defined by its disposition to charge my computer, so that it may be deemed a good or bad charger based on how well it charges your PC. Thus, charging PC batteries is the PC charger's function. Now, the PC charger also has the disposition to keep the cat warm as he lays on it. That disposition does not define your PC charger's nature and is contingent on other features of the environment other than itself. Hence, it is merely a disposition and not a function. A similar notion of function has been linked by Burge (2009) to the emergence

of biological normativity.[3] This is built into the fact that functions intrinsically manifest how they *ought to* be performed.

Drawing from Martinelli (2023), we can build a distinction between three levels of agency based on the agent's relation with its functions.

- Active agency is the ability to realize one's function through interaction with the environment and with no intervention by external systems. A thermostat, for instance, is capable of increasing the room's temperature when the temperature goes too low, without any resort to reason or perception.
- Reactive agency is the ability to realize one's function by selecting one's courses of action autonomously. A chess-playing machine, for instance, is capable of pursuing the goal to win the current match by autonomously selecting its next move according to an available array of legal moves, based on the current state of the board.
- Proactive agency is the agent's ability to manipulate its own functionality and redirect its functions to adapt to the environment. A human person, for instance, can resort to go out and have a walk without any real prompt from the external environment: she has simply elected to have a stroll and set a new goal to pursue. After the new goal has been set, her action may be deemed good or bad in relation to the quality of her walk.

On a similar note, Ezenkwu and Starkey (2019) distinguish between a high-level form of autonomy and a low-level form of autonomy. High-level autonomy is the ability to set one's own goals in one's interaction with the environment, as opposed to merely realizing pre-determined goals imposed externally (Popa, 2021). This is the core difference between tools and fully autonomous agents, if you will (Baird & Maruping, 2021). These notions, respectively, compose the marks of proactive agency and reactive agency. This is because functions operationally represent goals for an agent, as its actions are essentially directed to realize them. To have pre-determined goals amounts to serve built-in functions, like reactive agents do; to be free to self-identify the goals to pursue means retaining the capacity to redirect one's functions within the environment.

Now, proactive agency, or high-level autonomy if you will, is the condition to interact meaningfully with the environment. By 'meaningfully' we follow Di Paolo (2005) and mean that an agent makes sense of the environment if the kind of response the agent gives to a certain input aims at maintaining the internal states of the agent in spite of environmental perturbations. Agents in real-world environments may interact with their environment meaningfully only if they can deal with any kind of perturbation.

As we have already mentioned, normativity itself is already an analogical notion which can be distinguished by different regimes of application. In the next section, we will show that the strict sense of normativity corresponds to the one we use

[3] Darwin himself was aware of the analogical dimension of the terms he used in formulating his theory of evolution. On this point, see da Silva Oliveira (2022).

when we evaluate the actions of free and rational individuals, as only these beings genuinely have a relationship with what '*ought*-to-be-the-case' as opposed to what '*is*-simply-the-case'. However, as suggested by the notion of function above, there is a possible ground for normativity outside the realm of agency alone.

If the heart's function is to pump blood, we may say that the heart 'ought' to pump blood in order to be considered as a heart at all. Even more, a heart may be called good or bad, or, rather, may be said to function well or badly, based on how effectively it pumps blood, i.e. how fully its function is realized (Arp & Smith, 2008). Yet, it is clear that in pumping blood, the heart follows no rules and cannot choose to transgress its function while continuing to be a heart. This kind of functionality devoid of agency is the basis of what we have called 'analogical normativity'. Artifacts such as houses and chairs 'must' be able to fulfill certain functions to be considered as houses and chairs. For example, a house can be red or white, have one or fourteen rooms, one or five floors, but it must ('ought-to') in all cases have insulated walls, a watertight roof, and solid foundations to be a house. The difference between a good house and a bad house, e.g. predicated on the quality of insulation and waterproofing (compare a hut in the savannah and a residential villa), is here at best metaphorical and does not have a proper moral character.

The attribution of normative predicates to artifacts is possible because these objects reflect or participate in the actions of normative beings in the primary sense: a good house is a house that reflects the skills of the architect and builders and participates in the unfolding of a good life for those who inhabit it. This comes from the fact that the house fulfills the function that defines it. The analogical extension of normativity from its primary sense to its technical sense concerns this precise point: an artifactual *function* is analogous to a *rule* that a rational being follows in performing an intentional *action*: a good heart is a heart that corresponds to the function that defines it; as in the case of a house, it may be larger or smaller, noisier or quieter, but it 'ought' to pump blood to be a heart. The sense of this second analogical extension is as follows: organs and organisms instantiate functions *as if* they were the product of the activity of a craftsman manufacturing an object to serve a certain purpose.

2.3 Artificial Intelligence and Artificial Normativity

As technical objects, AI algorithms can thus be integrated into the logical space of technical normativity: they can be good or bad, in an analogical sense of these predicates, depending on whether they fulfill their functions or not. The question is whether they can be qualified as fair or unfair according to a similar analogical extension that goes from the fairness of persons to the fairness of institutions. In the case of the biological entities and artifacts that we considered earlier, they do not seem to be able to be qualified as (un)fair. A heart affected by a pathology could be called a bad heart, but it does not seem possible to say that it would be unfair, even by analogy, to the organism to which it belongs, and similarly for a poorly insulated house that would not be called unfair to its inhabitants. This raises another question, namely what are the characteristics of things that are not single natural persons but can be (un)fair? Before addressing this question, we can already position AI

algorithms in relation to the artifacts and biological entities that we mentioned. In this sense, a first thing to note is that biological organs such as hearts and lungs are defined by functions that are not the result of a design, while artifacts such as houses and hammers have functions that are the product of a design. An important difference between these two types of things that instantiate functions, in addition to their relationship to a possible design, is that organs have automatic functioning (they do not require intervention from an external agent to function, they possess active agency in the sense discussed above), whereas an artifact such as a hammer has entirely 'allomatic' functioning (it never works without intervention, without direct and constant manipulation by an external agent, i.e. it is entirely devoid of agency). The hypothesis we would like to propose is that we have here the reasons why biological organs and some artifacts cannot be said to be (un)fair although they can be said to be good or bad due to the functions they instantiate. Namely, organs cannot be said to be (un)fair because they do not participate in any design that would be at the origin of their functions, and artifacts like hammers cannot be said to be (un)fair because they are entirely devoid of agency, so it is always the agent who manipulates them who can immediately be held responsible for a (un)fair use of these tools. In both cases, (un)fairness cannot be predicated by analogy because there is no *participation*: the organ does not participate in any intentional design, and the simple tool does not participate in any action, in the relevant sense of participation, as it *immediately* is part of the action.

In regards to AI algorithms, we have shown in the previous section that they possess a form of reactive low-level agency. This characteristic sets them apart from simple tools like hammers. However, we have also shown that this form of reactive low-level agency is paired with a form of bounded autonomy that comes from the fact that these AI systems are the product of a design. Therefore, by virtue of this second characteristic, they also differ from biological organs. Do these two characteristics allow us to compare AI algorithms to things that can be said to be (un)fair by analogy? Let's take the example of an economic system: even in cases of largely planned economies, it can be doubted that such systems are entirely the product of design and it is possible to argue that they always retain a dimension of spontaneous order. It is also not easy to isolate the functions that they instantiate, so it seems difficult to judge their level of agency, or even to say that they possess one regardless of the sense given to this term. However, an economic system can be (un)fair in that it participates in certain (un)fair actions, makes them possible, or manifests them. Unlike tools like hammers, this same system also enjoys a degree of independence from the actions and existence of particular individuals: no individual has to act constantly and intentionally so that the economy works, and the same system can perpetuate itself over many generations, indicating also that it is independent of the existence of particular individuals. Thus, a relationship, or a plurality of mediated relationships to the actions and existence of concrete individuals seems to be a condition for such a system to be qualified as (un)fair. Although AI algorithms may initially appear closer to tools than to economic systems or public institutions, they resemble them precisely in that they also have a degree of independence from the actions and existence of specific individuals and yet always have a mediated relationship to these individual actions and existences.

For this reason, it seems to us appropriate to analogously qualify certain AI algorithms as (un)fair. However, it must be immediately emphasized that this (un)fairness does not imply any form of responsibility on the part of the AI algorithms. Just as it is possible to say that a house is a bad house without attributing responsibility for its flaws to it (while saying that an action is bad is inseparable from attributing responsibility for it to an agent), it is possible to say that an AI algorithm is unfair without attributing any responsibility to it.

3 What Does It Take to Be (Un)Fair?

We have seen, by introducing the Aristotelian notion of analogy, that analogous terms like 'health' have a primary sense from which analogical uses derive. In the previous section, we sought to demonstrate how the analogical derivation of normative predicates such as fairness can occur in the case of AI algorithms. A proper understanding of this extension, as is the case with 'health' and the predicates derived from it, requires an understanding of the primary sense of 'normativity' and 'fairness'. Indeed, when we ask about the source of fairness or unfairness in AI algorithms, the first step is to clarify the sources of fairness or unfairness in general, and more broadly the sources of normativity (that is, the set of concepts and phenomena implicitly or explicitly involving a form of 'ought', for example, an unfair action is an action that *ceteris paribus* ought not to be committed). In order to orient ourselves in the approaches followed by the majority of the aforementioned literature, we will in this section spell out the conditions that an entity must meet in order to possess proper normativity. This is the sense of normativity that we will thus deploy in this section and the next.

This first step should allow us to answer the question of the location of fairness or unfairness that we already perform as human beings, independently of the type of tools we use. By location, we mean here the delimitation of a logical space within which we admit certain phenomena or events that we can qualify as fair or unfair and exclude other phenomena or events to which these notions simply do not apply. In this section, we outline such a delimitation by defining a set of criteria that a phenomenon or event must meet to be integrated into the logical space within which the application of normative notions such as fairness makes sense. The argumentative and explanatory trajectory of this section is, in a sense, top-down: we explain what the normative point of view (that is, the point of view that is within the normative logical space) consists of and adopt it to show *that* we exclude AI algorithms from things that can be qualified as fair or unfair in the primary sense of the term. The next section complements the explanation of this exclusion by developing a bottom-up argumentative and explanatory trajectory: it shows *why* AI algorithms fail to meet the criteria for inclusion in the logical space of normativity, namely that they do not possess the type of organization necessary to ground the normative capacities they should possess to be normative agents. The convergence of these two lines of argumentation allows us to establish that the source of fairness or unfairness in AI algorithms cannot reside in the algorithms themselves, and that there are clear reasons in support of this thesis. AI algorithms can only be called fair or unfair by

analogy and insofar as they entertain relations of participation with the source of proper (un)fairness. In turn, we can also define more clearly the source in question, which must possess a certain type of physical organization and meet the criteria that we will now explain.

3.1 Conditions for Proper Normativity

The concept of fairness, like all other normative notions, is intimately linked to that of consciousness.[4] There is no sense in saying that a tool, let's say a hammer, considered in itself, treats a nail correctly or incorrectly. There are correct and incorrect uses of hammers, but these uses are then attributed to an agent endowed with consciousness and personhood, which ensure that they can be held responsible for their actions. Similarly, one cannot say that a nail has been the victim of injustice because it does not possess any of the relevant attributes for treating a being as a moral patient. Once again, these attributes are mental: setting aside for a moment the question of personhood and agency, at least sensitivity must be attributed to a being to make it a moral patient. In short, and to return to the side of the agent, there is no normativity without mentality: an action, an utterance, or a thought can be evaluated as correct or incorrect, just or unjust, good or bad, fair or unfair only if it can be attributed to a being possessing agency, and most importantly *proactive* agency as per the previous definition. This being must have reflective consciousness, that is, be capable of forming representations of her own situation in the world. This reflexivity, that is, the ability to have and entertain a sense of self, is also the condition for this being to be held, by herself and by others, responsible for her actions, utterances, and beliefs. The crucial point in what we have just established is that normative agency implies some form of responsibility, which in turn implies that there is a self that can be recognized as the author of its actions, and finally, that this self depends on consciousness. Indeed, if a being were not conscious, in other words, if it were not open to the world, it could not represent its own situation in the world, i.e. it would not be self-conscious. It is in this precise sense that we said earlier that all normative notions are linked to consciousness: it is a condition of possibility for the formation of the self to which responsibility can be attributed.

Two other characteristics of beings that entertain a relation to normativity must still be emphasized: rationality and freedom. A person can only be held responsible for her actions, words, and beliefs on the condition that she can justify them in an intersubjective context and according to certain normative criteria of rationality. For example, an agent would be able to justify her use of a hammer to break a window by saying that it was rational to do so at the moment she did

[4] In the sense in which we will use this notion, a being is said to be conscious if it possesses states characterized by a qualitative experience (something it is like to be in that state, what it is like to see the redness of an object for example) and intentionality (the fact of being directed towards something, of being about a part of the world). The questions related to the relationship between these two characteristics (such as those asking whether the qualitative dimension of the experience depends on the intentional ability to represent other things) are outside the scope of our reflection.

it because she had to face danger and escape as quickly as possible. Conversely, we would not treat someone who systematically failed to provide justifications as a rational and responsible being. In the context of our reflection, rationality is thus grounded in the ability to give reasons when reasons are asked for. Treating a being as a rational and responsible being finally amounts to interpreting what they do and say as not being constrained by purely mechanistic causality. For example, two people could suffer similar bodily harm due to an object falling on their head, but in one case, the object fell due to the wind, and in the second case, due to the deliberate and malicious action of a person holding the object. There is no sense in attributing responsibility to the wind, but obviously, the person who dropped the object is responsible. We not only believe that this person is responsible for the harm suffered by the victim, but also that they acted wrongly because they should not have done that. Finally, if we say that they should not have done that, it implies that they could have done otherwise, meaning that they had genuine alternatives available and they could and should have chosen them. Imputing responsibility therefore implies treating the agent as possessing a significant degree of freedom of action. McDowell (1998) perfectly expresses the ideas presented so far when he imagines what a wolf should possess to be considered a rational being:

> A rational wolf would be able to let his mind roam over possibilities of behaviour other than what comes naturally to wolves. ... [This] reflects a deep connection between reason and freedom: we cannot make sense of a creature's acquiring reason unless it has genuinely alternative possibilities of action, over which its thought can play. ... An ability to conceptualize the world must include the ability to conceptualize the thinker's own place in the world; and to find the latter ability intelligible, we need to make room not only for conceptual states that aim to represent how the world anyway is, but also for conceptual states that issue in interventions directed towards making the world conform to their content. A possessor of *logos* cannot be just a knower, but must be an agent too; and we cannot make sense of *logos* as manifesting itself in agency without seeing it as selecting between options ... This is to represent freedom of action as inextricably connected with a freedom that is essential to conceptual thought.

To sum up: proper normativity appears when beings are endowed with:

1. proactive agency
2. reflexive consciousness
3. freedom
4. rationality
5. responsibility

A corollary is that these beings must be embodied and have a bodily constitution that allows them to be both open to the world and capable of acting in it. For example, a rational being must have a bodily constitution such that, *ceteris*

paribus, when they perceive another person in danger, they draw the correct inference that this person needs help, and act accordingly to help them. A person who, *ceteris paribus*, perceives the situation but remains immobile or who acts as if to help another person when they perceive no threat would not be considered a rational agent. We therefore have no reason to believe that beings entirely devoid of perceptual openness to the world and motor capacities can possess any form of responsibility. In the next section, we will explore in more detail the specific bodily constitution that an entity must have in order to access normative agency.

Conditions (1) and (2) are the chief ones. In fact, they are the requirements for freedom, rationality, and responsibility to take place. As we take it, for something to act freely, responsibly, and rationally, it must first be able to act and be aware of its actions. (Proactive) agency and (reflexive) consciousness, in their turn, are the two faces of the same coin, as only a sense of self can ground fully-fledged agency in the world, and, vice versa, consciousness is expressed in the performance of the agent's interaction with its environment (Popper, 1994). However, consciousness and agency are independent insofar as beings can have a mind without possessing conditions (3–5). It should be noted that these latter three characteristics, which are in addition to reflexive consciousness and proactive agency, seem to go together in such a way that one cannot possess one without possessing the other two. As suggested earlier, the beings endowed with all five characteristics must also be social beings, as their rationality only makes sense within social and linguistic practices of justification, of giving and asking for reasons. This conception of normativity reflects ideas developed within a philosophical tradition that ranges from classical American pragmatists and the later Wittgenstein to philosophers of the Pittsburgh School (such as Robert Brandom and John McDowell), passing through Wilfrid Sellars and Donald Davidson. The concepts and theses developed within this tradition are particularly relevant here because they make explicit the criteria we use when we employ normative concepts and when we treat a being as having a rational mind.

As for the question of whether a machine can be conscious, the many positions defended in contemporary debates can be located on a spectrum ranging from the pure and simple impossibility for an artificial being to be conscious to the idea that some machines are already conscious or proto-conscious. According to us, a bad way to approach these questions would be to start from a metaphysical conception of the mind (whether this conception is materialist or dualist) to deduce a set of truths about the nature of physical systems that can or cannot have such a mind.[5] Therefore, we will not put forward any argument that seeks to a priori include or exclude artificial beings from the realm of beings that can in principle have mental states and, therefore, from the realm of beings that can in principle be held responsible for

[5] A critic could argue that since the first criterion we have used to attribute rationality and responsibility to a being is the possession of a mind, we should know exactly what the mind ultimately *is* But we do not believe that we need to answer this question, just as when we say that rocks, computers, plants, animals, and human beings have physical states, the proposition is perfectly intelligible independently of an answer to the question of whether these physical states are ultimately composed of discrete particles, fields of force, or other primitive entities.

what they do, say, and think. In the next section, we will establish certain restrictions regarding the bodily constitution that beings must have in order to be able to have mental states, but these restrictions are independent of the artificial nature of the being in question.

3.2 The Delimitation of Purposeful Action

Once the questions specific to the metaphysics of mind are set aside, it is possible to approach the question of artificial intelligence and its relationship to fairness from another angle, which we believe is more interesting. That is, from the point of view that begins by asking what criteria we already use to attribute consciousness and responsibility in our daily engagement with the world and the beings that populate it. Or, in other words, what are the presuppositions of the normative point of view that make it possible? Once these criteria are established, as we have done above, it is possible to verify if a being can correspond to them. For example, it makes no sense to apply such criteria in interpreting the movement of a hammer falling to the ground. The hammer does not choose to fall to the ground when it is no longer supported by another object, and all changes in the hammer's position correspond to this scheme; in short, it is not free and therefore cannot be interpreted as a being responsible for its actions. A conscious being can also be devoid of responsibility: an animal predator killing its prey is not considered to be committing murder, in the relevant sense of the term, because it does not possess the capacities that would allow it to consider alternatives to its actions and to choose between these alternatives based on criteria of rationality. Some movements of rational beings also escape the framework of interpretation in question: the patellar reflex triggered by the impact of a hammer against the kneecap is something that can *happen to* a rational person, but it is not an *action* that can be *attributed* to her. Brandom (2019) provides a useful summary of Davidson's theory of action by formulating five principles that define the specific type of events that actions constitute:

1. One and the same event can be described or specified in many ways.
2. One important way of identifying or singling out an event is in terms of its *causal consequences*.
3. Some, but not all, of the descriptions of an action may be privileged in that they are ones under which it is *intentional*.
4. What makes an event, performance, or process an *action*, something *done*, is that it is *intentional* under *some* description.
5. What distinguishes some descriptions as ones under which a performance was intentional is their role as conclusions in the process of *practical reasoning*.

Taking the example used by Davidson (2001), moving one's finger, turning on the light switch, and thereby alerting a burglar can be described in such a way that they constitute parts of a single event. But while moving the finger and triggering the switch were intentional, alerting the burglar (whose presence was unknown to the agent) was not. What makes an event an action is that there is at least one description

under which it is intentional, and what makes a description a description of an intentional event is that it describes the event as the conclusion of practical reasoning. In this example, turning on the light was something that the agent had a reason to do based on a certain goal (e.g. going to the kitchen to get a glass of milk). Although alerting the burglar is something the agent did according to a certain description of the event, it is not part of her action and she could not be held responsible for the burglar's escape. There is thus a boundary within the world between a type of events to which hammer falls, animal predation, patellar reflexes, and unplanned burglar's escapes belong, and another type of events that are attributed to agents and evaluated according to norms.

Of course, this distinction between two types of events is not exhaustive. Important subtypes exist within these two categories. Within the events that belong to the first category, i.e., events that are not normative in the strict sense of the term, an important difference exists between events that instantiate functions and those that do not. In this case, as for the principles of action that we have mentioned earlier, one and the same event can be described or specified in many ways according to its causal consequences. For example, it is possible to describe the event corresponding to the beating of a heart in terms of the sounds caused by this beating. However, this description does not allow us to identify the function of the heart, which is to pump blood, not to make noise. As in the case of action, some genuine causal consequences of the event are not relevant to defining the function because they do not belong to its characteristic outputs. A finer classification of events should therefore distinguish between actions, functions, and events that contain neither action nor function. According to such a classification, only actions can be qualified as (un)fair in the primary sense of the term because they are intentional and therefore attributable to a responsible agent. Simple events can only be qualified as (un)fair in an improper or, at best, metaphorical sense, while certain functions can be qualified as (un)fair by analogy under certain conditions.

But for now, let's stick to the distinction between purely causal events and events involving normativity in the strict sense. What does it mean, more precisely, for an event to belong to the second type? To borrow Sellars' phrase (Sellars, 1963), when we speak of an action in normative terms, we are not merely describing or explaining it, but we are 'placing it in the logical space of reasons, of justifying and being able to justify what one says [or thinks, or does]. Any event that cannot be placed in this logical space, the space of reasons, is not an event that can be evaluated in terms of norms, whether they are logical, epistemic, practical, legal, or ethical. And for an event to be situated in this space, it must ultimately be attributable to a person capable of participating in the social and linguistic practices of giving and asking for reasons. We perform these kinds of practices daily: from the courtroom where the judge asks the accused for the reasons behind her actions, to the most trivial of conversations where we ask our interlocutor why they claim their colleague at the office is incompetent. To answer the question of whether an artificial being can be held responsible for a fair or unfair action or decision, all we need to know is whether we can interact with this being and interpret it according to the rules of the game of giving and asking for reasons underlying all our normative practices. Let's call the claim that only beings who can participate in the practices of interpretation and

mutual recognition specific to the space of reasons can be held responsible for their actions the 'normative pragmatic view'.

4 Agential Limitations Behind AI's Limited Liability

Responsibility, freedom, rationality, consciousness, and ultimately proper normativity are features of agents whose interaction with the world is so sophisticated that they can deal meaningfully with real-world environments. Even rudimentary forms of mentality ground normativity because they equip agents with the kind of intentionality and subjectivity that make advanced interaction with the world possible. This is why the most basic biological organisms, like bacteria, may display primordial versions of biological 'normativity', arguably in contrast even with advanced AI systems like ChatGPT or AlphaGo.

In this section, we explore the reason why AI systems cannot qualify for proper normativity. We advance two claims. (1) Organisms, and possibly other kinds of entities with comparable organization,[6] are the right kind of entities that can develop the agential dispositions necessary to interact meaningfully with their environment. (2) AI systems lack the characteristic organization and structure of organisms, thereby fail to qualify for proactive agency and hence proper normativity. We are then going to expand the last point as we transition into the next section.

4.1 Interaction with the Environment

There are some kinds of events in any given real-world, unrestrained system that are unpredictable in nature. By 'unpredictable' we mean that the content of the perturbation cannot be possibly inferred a priori from the epistemic perspective of the agent. And, in order to respond to new stimuli for which the agent possibly lacks any kind of data to rely on, they need to be able to set their own goals to tackle the situation creatively. In other words, proactive agents may respond meaningfully even to perturbations that evade their present goal's phase space. For instance, a chess-playing machine would halt if I responded to its last move by just throwing a sandwich on the chessboard. On the contrary, a human player would be able to deal a meaningful response, e.g. asking for explanations, walking away, or proposing new rules to include the sandwich as a legal piece. The difference is that the human player is resilient to a dynamic change happening in the situation at hand (Felin et al., 2013; Landgrebe & Smith, 2022).

The conditions for normativity that we have advanced in the last section imply the capacity to meaningfully orient oneself in the environment—that is, with

[6] The A-Life project, for instance, aims at achieving full-fledged agency for AI by constructing artificial organisms. As per what follows, the following arguments might be dependent on the success of the A-Life project in the future: if artificial organisms will be possible, we contend, also full-fledged AI responsibility, intentionality, and agency may be possible. See Landgrebe and Smith (2022) for reasons against the plausible success of A-Life.

awareness of the situation, hence rationally, freely, and responsibly. In order for an agent to be normative, it is necessary to *understand* the totality of the surrounding, including unpredictable events, and to endogenously assign weight to elements of the environment. The process of 'assigning weight' to elements of the environment amounts to the capacity of understanding the meaning of what is going on in the surroundings (Boettke & Subrick, 2002). Only by grasping the significance of an event or a new object that enters the agent's environment for the agent's own aims, it is possible for the agent to interact with it and to proactively set new goals to further its interests.

The capacity to follow rules is very important for executing predetermined goals in controlled environments. However, complex systems may require the identification of new goals as unpredictable events unfold in order for any agent to function properly, in what is commonly called the process of adaptation (Popper, 1994). Adaptation requires the ability to understand meaning, on top of following instructions (Searle, 1980). Please note that understanding, or the capacity to interact meaningfully with real-world environments like proactive agents do, does not guarantee *success* in adaptation. For instance, consider a sudden explosion in a shopping mall at rush hour. There is no reason to think any bystander would be able to correctly run for shelter or save other people rather than blundering in the panic of the moment. Anyhow, proactive agents like people would try to come up with solutions to the unpredictable circumstance, by letting go of their previous goals and rearrange their understanding of the environment to identify new goals (e.g. run for that emergency exit as soon as possible). The robot waiter of the local fast food franchise or the robot vacuum of the convenience store would not even try, and would halt on the spot or act at random, simply because that event was not programmed in advance as a possibility in their system.

But what does an entity capable of interacting with complex environments look like? Organisms are the right kind of entity to qualify for proactive agency and thus normativity. We rely on the operational definition of 'organism' from (Maturana & Varela, 1980). An organism is an entity composed by functional units (organs) whose functions are concerted to bring about the preservation of the whole. As they put it, organisms are 'autopoietic machines' that work for producing parts of the whole and maintining it, rather than producing something that goes beyond the system's individuality ('allopoietic machines'). This complex interrelation of functional units is so close that it grounds the organism's perception of the self as an individual opposed to the external environment.

The fact that a web of functional units is tuned to the composition of a coherent whole allows the organism to overcome, and manipulate, its own functionality. It is this kind of agential structure that endows organisms with the agential dispositions necessary to develop high-level autonomy. This notion of 'auto-referentiality' of organisms allows their functions to be endogenously redirected toward the realization of some goal they set in the environment. We see this, in terms of self-identification of goals, in the adaptation of organisms to their environments through biological evolution. Essentially, organisms adapt to their environment in the sense that they can develop new functions by manipulating their faculties and organs. Even at the most rudimentary end of the spectrum, amoebas may elect to

wiggle to the opposite side of the primordial soup because they have sensed that a speck of food has dropped.

This action implies something very important: the capacity on the part of the amoeba to set a new goal and pursue it. This has happened in response to an event that was unpredictable a priori. And this self-determination would not have been possible in the absence of an organizational structure oriented to the systematic preservation of the whole. In other words, the array of functions composing the amoeba caused the emergence of a new higher-order function, followed by the organism as a whole, to deal with the new situation (Burge, 2009). While, of course, this sort of agency is too weak to constitute true normativity on the part of the amoeba, it is the organic structure that grounds the preconditions for it.

4.2 The Organic View against AI

Put together, these claims are a defense of the so-called 'organic view', i.e. the claim that only organisms may qualify for (moral) responsibility (Torrance, 2008). Interestingly for our case, List (2021) has recently defended a similar brand of the organic view by stating that phenomenal consciousness is a necessary condition for moral responsibility, and that (for the moment) only biological life can ground phenomenal consciousness. We find that the organic view is a good starting point for our own claim based on the following reasons.

AI systems are not organisms, and not just because they are not biological entities. They are simply built differently organization-wise, as we shall see in detail in section 5. This has to do with the fact that they are optimizing algorithms, while organisms are brought about by evolutionary processes. As Laukyte (2017) and List (2021) note, this casts a form of 'bounded autonomy' onto AI systems—what we would term 'low-level autonomy' or 'reactive agency'. The fundamental reason is that AI agents are doomed to serve predetermined goals, and that in general their attitudes (quasi-beliefs, quasi-desires, quasi-intentions) are determined by the attitudes of other human beings, e.g. the developers that programmed the machine or that designed the system, or the people that generated the data on which the machine is trained. The fact that machines are bound to the phase space inscribed in the dataset they're trained on or on their program is an indication of this.

Bounded autonomy essentially means that AI systems have a mediated relationship with the pre-intentional world, or what Searle (1983) would call the 'background of intentionality'. Our intentional states are ultimately grounded on a background of inarticulated, unconscious knowledge that substantiate our direct encounter with the world, composed for instance by tacit know-how or perceptual experiences. Human agents, and organisms in general, have immediate access to the pre-intentional world and directly develop their own attitudes and intentional states. AI systems (and collective agents) only have an access to the pre-intentional world that is mediated by other intentional agents, which dooms them to have attitudes that are explicitly or implicitly funneled by the goals and intentions of others.

5 Normative Limits of Artificial Intelligence

In the previous sections we have worked out the necessary conditions of an entity to be considered a normative being and be endowed with responsibility. In what follows we will reiterate that at the current level of development, the modern AI models and algorithms do not meet these conditions and cannot be considered fair or unfair in the true sense. Unlike in the previous sections, where we have argued our point with little assumptions as to the form the AI takes, in what follows we will focus on a particular technical implementation of AI and will base our argument on the technological details of these systems. We will take the example of Machine Learning, arguably the most modern and successful approach coming out of the field of AI. We will focus particularly on the connectionist paradigm and its most successful methodology—deep learning. As such we will not address some of the other AI paradigms such as symbolic AI, which, being much older, has been extensively discussed in literature and whose status as producing non-normative entities is rather clear. In fact, in recent years, the most successful AI applications and at the same time the ones that have been able to pass different forms of the Turing test and have been suspected of already achieving consciousness (Lemoine, 2022; Griffiths, 2022) are all based on deep neural networks. To name a few examples AlphaGo (Silver et al., 2016), LaMDA (Thoppilan et al., 2022) and ChatGPT (Liu et al., 2023) are all relying on deep neural architectures for their unprecedented success. Therefore, if we are able to argue convincingly against deep architectures being endowed with normativity we will have addressed the status of, at present, the most pertinent candidates to that role. In the following we assume the reader has a basic understanding of deep learning algorithms, we explain the features most relevant for our argument but refer the readers to (Goodfellow et al., 2016) for a full introduction into the field.

5.1 Machine Learning as Optimization

The crux of our argument is identifying any deep learning based form of machine learning with optimization. What follows from that identity is that the inherent limits of optimization apply to modern day AIs. Thus, these AIs are significantly limited, especially in terms of freedom and proactivity, which according with our previous section are some of the minimal conditions for normativity. We highlight that we do not claim it is impossible that at some point there will emerge an AI that is not solely based on optimization. However, this is the case at this point in time, at least for any successful AI designs.

Let us start with making the case for interpreting deep learning models as optimization. First, optimization itself is to be understood as a process by which the best object (in terms of a certain objective) is chosen from a set of objects. More formally we follow the definition of optimization provided in (Carissimo & Korecki, 2023) and reproduced here.

Definition 1 Optimization is a process of choosing $x \in M$ such that $\forall_{m \in M} f(x) \geq f(m)$, where M is a set of objects and f is a total order on M.

While the above definition attempts to be as general as possible, in mathematics and science optimization is often equivalent to a process of maximizing or minimizing a certain function (that is finding its minimum or maximum).

Now, it has been known since the 90s that neural networks are universal function approximators (Hornik et al., 1989). As such, under certain conditions, they are able to approximate a variety of functional relationships between a given class of inputs and outputs. This is the form in which neural networks are employed within AI solutions. Specifically, they are used to approximate some function. This function, depending on the particular flavor of machine learning, might, for example, represent the relationship between states and actions and rewards or the relationship between inputs and intended outputs. By exposing the AI to data or an environment, this function is optimized with respect to a certain chosen value using a variant of stochastic gradient descent (Bottou, 2010). The value chosen might be a certain error between what is intended and what is output or a measure of success often referred to as reward. In concrete terms the deep neural network might, for example, be used to minimize the classification error on a training set in a classification task or maximise the sum of discounted cumulative reward in a reinforcement learning task.

Based on the above we note that the essential component of modern day AIs is a deep neural network approximating some function and the optimization process that is used to maximize or minimize the said function. This process defines and conditions any behavior the AI system might exhibit. It will be behind the values pursued by the AI, the actions it might take and the subsequent inputs that it might expose itself to (assuming it is deployed in an embodied way in the world). But this defining function itself and the objective along which it is supposed to be optimized are not chosen by the AI itself.[7] Thus, the condition of proactive agency introduced in Sect. 4 is not met as the AI agent, even if it is embodied and able to act in the world, does not have the ability to choose its own goal or modify it. It can always only pursue the pre-determined objective, an immutable property that it has no access to and that in fact defines the crux of its function.

From the lack of autonomy and proactive agency it follows that such an AI agent would also lack freedom. As such, any action it might take must necessarily have been taken by its design and it has no freedom to choose a different action than the one that has been found to be optimal with respect to the pre-defined objective. Of course some AI agents, deployed in the real world, might exhibit certain levels of stochasticity leading to an illusion of choice (as in choosing a different action in conditions which appear the same). This, however, has nothing

[7] The function is chosen by the designer of the AI, even in cases such as the e.g. Generalized Adversarial Networks, self-play or surrogate losses the structure of the reward is chosen by the developer and not in any way by the AI itself.

to do with true freedom as the agent does not have a choice but rather might incorporate a certain pseudo-random component in its optimization process.

It is worth noting that optimization is not a process through which true proactivity or freedom can be achieved as the objective always needs to be specified a priori. One could imagine introducing meta-objectives based on which the agent then selects a lower level objective giving it an illusion of choice. It is an illusion, because the choice is still defined by the meta-objectives which needs to be specified externally and a priori. An attempt to let the agent specify its own objectives through optimization would lead to an infinite regress of specifying meta-objectives.

Since any deep neural network based AI agent will essentially be an optimization agent it cannot be considered as a normative being due to the lack of autonomy, proactivity and freedom. However, as outlined in Sect. 2 it is still possible to refer to the AI as unfair by analogy. It is also clear that the responsibility for such an analogously unfair AI lies with different entities than the AI agent itself. In the following section we will attempt to identify the source of that potential 'unfairness' in AI.

6 Source of 'Unfairness' in AI

While we have provided strong arguments for why modern AI algorithms cannot be considered unfair or responsible in the same sense we consider humans as such, it is clear that these algorithms, when considered as tools, do posses certain inherent properties, which can in turn lead its users to (perhaps unaware) commit unfair actions. The 'unfairness' that we will discuss here is analogous to that which can be exhibited by institutions and is generally related to a kind of systemic 'unfairness' or ignorance. For the following we will refer to the '(un)fairness' in AI indicating its analogous sense and distinguishing it from (un)fairness in the true sense.

We have already established the optimization view of the deep learning based AIs. We have also pointed out that optimization itself is limited as a paradigm for achieving subjectivity and normativity. It is further limited as a paradigm for dealing with open, complex systems such as human societies, which are precisely the kind of systems where the questions of fairness is posed (Carissimo & Korecki, 2023). These limits of optimization correspond well to the sources of 'unfairness' in AI. Following (Carissimo & Korecki, 2023) the limits are:

1. Object limit: pertains to the relation between optimization and the phenomenon, which is facilitated by the model M.
2. Objective limit: refers to the relation between optimization and the user, encapsulated by the choice of the optimality criterion f.
3. Process limit: when the process of optimizing itself affects either the model M or the optimality criterion f. That is, by selecting objects m from M or by applying f to m the user affects the phenomenon, which in turn may affect the model and/or notion of optimality.

In terms of the sources of 'unfairness' they are intricately linked with biases and we identify two key elements of AI algorithms that can lead to any potential biases, namely: the objective (also referred to as loss or reward) function and the data the algorithm is exposed to (the inputs). These elements allow us to identify the sources of 'unfairness' in a way that corresponds to the limits of optimization. We name them accordingly:

1. Object source: pertains to the 'unfairness' induced by the data that the AI is exposed to or by the model that is chosen.
2. Objective source: refers to the source of 'unfairness' stemming from the objective function defined by the designer of the AI.
3. Process source: when the AI itself affects the data it is exposed to. Or through its actions it modifies or subverts the intended goal of the objective.

In this section we will introduce these three sources, discuss them and give examples of cases, where establishing responsibility is easy as well as examples of edge cases, where establishing responsibility is especially challenging. Lastly, drawing on the arguments made we make the case for the following statement: *An unfair society is bound to always make 'unfair' AI*. or *AI is (un)fair when society is (un)fair.*

6.1 Edge Cases

Each section will present a few edge cases. Their general structure will be to assume the AI is initially 'fair'. We will define this as follows. We assume fairness can be measured, and is operationalized in a set of fairness-functions J (some examples of such functions as applied to a particular learning setting can be found in Rychener et al. (2022)). These fairness-functions $j \in J$ can be applied to the *data*, D, and the *objective function* of an AI algorithm. An AI is initially 'fair' if and only if it is considered fair by every fairness-function $j \in J$; its creators considered pre-stated criteria of fairness, and ensured that their AI satisfied those criteria. From this assumption, we discuss cases when the AI could still become unfair, thus complicating the attribution of responsibility.

6.2 Object Source

We identify the first source of 'unfairness' in machine learning as the data itself or the model choice. Indeed, the choice of the model can lead to significant issues, if e.g. the chosen model is not expressive enough. A good example here would be using a neural network that is not deep enough and so cannot model the intended phenomenon with sufficient accuracy. Moreover, the data used to train a given model has a strong influence on subsequent classifications or actions selected by the model. It has been often stated that the machine learning algorithms are only as good as the data that was used to train them. For one, the data establishes the scope of the algorithm, the space of classes that it can recognize and the inputs that it can be expected to process correctly. If an algorithm trained

to distinguish cats from dogs is presented with an image of a human it certainly cannot be expected to function well. Applying a given algorithm to inputs which are out of its scope can lead to 'unfairness' stemming from misapplication (this is analogous to using a tool in a wrong way, such as using a hammer to kill a person). The issue nowadays is that due to the incredible amount of data being used in training the scope of the large learning models is not clear even to its designers. Furthermore, even when applied within its scope, the algorithm will still be heavily dependent on the data. If the data itself was biased or unbalanced in any way, this bias will be present in the algorithm itself.

Referring to biases in the data has become very common as of recent. The reason for data being biased can be manifold and we will not give a full treatment of this topic here, but rather point out several examples of such reasons. We refer readers to Ntoutsi et al. (2020) for a full review of the issue of biases in data. As an overview, data can be biased due to the way in which it was gathered (methodological mistake) where for example members of a certain demographic were more likely to provide answers. Here it seems important to note that the concept of bias can be considered as relative and perspective-dependent, where the data is biased with respect to some intended representation. For instance data that can give a good representation of a certain demographic would be expected to mirror statistical properties present in that demographics. Imagine some data represents a female-dominated group. Data points coming from females would be expected to dominate and a data set with a 50–50 split between male and female data points could potentially be biased (in that it would over-represent the male data points). On the other hand if the same data set was to be used in a decision making process it could be considered biased to some extent as the opinion of the minority group (males in this case) would be likely to be under-represented. Here we see that the same data set can be considered biased or not depending on its intended use.

Considering the kind of data many machine learning algorithms are trained on (e.g. huge Internet web-crawls for large language models) they can be considered as a mirror to the particular society that generated the data. Thus, any biases present in the given society will be reflected in the model trained with that data. Similarly, the minority positions and perspectives are likely to be lost if not aligned with the majority's positions. In this sense machine learning algorithms exhibit the 'tyranny of the majority' property present inherently in most democratic systems of governance. When considering the data as the source of 'unfairness' it appears clear who is to 'blame' for potential lack of 'fairness' of the algorithm. While it seems difficult to assign responsibility to the society as a whole the data-induced biases of algorithms are in fact only a reflection of same biases in the data-generating process.

On the other hand, the data used for training is of course selected by a human being or a group of humans (though the process continues to become more and more automatized). As such the people selecting the data also bear some responsibility for making sure it is appropriate and representative for their intended use case. A selective filtering of data could also be a source of biases in the model just as much as the inherent bias of the data-generating process.

6.2.1 Assigning Responsibility

Let us now discuss the problem of assigning responsibility when the 'unfairness' of an algorithm is stemming from the object source. In most cases the responsibility for the 'unfairness' stemming from the biased data falls directly with the designers and more generally the entity responsible for creating the given AI. The designers can be reasonably expected to make sure (as a form of due diligence) that the data used for training is not biased. There already exist tools that facilitate the process of finding biases in the data sets (Bellamy et al., 2018) and applying them should be a routine process in designing data-driven learning AIs. This due diligence is of special importance for AIs that are to be deployed in potentially dangerous settings (e.g. self-driving cars) or any kind of situations, where AI's decision can affect humans.

The world changes but the data does not The data is outdated (pun intended), and no longer reflects the 'world' properly. The world has changed and invalidated the data that was used for training. In this case the responsibility of the designers would seem to be perhaps proportional to the predictability of the said change. If the designers could not have been reasonably expected to predict the world changing in such a way their responsibility is limited. *The order of the data matters* During the training of AIs it is common practice to randomly sample points from data. This is typically done to ensure that the AIs are not biased by the order that the data is presented to them. To overcome this potential source of bias an AI may be trained and re-trained many times with different data orderings. In the end an average is taken over all of the trained AIs. But what if the randomization process does not work as expected? This means that the data is presented to an AI in a particular order which influenced the training and results which may not have been achieved by a different order, as in curriculum learning (Bengio et al., 2009; Wang et al., 2021). Therefore it is not enough to know that the data is not biased, but the order must also be un-biased, and figuring this out a priori may be prohibitively expensive as it may require training the AI model in full.

6.3 Objective Source

The loss or reward function is used to quantify the accuracy of a given classification or action done by the algorithm. In a typical training setting data is presented to the algorithm and it attempts to guess the label of each data point (we will deal with the setting where the agent is embodied and deployed in the Sect. 6.4). The loss is then used to adjust the parameters of the algorithm so that the correct label is guessed or the best action chosen for each training point. The loss function can penalize certain mis-classifications more heavily than others thus potentially introducing a bias (which might be intended or not). Variations of such biased loss functions have been used to address the issue of training data imbalances (Jain et al., 2021). It is quite apparent that biasing a loss function could lead to both fairer outcomes and less fair ones, depending on the specific bias introduced. Thus, a biased loss function could be one source of potential 'unfairness'.

A more insidious and harder to identify source of 'unfairness' comes from a loss function that has not been intentionally biased. For AI models, which are trained using labeled data the loss function depends strongly on the labels themselves. In that case the biases potentially present in the labels would be classified as belonging to the object source as they are part of the data. On the other hand many AI algorithms do not rely solely or at all on labeled data but rather use a reward function to quantify the benefit of certain actions given a certain state. Many such reward functions, selected by the designers, have been shown to lead the AIs into unintended behaviors, which can certainly end up being 'unfair'. Imagine a traffic signal controlling AI, whose goal is to maximise the sustainability of the traffic system by minimizing emissions. Such an AI actually learns to stop one arm of traffic at an intersection for ever, since the cars generate most emissions when stopping and starting (Korecki et al., 2023). This is an example of reward mis-design or mis-specification. A similar phenomenon, reward mis-generalization, occurs when the AI finds ways of satisfying the reward that were not intended by the designers. AI with such reward are at a very high risk of leading to highly 'unfair' situations (as in the traffic signal example, where half of the users of the system will never be allowed through an intersection, while the others get an eternal green signal). In general reward design is a growing area of interest in the AI community and there are many issues associated with it. For one the reward functions are often, in a sense, over-fitting to particular algorithms and architectures making it difficult to realize, if a given reward is applicable generally or only to particular types of algorithms (Booth et al., 2023).

A further issue, that commonly occurs in practice, is that in many cases it is not possible to directly optimize the intended goal function. In such cases a 'surrogate loss' is chosen. Examples of such 'surrogate loss' could be using grades as a surrogate measure for intelligence or healthcare costs for a proxy for measuring health. The choice of the 'surrogate loss' can further lead to consequences, which might be hard to predict at first (Ledford, 2019).

6.3.1 Assigning Responsibility

The problem of assigning responsibility for AI algorithms which are 'unfair' as a result of reward mis-design appears to be a more challenging problem than in case of the object source. After all the essence of the problem is that a reward, which appears reasonable (e.g. minimizing emissions in a traffic system) might lead the AI to learn to act in a way that is highly 'unfair'. Here again the designers and entities producing given AIs (e.g. corporations or governments) must be expected to follow with due diligence. The research into ways of avoiding reward mis-design is a growing field and already some procedures have been proposed (Knox et al., 2023). The difficulty of the problem as well as the risk of it occurring even with good intentions of the part of designers also points to the importance of diligent testing and perhaps the need for external certification entities. Such entities could run the required tests to detect any mis-designs in the reward, certify the AIs and take responsibility for the appropriateness of the reward design. *A global objective harms the individuals* Imagine a recommendation system, which operates to maximise the global objective function which in this case is social welfare (Carissimo et al., 2023). In principle this

sounds fair—an equal treatment of all. However, to follow the objective the recommender may need to make mis-aligned recommendations which may seem manipulative. Is it then fair to manipulate the individuals to achieve a fair global outcome? *The world has changed but the objective has not* In an open system it is reasonable to expect that the desired goals change over time in adaptation to the current situation. It is a similar case as when the world changes and the data does not. Here as well the responsibility of the designers would be proportional to the predictability of the change.

6.4 Process Source

The process source is perhaps the most interesting and least explored source of 'unfairness' in AI models and algorithms. As such the process source might involve both the data and the reward but refers exclusively to an agent, which is deployed in an open environment and able to interact with it. If the environment is open and sufficiently complex it becomes impossible to predict a priori what kind of data the agent will be exposed to. If the agent is continuously learning, the data it interacts with might influence its subsequent decisions as well as the subsequent data it is exposed to (feedback loop). In a similar way as the agent explores an open world it might find ways of satisfying its reward function in ways that were not thought of by the designers.

Consider the Microsoft chat-bot Tay (Suárez-Gonzalo et al., 2019), which became racist, sexist and radically politicized within hours of its deployment to Twitter. Tay's character, unintended by its developers, was the result of a coordinated effort on the part of some Twitter users, who tried to affect it by the kinds of conversations they had with it. As a result Tay has been taken offline after around 16 h of functioning. Moreover, the subsequent chat-bots deployed by different companies have been carefully prepared to be able to withstand these sort of adversarial behaviors on the part of their users [not fully successfully (Borji, 2023)].

The process source points to a highly relevant tension between continuous learning in deployment and the increased risk of developing 'unfair' behaviors. Since none of the current AIs understand the concept of fairness (they are only able to optimize for some preset goal) it is impossible for them to judge whether a certain learned behavior is fair or unfair. This further confirms our interpretation of AIs as non-normative eternal optimizers. Therefore, for the AI models that are continuously learning in open environments significant work is performed by the human annotators who are supposed to (among other tasks) eliminate harmful content from the training data (Borji, 2023). Thus the level of openness is somehow restricted by the actions of human annotators (again highlighting that the AI is by no means free).

Here we come back to the statement that we made in Sect. 6 that *AI is (un)fair when society is (un)fair*. Indeed, an AI deployed in a perfectly fair society, where no one would attempt to skew its models by injecting it with biased data, would likely not need an army of annotators. Alas, our society appears not to be fair and so the AIs that we produce, as much as we might struggle, will end of mirroring some if not all of our own unfairness.

6.4.1 Assigning Responsibility

For the process source assigning responsibility appears the most challenging out of all the potential sources of 'unfairness'. Before, we never indicated users as potentially responsible. Indeed, how could the users be expected to know that the data used to train the AI was biased or that the reward was mis-designed. In the process limit perhaps, when we consider AIs which can be continuously learning, a user, who knowingly exposes the AI to biased data could potentially be considered responsible to an extent. Here we refer to Borji (2023) where the authors claim: 'We conclude that, while all the actors interacting with the chatbot share the responsibility of its actions, it is only Microsoft who must account for these actions, both retrospectively and prospectively.' While some of the responsibility falls onto the users, especially these who intentionally worked towards biasing of the algorithm, only the company deploying the algorithm is accountable.

The AI continues learning and exploring An AI that is constantly learning will be sampling experiences from the environment in which it finds itself. It may learn categorizations of things that the creators no longer have control over. It may also explore the environment in ways which were unpredictable with the use of (pseudo) randomness. An AI which spends a long time in such a state may get arbitrarily far from the original state it found itself in when it satisfied all of the fairness criteria. In such a situation, who is accountable for the things the AI has learned when the experiences and data points were self-selected by the algorithm in interaction with a complex environment? *The AI is attacked* As in the example of Tay given above the AI might come under deliberate attack by some of its users. As indicated the responsibility is shared by all interacting with the AI but the accountability falls on the designers.

7 Conclusions

In Sects. 2 to 5, we have shown that due to the normative and physical characteristics that any system must possess to be treated as a responsible entity, AI algorithms cannot, in principle, be attributed such responsibility. However, there have been increasing voices, especially in popular media, that seem to implicitly assign some responsibility, agency or even normativity to AI. We take the opportunity here to highlight the danger of such discourse. By assigning any of these properties to AI it becomes possible to mis-assign responsibility to it. It opens up a possibility for people to avoid responsibility and mis-place it on the AI. This is why we believe that the present work is of great relevance in delineating the boundaries of AI's normativity and arguing for its utter lack of responsibility. By following our arguments one can see through anyone who would try to avoid responsibility for one's actions by referring to AI (e.g. 'I was just following what the AI told me to do.').

Even though they lack proper normativity, we have shown in Sect. 2 that AI algorithms can be appropriately qualified as (un)fair in an analogical sense. An important point we noted is that such analogical predication presupposes that the system in question has a relationship of participation with a set of actions that can be

qualified as (un)fair in the primary sense of the term. The conclusion to be drawn from this semantic argument concerning the (un)fairness of AI algorithms is that a proper understanding of the ethical implications of these technologies' functioning requires the examination and clarification of their relationships of participation with the authentic source of normativity.

The arguments developed in Sect. 6 regarding the various sources of unfairness for AI algorithms outline the basis for such an investigation. The different sources represent modes of participation for AI algorithms through which they can be qualified as (un)fair by analogy. The complexity of these relationships becomes evident in light of the challenges we have raised regarding the attribution of responsibility and the edge cases for each of these sources. A deeper investigation of these relationships is therefore required and, in our view, should provide the framework for further inquiries that will focus on each of these sources and the problems of responsibility assignment they delineate.

This being said, one further interesting point that we can gather so far is that when it comes to assigning responsibility the AI is much different from most of the technologies that humanity has interacted with so far. In this line of thought, according to us, it would be incorrect to attribute potential negative effects to the user of this type of technology. Unlike other types of technology,[8] one should not assume that the user is expected to use a given tool in the correct way and follow all normative rules while using it so that if the user misuses the tool or uses it in a way that breaks the rules, they are held responsible for the effect of such use.

The reason for this difference with AI is that the users cannot be reasonably expected to be able to predict the effects of using AI (in some cases not even the developers are able to make such predictions). The AI models of today are essentially black boxes whose actions are not easily determined. On the one hand this could be considered the goal of the AI, so as to create entities, which can surprise us and act with an illusion of perceived independence. On the other hand deploying such surprising entities, especially in open complex environments such as human societies, makes it challenging to account for their actions within a normative framework that these societies have been used to.

As we have interpreted the AIs of today as essentially optimization processes one could ask if the solution to the 'unfairness' problem was not as simple as setting fairness as the objective for the AI to optimize. The question is a valid one, nevertheless, we argue that any form of optimization that would explicitly account for fairness would run into significant issues. For one optimization operates exclusively on quantities, but quantifying fairness is a problem in itself (at the very least there is no one accepted way of quantifying or even defining fairness). Secondly, the optimization would need to explicitly optimize for a given goal and on top of that fairness.

[8] For instance even some inherently dangerous technologies, such as weapons, are the responsibility of their users. Although here the matter of the ultimate moral responsibility is more dubious since one could claim that the main intended use of weapons is always wrong. Through that reasoning one could also assign responsibility to the manufacturers. Nevertheless, as it is, all over the world, the responsibility for murders and mass shootings is assigned to the shooters, not the manufacturers.

This would necessarily be a multi-object optimization (unless the fairness quantification and the goal would be collapsed in a fixed way into a single value). This on the other hand would lead to a Pareto front of solutions, where one could sacrifice fairness for the performance on the goal and the other way around. How then would the final solution be chosen? In many cases it would be likely that sacrificing fairness would lead to better performance on the goal and vice versa making it highly unclear where the balance lies. Nevertheless, it is important to note, that in some particular cases it might be possible to actually account for fairness in a way that also benefits accuracy (Blum & Stangl, 2019; Dutta et al., 2020; Wick et al., 2019).

Furthermore, a danger similar to that of responsibility avoidance mentioned above is that of the society relegating more and more crucial functions to the control of powerful AI systems that are themselves beyond control (or are controlled by the select few). Then the controllers of such society [if there even were any, see (Forster, 2021)] could easily avoid any responsibility by referring to the powerful AI that controls everything (while they are the ones doing the controlling in the end). Such hypothetical situations are important to consider to make it clear what the risks are of failing to clearly establish who is responsible for the AI.

This relates to the general problem of 'algocracy' or the rule by algorithm. Algocracy is a term used to refer to a society, which is run and controlled with the help of algorithms and advanced AI models. Based on the arguments we have presented in the preceding pages, algocracy is a dangerous oxymoron. To the extent that AI algorithms do not act, in the proper sense of the term, but can only function by participating in actions, they cannot exert any *krátos*. In a society run and controlled with the help of AI, the true rulers are those, who control and choose the models (be they AI or otherwise). However, as in the previous case, it is very easy for these rulers to avoid any responsibility for their actions and just blame any wrongs on the models and AIs. Moreover, it is often envisioned the people living in such a system would likely be unable to affect or select the models that are being used. Perhaps a solution to both the risks of dictatorial rule by algorithm and the problems of AI responsibility could be to some extent addressed by creating AIs which are more user-driven. By that we mean AIs on which users can have direct and meaningful influence (e.g. setting the reward function to be pursued). Responsibility assignment would also be easier to make for user-driven AIs. In such cases the users would have more control but also more responsibility over the effects that the AIs bring.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich. Funding was provided by HORIZON EUROPE European Research Council (Grant No. 833168).

References

Aristotle. (1998). Metaphysics, books g, d and e (C. Kirwin, Trans.). Clarendon Press.

Arp, R., & Smith, B. (2008). Function, role and disposition in basic formal ontology. In: Proceedings of bio-ontologies workshop, intelligent systems for molecular biology (ISMB), Toronto (pp. 45–48).

Baird, A., & Maruping, L. M. (2021). The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts. *MIS Quarterly, 45*, 315–341.

Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv: 1810.01943

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. Proceedings of the 26th annual international conference on machine learning (pp. 41–48).

Blum, A., & Stangl, K. (2019). Recovering from biased data: Can fairness constraints improve accuracy. arXiv:1912.01094

Boettke, P. J., & Subrick, J. R. (2002). From the philosophy of mind to the philosophy of the market. *Journal of Economic Methodology, 9*, 53–64.

Booth, S., Knox, B.W., Shah, J., Niekum, S., Stone, P., Allievi, A. (2023). The perils of trial-and-error reward design: Misdesign through overfitting and invalid task specifications. AAAI conference on artificial intelligence.

Borji, A. (2023). A categorical archive of chatgpt failures. arXiv:2302.03494

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of compstat'2010: 19th international conference on computational statistics, Paris France, august 22–27, 2010 keynote, invited and contributed papers (pp. 177–186).

Brandom, R. (2019). *Heroism and magnanimity: The post-modern form of self-conscious agency*. Marquette University Press.

Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research, 79*(2), 251–278.

Carissimo, C., & Korecki, M. (2023). Limits of optimization. *Minds and Machines, 6*, 1–21.

Carissimo, C., Korecki, M., & Dailisan, D. (2023). Strategic recommendations for improved outcomes in congestion games. *ResearchGate*. https://doi.org/10.13140/RG.2.2.33145.24167

da Silva Oliveira, D. G. (2022). An alternative view for scientific models based on metaphors: A case analysis from Darwin's use of metaphors. *Principia: An International Journal of Epistemology, 26*(2), 347–373.

Davidson, D. (2001). *Essays on actions and events*. Oxford University Press.

De Poehl, Z., & Royakkers. (2015). *Moral responsibility and the problem of many hands*. Routledge.

De Sio, M. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy Technology, 34*, 1057–1084.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*, 429–452.

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., & Varshney, K. (2020). Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. International conference on machine learning (pp. 2803–2813).

Ezenkwu, C.P., & Starkey, A. (2019). Machine autonomy: Definition, approaches, challenges and research gaps. Advances in Intelligent Systems and Computing.

Felin, T., Kauffman, S. A., Koppl, R. G., & Longo, G. (2013). Economic opportunity and evolution: Beyond landscapes and bounded rationality.

Forster, E. M. (2021). *The machine stops*. Phoemixx Classics Ebooks.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Griffiths, M. (2022). Is lamda sentient? *AI & Society, 39*, 1–2.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.

Jain, B., Huber, M., & Elmasri, R. (2021). Increasing fairness in predictions using bias parity score based loss function regularization. arXiv:2111.03638

Knox, W. B., Allievi, A., Banzhaf, H., Schmitt, F., & Stone, P. (2023). Reward (mis) design for autonomous driving. *Artificial Intelligence, 316*, 103829.

Korecki, M., Dailisan, D., & Carissimo, C. (2023). Dynamic value alignment through preference aggregation of multiple objectives. arXiv:2310.05871

Landgrebe, J., & Smith, B. (2022). *Why machines will never rule the world; Artificial intelligence without fear*. Routledge.
Laukyte, M. (2017). Artificial agents among us: Should we recognize them as agents proper? *Ethics and Information Technology, 19*, 1–17.
Ledford, H. (2019). Millions affected by racial bias in health-care algorithm. *Nature, 574*(31), 2.
Lemoine, B. (2022). Is lamda sentient?—An interview. Medium. Fecha de publicación.
List, C. (2021). Group agency and artificial intelligence. *Philosophy and Technology, 4*, 1–30. https://doi.org/10.1007/s13347-021-00454-7
List, C., & Pettit, P. (2011). *Group agency. The possibility, design, and status of corporate agents*. Oxford University Press.
Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., Ge, B., et al. (2023). Summary of chatgpt/gpt-4 research and perspective towards the future of large language models.
Martinelli, E. (2023). Toward a general model of agency. Argumenta.
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living (Boston studies in the philosophy of science)*. Springer.
McDowell, J. (1998). *Mind, value and reality*. Harvard University Press.
Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., et al. (2020). Bias in data-driven artificial intelligence systems-an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1356.
Popa, E. (2021). Human goals are constitutive of agency in artificial intelligence (AI). *Philosophy & Technology, 34*, 1731–1750.
Popper, K.R.S. (1994). Knowledge and the body-mind problem: In defence of interaction..
Rychener, Y., Taskesen, B., & Kuhn, D. (2022). Metrizing fairness. arXiv:2205.15049
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417–424.
Searle, J. R. (1983). Intentionality: An essay in the philosophy of mind.
Sellars, W. (1963). *Science, perception and reality*. Ridgeview Publishing Company.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., & Den Driess-, Van. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.
Suárez-Gonzalo, S., Mas Manchón, L., & Guerrero Solé, F. (2019). Tay is you: The attribution of responsibility in the algorithmic culture. *Observatorio, 13*(2), 14.
Thoppilan, R., Freitas, D.D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Le, Q., et al. (2022). Lamda: Language models for dialog applications.
Tollon, F. (2023). Responsibility gaps and the reactive attitudes. *AI and Ethics, 3*, 295–302.
Wang, X., Chen, Y., & Zhu, W. (2021). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4555–4576.
Wick, M., Tristan, J.-B., et al. (2019). Unlocking fairness: A trade-off revisited. Advances in neural information processing systems.