PERSONAL AUTONOMY AS AN ETHICAL FOUNDATION FOR OPAQUE ALGORITHMIC DECISION SYSTEMS

Francisco Lara

Abstract

AI is becoming a highly efficient instrument for decision-making in relation to the distribution of goods, services or prerogatives in different public and private administrative systems. The problem is that the greatest efficiency in this area is obtained thanks to black box algorithmic systems for which, due to their technical characteristics, explanations cannot be provided of how they have made their decisions. This has led a number of scholars to actively question the use of such systems, arguing that the lack of explanations in important decisions for the subjects poses a serious threat to their autonomy and, with it, an attack on their dignity.

In this article the basic idea is accepted that the opacity of these systems implies, in principle, an erosion of personal autonomy. However, it is also argued that this idea does not rule out the possibility that the lack of explanations may at times be justified. To support this thesis, we first analyze the interpretations of three basic criteria (agential, justificatory and normative) that have given rise to the aforementioned position, based on dignity, which is the object of critique here. Alternative interpretations of such criteria are then given, from which to deduce a certain flexibility in the demand for transparency in algorithmic decision-making systems. Finally, three principles are derived from this proposal to ethically regulate the use of this type of system.

Keywords: AI Ethics – Explainable AI – Opacity – Decision making systems – Personal autonomy.

"Algorithmic governance" encompasses a broad spectrum of practices that order and regulate social matters through the use of AI. Part of this new reality is related to the use of algorithmic systems in administrative, judicial, healthcare or commercial instances for decision-making regarding the concession of certain goods, services or prerogatives to individuals. These consist of bank loans, jobs, work bonuses, conditional release of prisoners, donated organs... The idea behind the use AI for these types of decisions is increased efficiency and precision, overcoming the physical and epistemic limitations, biases and, in some cases, the arbitrariness of humans.

The problem is that the great benefits of entrusting these decisions to AI would be achieved with algorithmic systems that, because of their self-learning capabilities, end up being highly flexible, but also extremely opaque. They are referred to as "black box" systems because they can produce responses different to those predicted and for which experts may have no explanation. These systems, which would improve social decisions at the cost of less transparency, will be referred to as "Opaque Algorithmic Decision Systems", hereinafter OADS.

OADS have usually been questioned due to the distrust they generate, either because of the simple fact they do not provide explanations for decisions, or because their lack of interpretability could prevent the detection of deficiencies or morally unacceptable patterns (Biran & McKeown 2017, Doshi-Velez & Kortz 2017, Vredenburgh 2022, Domingos 2015, Ribeiro et al. 2016, Lyons et al. 2016). Here, however, we are interested in another type of criticism of OADS; namely, that due to their lack of transparency, their

use can pose a direct threat to personal autonomy. These systems would negatively affect the capacity of human beings as regards self-determination, ability to develop own sense of worth, make decisions about what matters most to them and act accordingly (Rubel et al. 2021: 21, Colaner, 2022; Vaassen 2022, 7-8)[1]. This criticism is independent of the previous ones and is evident in the fact that a reliable, accurate and unbiased OADS could still be questionable because, by not providing explanations of decisions made, it entails, insofar as it undermines the autonomy of affected parties, an attack on their dignity. This has led some authors to contend that because of this fact, and irrespective of their potential benefits, OADS should be outlawed entirely (Rudin 2019, Vredenburgh 2022), or at least in certain important contexts (Selbst & Barocas 2018, Colaner 2022: 237, Grant et al. 2023: 4, Rubel et al. 2021: 70-71). From this point onwards we shall refer to this as the "Dignity Autonomist Argument" (DAA).

The essential aspect of this argument against OADS could be characterized, in my view, by a particular specification of the three basic criteria set out below.

a) The first criterion is agential, regarding *who can be decision makers* in public contexts where OADS are used. This premise is understood by advocates of the DAA in strictly anthropocentric terms: OADS are unacceptable because, at least in regard to certain matters of great relevance to the humans concerned, decisions can only be made by other humans. This premise would be unrelated to the issue of opacity, but it is a precondition and defining consideration for the DAA.

b) The second criterion would be justificatory and seeks an answer to the question of *why decisions must be explainable*. Those who defend the DAA argue that this is the case because opacity infringes upon personal autonomy and this is serious because this autonomy constitutes an essential feature of human nature which, therefore, dignifies it.

c) Lastly, there is the normative criterion, which would demand a pronouncement *on the type and degree of the obligation to give explanations* that would result from the recognition of autonomy as a value. The DAA would incline towards a deontological perspective, from which moral agents would be strictly bound to prevent such decisions from being made by algorithmic systems that are not transparent.

To summarize, according to the DAA, the fact that certain important decisions regarding citizens are made by artificial systems and, furthermore, no explanation is given for these, shows a lack of respect for that which ennobles humans, an affront to their dignity as autonomous beings. For this alone, the very use of algorithmic systems to reach such decisions is, in itself, incorrect.

In that which follows, I shall endeavor to demonstrate that this strict interpretation of autonomy is not valid, and that the relevance of autonomy to the problem of ODAS should be understood differently. To this end, in the following three sections I shall critically analyse each DAA interpretation to the three criteria previously set out, and shall

[1] This critique is different from those others that, also based on human autonomy, would be critical of OADS on the grounds of an alleged right to be informed (Selbst & Powles 2017, Bryce et al. 2017, Kaminski 2019, Lipton 2018), which not everyone shares (Wachter et al. 2017), or of their possible use for manipulation (Yeung 2017, Danaher 2016, Susser et al. 2019).

propose a different interpretation for all of them, but from the shared idea that the explainability demanded of OADS comes from the moral value of autonomy. This alternative interpretation of the criteria will be based on non-anthropocentric and non-deontologistic postulates, thus allowing for a more flexible defense of the requirement to provide explanations for the use of OADS. The proposal as a whole will be referred to as "Consequential Autonomist Argument" (CAA). In the fourth section, by way of conclusion, I shall summarize my proposal and, from my interpretative proposal of the three criteria, present three ethical principles for the regulation of OADS.

1. Should Algorithms Make Important Decisions?

The DAA would be characterized, in its response to the aforementioned agential criterion, by a refusal to allow algorithmic systems to make important decisions that affect human beings. Relieving human beings of such responsibility would be disrespectful or an affront to the dignity of the subjects affected by the decisions.

The intuitive power of this approach must be recognized. But neither should we forget that intuitions are, by their very nature, thoughtless, usually consisting of deep-rooted beliefs that respond to socializing processes rather than reflexive acts, and are therefore always conditioned by shared, traditional and customary norms. These processes would explain, for instance, the generalized belief that there are certain things machines should not do, that this would somehow make our existence less meaningful. However, as history has shown, these types of beliefs, radically opposed to social change caused by technological progress, gradually soften and end up being replaced by others favorable to change. Technologies that have brought with them the modification of important human behaviors through the mediation of devices, such as the non-visual long-distance communication permitted by the telephone in its era, gave rise to immediate reactions of indignation and rejection that, over time, developed into more considered judgments of assumption. Time demonstrated that our intuitions against speaking to somebody through a machine were irrelevant and irrational, especially when one realized the efficiency of doing so. It is highly likely the same will occur with the substitution of humans by automated machines, as is progressively happening with social robots. Could this pattern of tolerance not then also be repeated with algorithmic decisions? For this reason, we should not simply accept the refusal to leave such decisions to the systems when such refusal is based solely on intuition.

Nevertheless, it could be asserted that it is not just a matter of intuition; rather, there are important reasons of an ethical nature to oppose algorithms occupying this role traditionally carried out by humans. Grant et al. (2023) point in this direction, arguing on the basis of reasons rather than intuitions, when they argue that this delegation of functions would be ethically unacceptable because, in situations with major repercussions, decisions will only be respectful, in accordance with "obligations of agential consideration" if they are made by "full-blown moral agents exercising their powers of moral reasoning, and that those agents deliberate in good faith a decision that respects the decision-subject's moral claims on the decision-making process" (23). With the concept of 'full-blown moral agents', these authors seem to refer to beings with deliberative capacities who can feel responsible for how their decisions affect the interests and rights of others. It upholds, therefore, that certain important decisions, such as whether or not to impose a criminal sanction, must always be made by agents who, when making decisions in such a regard, exercise their moral peculiarities and with the appropriate level of concern for such important decisions, taking a special type of

responsibility for the resulting decision, which nobody would take if the decision-making were delegated to an automated system (Grant et al. 2023: 23-24).

Moreover, Grant et al. (2023) add that making the decision in this way (human) "demonstrates an important kind of respect for the decision-subject: she both recognizes and gives appropriate weight to their status as a fellow member of her moral and political community in her deliberations." (24). To support this, they put forward the thought experiment of the existence of the Juror Substitution Policy, in which jurors would be given the choice of being replaced by an individualized algorithmic system that, deliberating in the same way and with the same information as the juror, would do so, and thus the jury could reach decisions on whether or not the defendants should be punished. In the case of members leaving decisions to their algorithmic substitute, even where the results were the same as if a human jury had deliberated on them, the mere fact of having delegated the decision to an artificial decisional system demonstrates an ethically unacceptable course of action, due to the unfair consideration with which the subjects concerned are treated.

This essentially raises two questions. The first is whether or not algorithmic systems could come to work as "full blown moral agents" when making their decisions. The second is whether, getting to the point where they behave as true moral agents, their decision would continue to be unacceptable due to originating from a machine.

In regard to the first, mention should be made of the fact that, in effect, it would at present be difficult to accept that an artificial system could be designed to make decisions similar to those that a moral (human) agent could make when deciding as such, despite the advances in so-called 'machine ethics' in this regard (Anderson & Anderson 2011, Moniz Pereira & Barata Lopes 2020). These advances are still in their infancy, as Grant et al. (2023: 26n) themselves point out. Nevertheless, the probability of these systems being considered agents increases considerably if, trusting in the immense possibilities of deep learning, we also rethink the very concept of what a moral agent is. Thus, a number of authors defend a "relational turn", according to which algorithmic systems could be conceived as moral agents as long as, having a particular and valuable social relationship with other bodies, we behave towards them as if they were (Coeckelbergh 2012, Gunkel 2012). On the other hand, leaving aside such disruptive approaches, the consideration of artificial entities as moral agents has also been defended from a functionalist and non-anthropocentric ontologism. From here it has been argued that, in the short term, algorithmic systems could behave in accordance with moral values from a weak conception of agency as attending to criteria of mere interactivity, autonomy and adaptability (Floridi & Sanders 2004). Some have gone further and come to defend that, in the long term and thanks to predicted technological development, there may be algorithmic machines that act according to moral criteria up to now only present in beings with minds, such as free will, rationality or even imagination (Verbeek 2005, Sullins 2011, Purves et al. 2015, Behdadi and Munthe 2020).

On the other hand, with little development and in a footnote, Grant et al. (2023: 26n) state their opposition to the second question posed above. They doubt there would be any progress in this matter if algorithmic decision-makers made their decisions as full moral agents, since it could always be argued they are not of the "*certain kinds of full-blown moral agents* -such as fellow citizens, a company representative, etc-" of which the subjects affected by the decisions plausibly deserve due consideration. This is questionable for two reasons. First, because it could always indirectly justify the specific status that would be required. In other words, even if the decision were made by the artificial decision-maker, it would respond to indications from those with relevant status, such as designers, groups of experts and stakeholders - or fellow citizens in the case of

substitutes for jurors. Secondly, it could be adduced that what is in the end relevant is not who makes the decisions, but that they do not go against certain moral demands. Let us imagine in the case of the substitute jury that it is made public that certain biases and deliberative deficiencies of some members of the human jury (who had not opted for their algorithmic peer) have caused a situation where they have not allowed the measures recommended by the system to be carried out, measures that turned out to be more just and efficient. In such a case, would we continue to think that algorithmic decisions are inconvenient or disrespectful if they are not finally adopted by human beings in general or relevant humans, in this case human jurors?

2. Why Must Algorithms Give Explanations?

According to the DAA, OADS would be unacceptable because, due to their lack of transparency, they would impede autonomy, a trait essential to human beings, thus conferring upon them a higher value in the face of any other thing that exists in nature. Colaner (2022: 234-6) argues this when, getting his inspiration from the "personhood argument" defended by Selbst and Barocas (2018) for the defense of privacy, he holds that opacity goes against human dignity, as it does not respect the need to know and participate that characterizes all of us. In the specific example of OADS, a type of objectification and vulnerable dependency on the affected individuals would occur because, on the one hand, they would be being impeded from understanding themselves why a specific decision has been made on them and, on the other hand, they would be incapacitated in terms of influencing the process with reasons connected to their possible non-conformity. Coloner (2022: 236) adds, furthermore, that if individuals are unable to significantly participate in the decision-making process, they cannot self-actualize. The nature of our species would include the capacity to govern ourselves and the community, and our full actualization as humans implies that it is we who make the important decisions about ourselves and our community.

It is important for the latter meaning of this critique to be clear. The argument is not that the lack of awareness and lack of participation on the part of the affected parties in the process is negative because it leads to results that are unfair or unacceptable to someone. The process could end up being extremely fair and beneficial to all involved, but, even so, it would be questionable for the sole reason that, as it is opaque, it infringes upon autonomy and, thus, upon the essential and valuable autonomy of human beings. Colaner (2022: 236) underlines this last naturalist meaning of the objection when he recognizes the Aristotelian influence in it, which derives the obligation to participate in deliberative processes in relation to our well-being and that of the community from the fact that the humans are a decision-making species.

However, I consider that basing the demand for OADS transparency on a naturalist justification such as that described by authors including Colaner brings with it the exposure of ADD to extremely consolidated criticism. On the one hand would be the traditional accusations aimed at naturalist theories that say their basic arguments move too casually from the realm of being to the realm of ought-to-be, committing what is known as the 'naturalistic fallacy' (Moore 1903). In contrast, there would be room to argue that these approaches are always accompanied by anthropocentric assumptions ultimately based on not particularly well-founded theoretical and, in many cases, clearly biased, commitments. The very choice of cognitive and deliberative capacity as essential to human beings is debatable. It is difficult for a capacity to be considered as "essential" in either of the two main senses in which the term is used in this debate. In principal, this

is due to the fact that because it is not exclusive to members of the human species, it cannot be something that characterizes it in a differentiating way. There are non-human animals that are also rational. Secondly, neither is it evident that this ability to be rational is essential because it concedes value or dignity to the subjects that possess it. Accepting this would mean having to accept a practical implication that their defenders would not accept willingly: that not all humans deserve to be treated with dignity because not all of them are capable of rational behavior. They would be obliged to argue that, at the risk of falling into a position that is questionable because it is based on a speciesism, certain disabled humans would, on not having that determining capacity, be deprived of the status that others deserve (Pluhar 1995, Dombrowski 1997).

Lastly, even overcoming the above criticism, one could always ask the defenders of the anthropocentrism underlying this naturalism to explain why rationality should be the distinguish feature. Would it not be sufficient for a being to have the capability of suffering (although not deliberate) for us to show concern towards it? Would the fact that it were unable to deliberate in a more sophisticated way be so morally decisive? (Bentham 1789: 282-3, Singer 1975).

Given these serious objections to the naturalism underlying the DAA, the best alternative, in my opinion, would be a prescriptive and utilitarian justification of autonomy, and the subsequent demand for transparency. This moral justification is based on a logic that demands consistency in our judgments and actions. As Hare (1952, 1963) argues, when we defend a moral decision, we do so from a commitment to equal consideration for the welfare interests of all concerned. This commitment extends beyond humans to include any living being with welfare interests, such as some non-human animals. However, this doesn't mean we treat everyone the same. The principle of "equal consideration of equal interests" requires a nuanced approach that allows for differentiated treatment. For instance, a person's interests are relevant when they stem from their rational self-determination—the ability to decide what's best for themselves. This requires a higher degree and quality of autonomy. In contrast, the autonomy of other sentient beings is limited to their most basic desires, like not suffering or dying. Therefore, while personal autonomy remains an intrinsic value, its justification is no longer based on naturalistic or anthropocentric views.

Therefore, the lack of explainability on the part of OADS would not in itself be incorrect because it dehumanized us; rather, it would be incorrect because it might in certain situations fail to meet the requirement of morality to contribute to the maximization of the well-being of all concerned. Without explanations, affected parties cannot determine for themselves and demand what really benefits them (according to their own judgment) (Raz 1986, Atkins 2000, Bratman 2000, 2018, Ismael 2016). In the words of Rubel et al. (2021), a lack of transparency would be in breach of the “informational prerequisite” of the autonomy of individuals to have access to information required to exercise their cognitive and practical agency in accordance with their self-chosen values and commitments. Especially in major decision-making contexts, concern for the autonomy of the individual would, in principle, require giving them the information that would enable them to assess and understand their situation in order to determine how to act (cognitive agency), as well as to behave according to their values and commitments (practical agency). In particular, in the case of decisions made by OADS, the lack of explanation would prove to be a serious hurdle for the autonomy of subjects. To this end, there would be a prima facie obligation to avoid this impediment by providing subjects with relevant information on how the system works. You can object to or accept the functioning of a system, by virtue of whether or not its decisions are wrong, immoral or go against your interests only if you know how it operates (Rubel et al. 2021).

Moreover, without this functional information from the systems, there would be a detriment to practical agency, depriving us of clues on how to modify our actions in similar decision-making situations and thus achieve better results. If we are not told why we have not been selected for a job position, we will not be able to undertake actions aimed at future success (Lombrozo 2011, Vaasen 2022: 6-7), and in this way would be denied opportunities to choose and progress, where appropriate, towards increasing our well-being.

3. To What Extent Should Algorithms Give Explanations?

It follows from the previous section that the critique of the naturalistic rationale of the DAA does not necessarily lead to a disregard of autonomy, an element that could be highly valued, as we have just seen, by virtue of the fact that, due to its relevance for personal well-being, affords meaning to and enriches our lives. Nevertheless, the recognition of any value can translate into different regulatory proposals. In the case of the DAA this is clearly deontological. According to deontologism, we cannot truly understand the fundamental importance of a value unless we see it as an absolute rule that restricts us from only considering the consequences of our actions. In other words, if a certain value is truly important, it should lead us to create strict, non-negotiable rules. If we fail to create these rules, we don't genuinely grasp what it means for something to be valuable (Nagel, 1986: 182). This interpretation suggests that the value of something isn't judged by its outcome, but by our duty to respect it. When we see something as valuable, our actions shouldn't be based on what will produce the best result, but on our obligation to follow the rules that protect that value. Therefore, true respect of something of value does not then permit us to negotiate with it, sacrificing it either to obtain another weightier value or in order to end up with things with that value (Fried 1978).

In accordance with this, from the perspective of the DAA, the mere fact of doing something contrary to human autonomy would be wrong, regardless of the best consequences that this could have, both for the promotion of that same value in other areas or subjects, as well as for the possible promotion of other values that are also highly relevant. It can thus be argued that the concern of the DAA for autonomy in opaque contexts is not contingent and is what follows from the contention of Grant et al (2023) when they argue that the obligation to show due consideration to subjects gives decision-makers a strong reason to avoid black box systems in many contexts.

Nonetheless, here we will defend a contingent proposal according to which, by virtue of technical progress and circumstantial aspects, the concession of certain amounts of opacity (and consequent undermining of autonomy) could be justified if it leads to improved balances in terms of efficiency or reliability with which to ultimately increase the autonomy of all. I believe that such flexibility better adjusts to the context and aim of OADS. If their intention is to basically decide on the allocation of benefits or obligations to certain subjects, their autonomy is not diminished by failing to provide them with omnipresent and complete explanations. They need explanations, of course, but only those which are causally relevant: those that indicate what essentially motivated the decision and what changes in the inputs *robustly* correlate to certain changes in the output of the decision (Strevens 2013, Vaassen 2022: 5). These types of explanations are important because they are crucial to safeguarding the autonomy of the individuals affected, and therefore the question of who determines when and how they should be provided is not a minor one. The affected subject cannot be the one to determine this because they cannot determine the relevance of the causal aspects prior to the decision.

Nor could this function fall exclusively to the algorithmic systems themselves, as this could generate mistrust among those affected. Therefore, from my point of view, it is something that should be the responsibility of an ethics committee composed of human experts in the subject matter being decided upon, computer designers and AI ethicists, who would establish in which specific situations the subjects affected by algorithmic decisions would have the right to receive causal explanations for such decisions.

In addition, explanations should not always include all the decisive causal factors, as some will not help to understand the decision or may even lead to confusion, either because too much information makes it difficult to grasp the essentials or because they provide irrelevant information (Vaassen 2022: 6). This should not initially be a problem, as there is agreement among experts that to be aware of these crucial regularities between changes in inputs and outputs does not require detailed knowledge of how the system works (Newel 1982, Campbell 2008), or knowledge of all possible correlations; rather, only those that are basic and relevant to satisfy the autonomy of the subjects. It should not be forgotten, either, that giving explanations is a costly process involving time and effort, which are not always available to us (Doshi-Velez & Kortz 2017).

Mention might also be made here, in defense of algorithmic systems not being required to be completely transparent, that it is ultimately something not demanded of humans when making decisions in the same settings, either. Why, then, must it be demanded of systems? (Zerilli et al. 2019, 2022). In both cases, complete transparency is impossible. In the case of humans, this is because their decisions also respond, in many instances, to black boxes. They are far more opaque than we would like to believe, given that a number of recent psychological studies raise doubts about whether we have reliable introspective access in regard to our own decision-making processes (Schwitzgebel 2019: 4.2.1). With this being the situation, relying on human decision makers would also involve a risk, similar to that associated with OADS, that prohibited inference rules derived from human biases would be implemented and not perceived.

On the other hand, it is not merely a question of an introspection deficit. Sometimes, opacity stems from the difficulty of human decisions. We often rely on judges, committees and doctors to make decisions that, because they are equally complex, as in the case of those associated with OADS, are not fully understood by those affected, and sometimes not even by experts (Zerilli et al. 2019, London 2019). On other occasions, even in less complex decisions such as those relating to job recruitment, it is also true that we make do with explanations that are not very detailed, in which imprecise information is given. For instance, that candidate X lacks the relevant professional experience, etc.

Thus, for all of the above, if we do not wish to be subject to an accusation of ‘double standards’, the transparency requirement for OADS should not be absolute and should be limited to restricted access to the explanatory keys to the decision (Lipton 2018; Cappelen & Dever 2021).

Another more theoretical objection to the non-contingent regulatory application of autonomy would be that it is based on an idealistic conception of the scope of application of norms. Behavior is required that fully respects the value of autonomy in a context wherein those to whom the norms are addressed lack the personal and social conditions to be able to act in accordance with such a requirement. This translates into the formulation of strict judgments that ultimately imply this lack of respect for autonomy and dignity, the avoidance of which was precisely the intention. Generally, flesh and blood subjects are empirically conditioned and, if we are not aware of this, we end up demanding more from them than we should and punishing them for what they do not deserve. Is it fair to blame someone who, due to their personal or social limitations, is not

completely autonomous? Would doing so not amount to undignified and disrespectful treatment?

Transferring this objection to the sphere of OADS, avoiding their functioning by virtue of their lack of complete transparency would be like demanding the ideal situation when the appropriate conditions for it are not forthcoming. All that would be achieved with this is would be to dispense with a much more efficient and impartial distribution of benefits and prerogatives than anything humans could offer. Unfair treatment would therefore consist of impeding a resulting state of things where people would equitably acquire, in a more efficient way, those goods that would progressively increase their spaces of self-determination and, with it, their capacity to demand and understand more explanations in the future. A lack of complete transparency would be justified by this horizon of a broader future autonomy.

To all of the above, as an alternative I propose an understanding of autonomy as a value to be promoted and, therefore, implemented as a contingent value. In this other model, full autonomy would not be a condition for making moral judgments, just a horizon to aspire to. From this consequentialist point of view, it would sometimes be justified, because of the possible counterproductive effects on the individual and society, to ignore the demand for explanations.

4. Ethical Principles for the Use of OADS

In the preceding paragraphs the essential lines have been drawn of a theoretical alternative to the DAA that, as already mentioned in the introduction, could be called, as a whole, the "Consequential Autonomist Argument" (CAA). This would share with the former the commitment to an obligation to regulate the opacity of OADS so as to safeguard the autonomy of human beings, especially those individuals affected by the decisions of these systems. However, if in relation to the CAA we abandon the anthropocentric and deontologistic assumptions, with a consequentialist justification of our obligations, the demand to uphold autonomy brings us to a different conception of when it would be ethical to use OADS. Below I shall set out those that, in my opinion, would be the three basic principles which regulate said use from the CAA.

1. ***Fundamental Liberal-Utilitarian Principle***. Defending OADS transparency from an idea of personal autonomy that appeals to dignity is open to criticism for its naturalist and anthropocentric tenets. In opposition to this, I have defended that the main underlying reason for decisions on matters of social distribution and administration, be they made by humans or algorithmic systems, should be the promotion of the autonomy of the affected individual, insofar as it is essential for increasing general well-being. This is ultimately based on the same prescriptive logic of the moral discourse. For all of this, it must be understood that the decisions of algorithmic systems will be valid to the extent they either do not reduce or they maximize opportunities for self-determination and self-achievement of personal well-being. Given the empirical evidence of how transparency in important decisions is usually critical in order to take advantage of such opportunities, it turns out that the explanation for such decisions is, prima facie, an ethical right of all affected subjects.

2. ***Principle of Conditional Agential Relevance***. As we have seen, initially there are no ethical reasons for presupposing that decisions that occupy us should only

be made by humans. Such presuppositions are usually based on little more than prejudices on the role that devices should play in our lives or in arguable conceptions of respectful treatment. Therefore, it would be irrelevant for decisions to be made by humans or algorithmic systems, as long as with them the *fundamental liberal-utilitarian principle* is respected.

Now, following this principle could justify in certain situations a moral obligation to choose an algorithmic system over its human equivalent if, compared to the latter, the use of the former implies a greater attainment of autonomy, in the sense that it can achieve better levels of precision in the accuracy of its decisions with the same level of opacity. This would not be strange nor unusual, as there is evidence that algorithms, whether more or less interpretable, often make more accurate predictions than human experts (in the case of medical predictions, see McKinney et al. 2020), who often choose from biased approaches (Jolls et al. 1998) and based more on experience and intuition than on statistical analysis (Meadow & Sunstein 2001, Klein 2017). Furthermore, black box algorithms in particular, used by OADS, can make even more precise predictions than linear models, as they base themselves on more flexible, particularized and complete data processing (Breiman 2001, Caruana et al. 2015).

Nevertheless, this should be taken with a great deal of caution, as these systems may not be as accurate in the field as they are in theory and the laboratory. This is, on the one hand, because on requiring more input data from the subjects on what to say, possible problems arise related to data quality and transcription and, with it, imprecision in prediction (Rudin 2019). On the other hand, the supposed benefits in precision are also more vulnerable to overfitting, which occurs when a predictive model generalizes incorrectly from patterns of correlation that are irrelevant in the context of applying the model (James et al. 2021, Grant et al. 2023, Shellenbarger 2019). This is particularly worrying when these patterns of correlation result in discriminatory conclusions, either because the training data are biased, or because the correlations are genuine but morally inadmissible (such as the case of racial discrimination arising from prisoner parole programs). Being so effective at finding and processing correlations, these systems lack human capacity to perceive possible discrimination resulting from statistical data on the group to which the subject belongs. Thus, even when the data confirm some statistical relationship between race and job efficiency or criminality, the human, in contrast to the system, would always be able to perceive the injustice of considering this evidence when hiring or approving parole to certain individuals. The black box system would not detect these immoral inferences from the data and, even worse, its opacity would not allow them to be perceived by humans (Grant et al. 2023).

This all points, therefore, to a conditional acceptance of the belief that the algorithms are permitted to make relevant decisions for humans because they are able to guarantee a greater autonomy of all affected parties. With technological advances, we may even eventually develop a method for determining in advance whether certain decisions should be left more to the system than to humans. This method would consider aspects such as the potential greater efficiency of the opaque system (given the volume and complexity of the information to be processed), its fairness (being free of algorithmic and human bias) and its contribution to increased autonomy (giving people more opportunities). Thanks to this method, even before decisions are made, depending on the particularities of the issue at hand, it would be possible to know whether or not it is appropriate for partially opaque systems to decide before humans. Thus, in general, considering the aspects mentioned above, we could be more

permissive with their use in medical fields than in purely administrative ones. But then the particularities of each situation could require us to make a more specific judgement. Thus, there may be cases of medical procedures that, because they do not require the processing of so much data, because they involve less serious illnesses or because they require a closer doctor-patient relationship, may make it advisable for a human being to make the decisions. Similarly, in the administrative sphere, a distinction could be made between decisions that significantly affect people's opportunities and those that do not. Even so, it should not be forgotten that these would be guidelines for a future method that would be free from the technological limitations mentioned above, such as those inefficiencies and possible algorithmic biases that could conceal the opacity of decisions.

3. ***Limited Explanation Principle.*** From the alternative proposal we have defended here, the right to explanation would not be absolute. Not to understand it so and demand full transparency would contrast with the limited demand for explanation that is actually required both in the uses and contexts of OADS and with the degree of transparency demanded when human decision makers are used for the same purposes. Moreover, as we have seen, to demand such transparency in an absolute way, from idealistic conceptions of human autonomy, may lead to a counter-productive result as regards the intended extreme defense of respect for individuals. To all of the foregoing, I have argued, from the proposed CAA that, in order to improve the level of general autonomy, it would be appropriate to allow a certain degree of opacity in those one-off situations in which opting for OADS may offset the immediate lack of explainability therein with a considerable increase in overall autonomy. To delimit such ad hoc permissibility of OADS, they should, from a clearly consequentialist perspective, be evaluated by virtue of their efficiency, fairness and reliability.

Thus, in regard to a possible increase in efficiency for greater autonomy at the expense of certain opacity, examples could be given of situations in which indicators for a good decision that are considered trustworthy stop being so when made public. These would be indicators of certain goods that are no longer considered as such if everyone begins to pursue it from the knowledge that it adds up to the assessment required for the decision. If the only reason for not enrolling on a certain training course outside that of improving a curriculum and therefore have more options of getting a job, but it turns out that the course in question only has an effect when followed with a strict self-improvement motivation (for example a debate course, according to Vaassen 2022), such an indicator on the evaluation scale would have lost its value. For this reason, being aware of this, and after confirming the possible greater goods that would be produced in terms of autonomy (with the best performance in the job on the part of the chosen citizen, for example), in certain cases not making certain explanatory factors of the decision public could be allowed.

Additionally, a system could also lose efficiency to the extent that transparency makes it obligatory for reasons to be understandable for non-experts and this has an effect on an intended restriction of the functioning thereof. Nguyen (2022: 339-350) indicates how the demand for public transparency in evaluative processes of social activities may put pressure on experts and become a missed opportunity for society to benefit from the best evaluative judgments they offer. This author gives the example of how, fearing incomprehension or suspicion on the part of the public regarding the evaluation that a philosopher could perform as an expert on the teaching of critical thinking in philosophy departments (their main purpose), they end up

performing an evaluation based solely on indices, more understandable for the public, such as the ratio of graduates or their success in the labor market. This would finally mean that the evaluator would not evaluate as an expert and that society would not benefit from this (338). Applying this idea to the sphere of algorithmic decisions, we can draw conclusions on the convenience of sometimes not demanding so much transparency. With experts in computation being aware that society will seek public explanations of decisions made by the system they have created, in order to avoid nepotistic, self-serving or biased motivations being hidden in the process, it is possible that they will not feel motivated to create more flexible mechanisms that permit the discovery of useful patterns, which are not easily observable and understandable by humans (Vaassen 2022). As such, the requirement for excessive transparency could, in some cases, prevent algorithmic systems from achieving better outcomes for all.

A second reason, apart from efficiency, for justifying a certain lack of transparency could be the resulting increase in equity that would be achieved. It may be the case that providing relevant causal explanations (in order for the individual to consider their evaluations and actions) makes access to certain assets easier. Let us think about situations in which the information provided on evaluative indicators could be better used in their favor by power groups or more affluent classes, either by acquiring merits more easily attainable by them, or exerting social or political pressure to modify the indicators at their convenience. It is however also true that, in contrast, the application of transparency measures would permit audits that examine the nature of data, factors and processes that could have caused discrimination of some kind in the decisions, There would thus be a need for a targeted exam that specified the convenience or not of transparency by virtue of its possible consequences (Zarsky 2016: 125).

Thirdly, I think that trust in the system could also be an important factor for demanding greater or lesser explainability from the system. It is true that, in principle, more transparency means greater trust. It is obvious that our mistrust of certain decisions, recommendations or predictions increases considerably when not accompanied by a plausible explanation (Symeonidis et al. 2009, Ribeiro et al. 2016, Doshi-Velez & Kortz 2017, Lipton 2018, Holzinger et al. 2020). Therefore, if algorithmic systems lack transparency in their resolution processes, it is normal that doubts and worries arise, either regarding the possible concealment within these processes of functional bias or bias deriving from the data (Kim et al. 2016, Biran & McKeown 2017, Buolamwini and Gebru 2018; Barabas et al. 2020, Vredenburgh 2022) or in relation to the very efficiency of the system, by denying technicians relevant information about its functioning that would allow modification when their effects were not those desired (Domingos 2015; Selbst & Barocas 2018, Ribeiro et al. 2016; Lyons et al. 2016, Coloner 2022: 232, Walmsley 2020, Vredenburgh 2022).

But this is not always the case. To begin with we must be cautious in this regard, as human trust is a complex concept and therefore difficult to measure and manage. Thus, trust is normally connected to biased heuristic processes and prejudices linked to factors such as pre-existing beliefs on the institution making the decision or peculiarities of the setting in which it is made. AI is not exempt from such prejudices, and they become highly evident when there is debate on its limitations and potential. This leads to strange reactions, such as the study by Ehsan & Riedl (2021), according to which trust in an AI system was greater when the explanation of how it worked was performed using much more difficult to understand mathematical and computational language. In this vein of strange conclusions on the relationship

between transparency and trust in the system, some authors have also observed that in order to avoid mistrust it would be more effective to provide information on what institutional or particular purposes the decision-making system actually serves, rather than specific explanations about how the system has made decisions (Selbst & Barocas 2018: 1130).

To summarize, from the CAA we present here, as a general rule human beings will be more autonomous if they receive explanations of the decisions that are made on aspects relevant to their values and lives. As a result, the use of OADS should, in principle, be avoided. Nevertheless, the CAA differs from the DAA in that it could justify the fact that certain explanations are not given as long as the ends adequately justify this. By increasing the autonomy of all, and providing the efficiency, equity or reliability of the system is not undermined, certain decisions could be opaque. Other relevant factors for determining when this exception to the rule could be given would be the degree of complexity of the decision and its relevance in the plans to be implemented by the subject. With all of these factors in mind, the ethical thing to do would be to study, in each particular case, whether or not it is convenient to use an OADS.

References

Anderson, M. & Anderson, S. L. (ed.) (2011). *Machine Ethics*. Cambridge: Cambridge University Press.

Atkins, K. (2000). Autonomy and the subjective character of experience. *Journal of Applied Philosophy*, 17(1), 71-79.

Barabas, C., Doyle C., Rubinovitz, J. B. & Kinakar, K. (2020). Studying Up: Reorienting the study of algorithmic fairness around issues of power. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 167-176.

Behdadi, D. & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, 30 (2), 195-218.

Bentham, J. (1789). *Introduction to the Principles of Morals and Legislation*. University of London.

Biran, O. & McKeown, K. R. (2017). Human-centric justification of machine learning predictions. *IJCAI*, vol. 2017,1461-1467.

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

Bratman, M. (2018). *Planning, Time, and Self-Governance: Essays in Practical Rationality*. Oxford University Press.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199-231.

Goodman, B. & Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, *38*(3).

Buolamwini J. & Gebru, T. (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research,* 81,1-15.

Campbell, J. (2008). Interventionism, control variables and causation in the qualitative world. *Philosophical Issues*, 18(1), 426-445.

Cappelen, H. & Dever, J. (2021). *Making AI Intelligible: Philosophical Foundations*. Oxford University Press.

Caruana, R., et al. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*.

Coeckelbergh, M. (2012). *Growing Moral Relations: Critique of Moral Status Ascription*. Palgrave Macmillan.

Colaner, N. (2022). Is explainable artificial intelligence intrinsically valuable? *AI & Society*, 37, 231-238.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology,* 29, (3), 245-268.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake the World*. Perseus books Group.

Doshi-Velez, F. & Kortz, M. (2017). *Accountability of AI Under the Law: The Role of Explanation*. Berkman Center Research Publication.

Dombrowski, D. A. (1997). *Babies and Beasts*. University of Illinois Press.

Ehsan, U. & Riedl, M. O. (2021). Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv preprint arXiv:2109.12480*.

Floridi, L. & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14 (3), 349-379.

Fried, C. (1978). *Right and Wrong*. Harvard University Press.

Grant, D. G., Behrends, J. & Basl, J. (2023). What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. *Philosophical Studies*. https://doi.org/10.1007/s11098-023-02013-6

Gunkel, D. (2012). *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. MIT Press.

Hare, R. M. (1952). *The Language of Morals*. Oxford University Press.

Hare, R. M. (1963). *Freedom and Reason*. Oxford University Press.

Ismael, J. (2016). *How Physics Makes Us Free*. Oxford University Press.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to Statistical Learning with Applications in R*. Springer.

Jolls, C., Sunstein, C. R. & Thaler, R. (1998). A behavioral approach to law and economics. *Stanford Law Review*, 50 (4), 1471-1550.

Kaminski, M. E. (2019). The Right to Explanation, Explained. *Berkeley Technology Law Journal,* 34.

Kim, B., Khanna, R. & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *NIPS,* 29, 2280–2288.

Klein, G. A. (2017). *Sources of Power: How People Make Decisions*. MIT press.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, 6(8), 539-551.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, 49(1), 15-21.

Lyons J., Sadler, G., Koltai, K., Battiste, H., Ho, N., Hoffmann, L., Smith, D., Johnson, W. & Shively, R. (2016). Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems*, July 27–31.

McKinney, et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89-94.

Meadow, W. & Sunstein, C. R. (2001). Statistic, not experts. *Duke Law Journal,* 51 (2), 629-646.

Moniz Pereira, L. & Barata Lopes, A. (2020). *Machine Ethics. From Machine Morals to the Machinery of Morality*. Springer.

Moore, G. E. (1903). *Principia Ethica*. Cambridge University Press.

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

Newell, A. (1982). The knowledge level. *Artificial Intelligence,* 18 (1), 81-132.

Nguyen, C. T. (2022). Transparency is surveillance. *Philosophy and Phenomenological Research*, 105(2), 331-361.

Pluhar, E. (1995). *Beyond Prejudice*. Duke University Press.

Purves, D., Jenkins, R. & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, 18 (4), 851-872.

Raz, J. (1986). *The Morality of Freedom*. Oxford University Press.

Ribeiro, M. T., Singh, S. & Guestrin, C. E. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1135–1144.

Rubel, A., Castro, C., & Pham, A. (2021). *Algorithms and Autonomy. The Ethics of Decisions Systems*. Cambridge University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Schwitzgebel, E. (2019). Introspection. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition).

Shellenbarger, S. (2019). Make your job application robot-proof. *The Wall Street Journal*. https://www.wsj.com/articles/make-your-job-application-robot-proof-11576492201

Selbst, A., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, 87(3), 1085-1139.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4).

Singer, P. (1975). *Animal Liberation*. The New York Review of Books.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science* Part A, 44(3), 510-515.

Sullins, J. P. (2011). When is a robot a moral agent? In Anderson, M. & Anderson, S. L. (Eds.), *Machine Ethics* (151-161). Cambridge University Press.

Susser, D., Roessler, B. & Nissenbaum, H. (2019). Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review*, 4 (2019), 1-45.

Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2009). Moviexplain: a recommender system with explanations. *Proceedings of the third ACM conference on Recommender systems*, 317-320.

Vaassen, B. (2022). AI, Opacity, and Personal Autonomy. *Philosophy and Technology*, 5, 88.

Verbeek, P. P. (2005). *What Things Do? Philosophical Reflections on Technology, Agency, and Design*. Pennsylvania State University Press.

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy*, 30(2), 209-229.

Walmsley, J. (2020). Artificial intelligence and the value of transparency. *AI and Society*, 1–11.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76-99.

Yeung, K. (2017). 'Hypernudge': Big Data as a Mode of Regulation by Design. *Information, Communication & Society* 20 (1), 118-136.

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology, & Human Values,* 41(1), 118-132.

Zerilli, J., Knott, A., Maclaurin, J. & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology*, 32(4), 661-683.