Human intervention in automated decision-making: Toward the construction of contestable systems

Accepted manuscript, ICAIL 2019

Marco Almada
University of São Paulo & Lawgorithm
marco.almada@usp.br

ABSTRACT

Concerns about "black box" machine learning algorithms have influenced why modern data protection laws and regulations on their establishment of a right to human intervention on decision-making supported by artificial intelligence. Such interventions provide data subjects with means to protect their rights, freedoms, and legitimate interests, either as a bare minimum requirement for data processing or as a central norm governing decision-aiding artificial intelligence. In this paper, I present *contestability by design* as an approach to two kinds of issues with current legal implementations of the right to human intervention. The first kind is the uncertainty about what kind of decision should be covered by this right: should intervention be restricted to those decisions with no human involvement, or should it be interpreted in a broader sense, encompassing all decisions that are effectively shaped by automated processing? The second class of issues ensues from practical limitations of this right to intervention: even within a clear conceptual framework, data subjects might still lack the information they need to the concrete exercise of their right, or the human intervention itself might introduce biases and limitations that result in undesirable outcomes. After discussing how those effects can be identified and measured, I then advance the thesis that proper protection of the rights of data subjects is feasible only if there are means for contesting decisions based solely on automated processing is not an afterthought, but instead a requirement at each stage of an artificial intelligence system's lifecycle.

KEYWORDS

Automated decision-making, algorithmic bias, machine learning regulation, contestability by design, privacy by design

1 INTRODUCTION

Modern data protection laws have introduced a right to human intervention on decisions based on automated data processing as a measure to protect data subjects from harms or undue constraints to their rights or interests. The current paradigm of such a right can be found in Article 22(3) of the European Union's *General Data Processing Regulation* (GDPR), which establishes that data subjects have the right to contest decisions based solely on automated processing. As a result, those subjects are provided with the means to request a reevaluation of any decisions that introduces undue constraints to their legitimate interests, liberties, and rights. However, those means only have a concrete impact if data subjects actually contest a specific decision, something that is only possible if the contesting subject can find out whether a decision that affects their interests or rights involves automated processing.

Identifying the role played by automated processing within a decision-making process can be difficult in practice, due to the many kinds of opacity [7] involved in decisions supported or made by artificial intelligence. As a reaction to this issue, the discussion on the existence and reach of a "right to explanation" of the decisions based on automated processing[1] ends up taking a central place in the debates regarding the right to human intervention, while questions about the adequate means for human intervention, its effectiveness and the rights and interests that it should protect do not receive the same attention within the literature.[2]

In this paper, I approach two issues about the reach of a right to human intervention. First, I explore what kinds of decisions are covered by such a right. The existing GDPR norm allows human intervention only for decisions *solely* based on automated processing of *personal* data, but this formulation leads to some questions: when is a decision solely, as opposed to only mostly, based on automated processing? Are there meaningful differences between the notions of "automated decision-making" and "decisions (...) based on automated processing"? I sustain that "automated decision-making" should be understood as a shorthand for the sort of decision that is covered by the right to intervention, rather than a full description, as there a decision can be based solely on automated processing even in cases when there is a (sufficiently constrained) meaningful human participation in the decision loop.

The second issue concerns the nature of the protection offered by a right to human intervention in automated decision-making. At least in part, human intervention is meant to solve decision problems that cannot be addressed within current computational capabilities [2], but it has also been justified in terms of "uphold[ing] human dignity" [23]. However, what happens when intelligent systems can make decisions that lead to fairer outcomes than a human decider or intervenor could achieve? This question is relevant not

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

[1] This debate [28, 4, 23] is part of a more extensive discussion focused on the GDPR. Despite the jurisdictional focus, the normative concepts discussed within this debate are relevant even outside the European Union, both because of the global reach of the GDPR and because of the influence it has had in lawmaking in other jurisdictions such as the Brazilian LGPD.

[2] That is not to say that there is no discussion of such questions, as Edwards and Veale [12] propose a shift away from a "right to explanation" and towards ensuring decision quality.

just from a design standpoint — as technical improvements in machine learning solutions may lead to superhuman capabilities in some applications — but also because of the biases present in human decision-making due to individual, organisational or societal factors.

To explore those questions, I draw from the GDPR to identify how the questions raised above may arise not just as incidental features of normative design but as conceptual issues with the idea of human intervention as currently understood. In Section 2, I explore the notion of automated data processing, discussing when this kind of intervention is applicable and what goals it should achieve. Section 3 then addresses some practical issues regarding human intervention: how much information is necessary for the exercise of that right? Can algorithmic biases be contained? Can human intervention lead to more biased outcomes than the ones produced by artificial intelligence systems? I then finish by discussing, in Section 4, some practical measures to safeguard rights and legitimate interests from the potential harms from automated decision-making. In particular, *ex ante* intervention during the AI design process may prove to be a more effective tool than *ex post* review of potentially harmful decisions.

2 THE RIGHT TO HUMAN INTERVENTION

The right to human intervention is usually[3] proposed as a means to contest decisions that rely on data processed through automatic means. The basis for such a right may vary between jurisdictions, but in the formulation adopted by the GDPR a data subject may contest a decision if it affects *rights* or *interests* legitimately held by them and if it is *solely* based on the *automated* treatment of *personal* data regarding the data subject. Understanding which decisions are subject to human intervention therefore demands further investigation of those shared elements.

What sort of interests count as legitimate for the purposes of this right? Current implementations of the right to intervention usually do not restrict themselves to legal interests. Instead, a significant impact such as automatic rejection of an online credit application provides sufficient grounds for contesting the relevant decision [9, 19]. As Edwards and Veale [12] point out, the GDPR uses a vague notion of "interest",which can be specified, at least in theory, through the use of soft law instruments such as the GDPR Recitals. However, the discussion on the kind of interest that should be taken into account is more related to general issues regarding the legal system to which a given data protection law belongs[4] rather than machine learning-specific issues.[5]

The notion of *automated processing* can also be a source of vagueness in both laws, as it appears to cover a wide range of computer operations.[6] Brkan [4] uses the expression "automated decision-making" to cover all decisions that are solely based on automated processing: as long as the ultimate decision rationale is provided through automatic means, this definition treats a decision as automatic even if it involves human intervention at other stages of the loop, such as data acquisition or interpretation of the actual decision made. In particular, a scenario in which a human controller merely rubber-stamps the decisions made by an automated system still counts as an automated decision [33], as the decision maker cannot or will not alter the actual result.

While "automated decision-making" is a useful shorthand for the sort of decision solely based on automated processing that is the object of the right to intervention, one must keep in mind that some of the relevant decisions do not fully remove humans from the loop. While this might be the case for many decision-making systems, such as high-frequency trading operations, it is possible, at least in principle, to build systems in which a human decision-maker can rely only on data obtained through automated sources. If this data does not require much interpretation, we are once again back to the rubber-stamping scenario, as the human "decision-maker" will play a merely *formal* role that is not far removed from what a random number generator could do.

If, on the other hand, the decision-maker has to perform a deliberate choice between the automatically-produced options, the decision cannot be described as fully automated, as the human controller will necessarily deploy some of their previous knowledge and values when making a non-trivial decision based on the automatically processed data.[7] Under a strict interpretation of what counts as the basis for a decision, such as the one proposed by Brkan [4], attributing any substantial role to a non-automated controller would be enough to avoid the need for human review, even if the outputs of automated processing still shape the actual human decision. That, in turn, would make it possible to circumvent the right to human intervention through organisational designs more sophisticated than simple rubber-stamping.

For example, if a human can choose freely between several possible scenarios, but all of those choices are generated by automated tools, their decision space is severely constrained. Nonetheless, the output decision would not be performed by an automated system, as the human actor can choose between one of several outcomes that might be very different from one another. Since, in this case, human choice can lead to *substantially* different outcomes, a narrow interpretation of GDPR Article 22(1) would then exclude this sort of decision from the scope of the right to intervention, despite the fact that the human decision-maker would still lack control over the *content* of the outcome, as the options were generated from what Mendoza and Bygrave [23] call "data shadows": model representations of the data subjects, which treat some aspects of a person that

[3] See, e. g., GDPR Article 22(3).

[4] Regarding the problem of aggregate decisions, Edwards and Veale [12] claim that "Such group privacy impacts are not dealt with well by [data protection] law—an area based on individualistic human rights—and are exacerbated by a continuing lack of provision for class actions in EU states".

[5] Part of the debate on whether robots should have some sort of legal personality involves the question of whether a robot can be said to have its own interests that should be legally protected [25, 6]. If that thesis becomes accepted, the notion of a legally protected interest would be enlarged, but current technology is still far from the point where it could be plausibly claimed that robots hold autonomous interests.

[6] As Brkan [4] points out, "The notion of automated decision-making is not a unitary concept, comprising only a particular type of decisions. Rather, it is broad, multifaceted and prone to be divided into several sub-categories", covering applications as diverse as the sorting of search results, online advertisement, decisions about bank loans, and high-frequency trading.

[7] This might be the case even if, usually, the system itself provides guidelines for deciding between those scenarios, such as accuracy metrics. If the suggestion is regularly accepted as-is, this becomes an example of rubber-stamped decision, but it is possible, at last in principle, that human deciders simply take those metrics as parameters to decide between (potentially incommensurable) expected results of possible choices.

can be metrified, but might fail[8] to capture significant aspects of human experience. This failure can, in turn, lead to unintentional harm or insufficient protection to rights and legitimate interests that were not adequately captured by the models used to represent the data subjects of automated processing.

A possible safeguard against such cases would be to interpret the "based solely on…" requirement (GDPR Article 22(1)) in a broader sense: as long as the only *outside* sources of information available to the decision-maker come from automated processing, the ensuing decision should be considered as solely based on that processing and therefore contestable. This interpretation does not overextend the scope of a right to human intervention, as the introduction of any substantial information sources that do not require automated data processing will still put a decision outside the reach of this sort of review. Even so, it prevents the instrumental use of organisational arrangements and the underdeterminacy of data to hinder data subjects from contesting decisions that fail to properly consider their rights and legitimate interests.

3 PUTTING A HUMAN IN THE LOOP

Vague definitions notwithstanding, the idea of a right to human intervention has been introduced into legislation as a reply to various demands and concerns about the roles that automated processing of data should play in modern societies. Human review is seen, for example, as an antidote to machine error: tacit human knowledge [26] and intuitions, which can be challenging to represent computationally, could help in the identification of mistakes committed by machines. From an instrumental perspective, then, human intervention is demanded as quality control, especially since failures in automated systems can lead to large-scale harms.[9]

Any defence of human intervention along those lines, however, is highly circumstantial: as Brennan-Marquez and Henderson [2] suggest, advancements in artificial intelligence technologies might lead to systems which can accurately codify human intuitions or even perform better, for any given criteria, than a reasonable human baseline.[10] In fact, it is not impossible to see a future in which machine learning algorithms could "be made to disregard discriminatory factors more effectively than humans." [19], as the nascent fields of algorithmic fairness and transparency build solutions and standards for non-discriminatory use of machine learning systems. If and when those goals are achieved, human intervention could make a system *more* likely to behave inadequately, resulting then in avoidable risk to rights and legitimate of the decision subjects.

A right to intervention, however, can be sustained on grounds other than the limits to computational accuracy and efficiency. Hildebrandt [15], drawing from the literature on "the foundational indeterminacy of human identity" [15] and also on the technical limits posed by the theory of computation, claims that there are aspects of each person that are incomputable, that is, mathematically impossible to describe by the kind of models that can be implemented within a computer. In that context, Hildebrandt [15] claims that data protection laws should ensure that persons are not reduced to the computable facts about them, a point we further explore in Section 4.

In another direction, Brennan-Marquez and Henderson [2] claim that democratic decision-making requires that those in charge of decisions should also be susceptible, at least in principle, to the effects of its decisions. Therefore, as long as machines are not able to "internalize the effects of judgment" [2], their decisions within a democratic polity — for example, those taken by a robotic juror [2] — should be subject to human oversight, such as the possibility of human intervention.

Both Hildebrandt [15] and Brennan-Marquez and Henderson [2] are, thus, concerned with "cyberphysical systems that reconfigure both us and our world" [15] without any legal means for contestation. Those concerns with the transformative impact of artificial intelligence do not necessarily dissipate with technological advancements; in fact, the development of new technical solutions will probably lead both to new means for intelligent systems to effect change in the world and to new approaches for human oversight of automated decisions, a development that could lead to an arms race between regulatory approaches and the design of artificial intelligence systems.

Still, a well-founded system for human intervention in automated decisions might prove itself ineffective if it is difficult to use. Under a broad interpretation of "intervention", as discussed by Kamarinou, Millard, and Singh [19], human action might be required not just as a means to replace the final decision made through automated processing, as humans are also involved — and therefore could act — in the earlier stages of system design, training, and testing.[11] It also might be difficult to access in practice the means that are formally available for requesting human intervention: "For example, if the data subject concludes an online contract with dynamic pricing, how can she request human intervention if the website does not provide for that possibility?" [4].

Proper application of data protection laws to decisions based on machine learning approaches will require answers to those and other questions related to the foundations and the extent of such right. In this section, I explore two of the relevant aspects: whether a "right to explanation" is necessary for human intervention and how to understand the effects of the right to intervention. First, I briefly consider what sort of information is necessary for the proper exercise of a right to human intervention, before discussing existing proposals for the measurement of algorithmic discrimination. Such measures can be relevant as a means to identify situations in which intervention is necessary, but they can also work as a control for the intervention process itself: since human intervenors might themselves introduce biases and errors to the decision-making

[8] Either as a result of inadequate modeling or, as Hildebrandt [15] points out, a deeper reflection of computational limits to what aspects of a human being can be adequately represented.

[9] A drastic example, described by Brennan-Marquez and Henderson [2], is the Petrov incident, in which the Soviet officer Stanislav Petrov decided to override an automated system built for detecting nuclear attacks against the Soviet Union. Instead of acting in accordance with the notifications he was receiving, he decided to treat them as a system failure, based on a gut feeling that was later proven correct and has prevented a nuclear war.

[10] As of the writing of this paper, so-called Artificial General Intelligence is still a long-term goal, but AI systems have already achieved super-human capacity in a broad variety of domains, such as chess and video games, or, to pick a more blatant example, arithmetic [32].

[11] This broader perspective is compatible with the claim made by Lehr and Ohm [22] that legal scholars should pay more attention to the software development stages that antecede the actual model deployment, especially considering that such stages do not flow linearly. Instead, design, training, and testing will usually continue to happen even after a machine learning model is put into production.

process, their performance should also be subject to external control. I thus finish the section by proposing that the fairness of automated decision-making should also be evaluated in comparison with the expected outcomes from replacing the automated system with a human.

3.1 The right to meaningful information regarding automated decisions

Current implementations of the right to human intervention require that data subjects take action against perceived undue constraints, such as actively contesting the decision.[12] The formal requirements for exercising this right depend on each particular jurisdiction, but in all cases, a data subject will need some information about the systems involved in the decision-making process, if only to identify whether there are grounds for intervention.

Acknowledging this demand, GDPR Articles 13–15 establish that a data subject must have access to information about the existence of any automated decision-making based on their personal data and, if such processing exists, a "right to meaningful information" [28] about how this processing happens. In the European Union, this right to meaningful information has often been interpreted in a broad sense. Selbst and Powles [28] claim that information meaningfulness should be evaluated on a "functional" basis: a piece of information regarding a decision based on the automated processing of personal data is legally relevant if a data subject would need that information to exercise their rights. Since, as discussed in Section 2, decisions based solely on automated data processing usually still happen within an intricate organisational context, compliance to even a narrow functional interpretation would require the disclosure of information concerning not only the technical aspects of how the relevant intelligent systems generate output, but also the data on which the automated processing happens and the human and institutional factors involved, as well as on how a data subject should proceed to effectively request human intervention.

To a certain extent, explainable artificial intelligence (XAI) approaches can provide information about the algorithmic aspects of decision-making. Recent work in XAI, as described by Mittelstadt, Russell, and Wachter [24], has been focused on one of two goals: transparency — that is, the representation of the inner workings of a model, or a part of it, in ways that can be understood by a human being — or *post hoc* interpretations of model behaviour. However, most of the current XAI approaches work by producing local approximations of the models that one wishes to explain, rather than the sort of explanation that is usual in human communication.[13]

To a certain extent, technological development could make those local approximations more understandable to the layperson, mitigating the sort of opacity that arises from technological illiteracy. Yet, transparency-based systems might face a more fundamental constraint, described by Burrell [7] as "mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation.". Admitting that explanation techniques can present an algorithm in the simplest possible form, that complexity reduction might not be enough to make the system understandable for private individuals or small organisations. If even a simplified model still proves to be a black box, the algorithmic explanation will neither be necessary for providing help against the more blatant violations of rights, freedoms, and legitimate interests — which will have directly noticeable effects — nor sufficient to protect those data subjects against more subtle harms or frustrations.

Against such highly complex systems, it might be more useful to use techniques that control system behaviour, such as those proposed by Kroll *et al.* [21]. In those approaches, the explanation of a given algorithmic decision does not show the inner workings of a system. Instead, it presents, in a way that can be understood by humans, the factors that led the system to take a given decision and the outcomes of the choice that was made, switching the emphasis to the control of the *decision rules* rather than the computation rules. Data subjects would then be able to receive reports about whether, given input data and relevant decisions, a system has followed the accepted rules. If the answer is "no", the affected data subject would have grounds for requesting human intervention, even if they do not know the specific details of how the algorithm worked.[14] Therefore, behavioral explanations of algorithmic models can, if properly coordinated with a more general explanation of the organisational context in which the decisions are made,[15] empower data subjects to identify whether intervention is necessary without having to understand (a simplified version of) the workings of a given decision model.

3.2 The impact of human intervention

Up to this point, we have discussed human intervention as a tool to safeguard rights that have been imperilled by automated decisions. However, does that tool provide an effective approach to the problem it claims to solve? Decisions based on automated processing, such as the profiling of large customer bases, usually affect a large number of individuals. So, as long as an appropriate framework for intervention is in place, individual requests only become a significant overhead to a company's operations if a substantial number

[12] GDPR Article 22(3) establishes that data controllers shall implement measures to ensure that the data subject "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision", but it still falls upon data subjects to actually make use of such rights.

[13] Mittelstadt, Russell, and Wachter [24] describe explanation as a mostly causal phenomenon, based on the contrast between the observed scenario and some (but not all) of the plausible alternatives. In spite of that, social phenomena are many times described in terms that are not necessarily reducible to causal narratives; legal norms are usually understood as providing persons (natural or otherwise) with reasons for acting in this or that way, and a description of those reasons is also seen as form of explanation, for instance, in a trial. Algorithmic explanation would, thus, benefit from a greater consideration of non-causal forms of explanation.

[14] Reducing the amount of information that a data subject must have to exercise or not their right to intervention is desirable not only for a corporation — that can preserve its trade secrets — but from a civic standpoint, as it lowers the barriers that might prevent less-knowledgeable citizens from seeking redress to harms. Techniques such as zero-knowledge proofs can, in principle, be used as a starting point for ensuring to data subjects that their rights, freedoms, and legitimate interests are preserved by a given algorithm, but their acceptance by the intended targets — especially the less math-savvy ones who have the most to gain from this sort of approach — requires a certain level of trust, which can be established by means such as external certification of the explanation methods (as suggested by GDPR Article 25(3)).

[15] Since those human and institutional decisions have quantifiable aspects, it might be interesting to also subject them to the sort of black box controls provided by *post hoc* explanation.

of individuals decide to exercise their rights,[16] even if manual review turns out to be several orders of magnitude slower or more expensive than the original processing approach. At first glance, then, human review seems to be a solution that ensures individual rights without undue intervention in general business practices.

This assumption that human intervention is a low-cost approach might fail if the requirements for an intervention place significant burdens upon data controllers or processors. Even if it holds, the ensuing human actions might be ineffective for protecting the rights of the claimant or, worse, actually leave them at a worse position than the one that would result from the previous automated processing,[17] if they somehow lead to outcomes that unduly affect the rights, liberties, or legitimate interests of the claimant. Therefore, the design of actual mechanisms for human intervention on decisions based solely on automated processing might benefit from adequate measures of their effects.

One particular category of harm that is usually considered as a critical motivator for human intervention is *algorithmic discrimination*, which happens when algorithmic decision-making treats individuals or groups in a different way based solely on a salient attribute or set of attributes, such as race or gender [14]. These outcomes might be a result of deliberate design choices — as an extreme example, a government might use data-driven approaches to enforce segregationist policies — or appear as an undesired result of the algorithms used within a system, the data sets used in its training or the training process itself.[18]

Many forms of discrimination — whether or not performed by algorithms — can lead to measurable effects, such as the financial cost of a denied loan or the additional time that an individual might spend commuting after being algorithmically priced out of their neighbourhood. In those cases, mathematical models of taste-based discrimination and statistic discrimination[19] can be used to monitor whether a given algorithm is producing unfair outcomes against an individual or group [14].

Still, not all aspects of algorithm discrimination lend themselves to straightforward quantification. Taking that into account, Zarsky [34] proposes an analytical framework that considers two dimensions of algorithmic decision-making: both the new issues that the decision-making process may introduce and the existing problems that automation may worsen. Within contexts where algorithms produce opaque decisions, those dimensions may be evaluated in terms of their efficiency — in an approach similar to economic treatments of the subject [14] — and their fairness.

Evaluating fairness within a computational context can be a difficult task. One possible approach for identifying adequate measures of fair decision-making is to capture how people perceive algorithmic decisions and what aspects they evaluate before saying that a decision is fair or unfair [1]; with such standards in hand, it becomes possible to build artificial systems that attempt to meet them.One key qualitative aspect that has emerged from preliminary studies [1, 36] is that "people do consider justice-related aspects of algorithmic decision-making systems, much as they do for manual decision-making processes" [1], which results in high levels of support for the careful management and governance of artificial intelligence technologies [36]. While those studies are not yet mature enough to provide a basis for wide-sweeping normative conclusions,[20] their results can be useful as a proof of concept for modelling fairness-related aspects of algorithm evaluation and as a basis for future experiments.

One direction that seems to require further enquiry is the fairness of large-scale automated decision-making. Some systems, such as the ones deployed in high-frequency trading, can make individual decisions that result in millions of dollars in profits or losses; others, like the choice of what ads a website shall display to a user, can be individually small [23], but their aggregate impact might be significant. In the latter case, any specific intervention will not produce many effects beyond the potential preservation of individual rights that would otherwise be unduly constrained. On the other hand, replacing artificial intelligence with a human being in the first case will place the human reviewer in a position where their decision can have a substantial impact and must, accordingly, be subject to stricter standards.

Monitoring the impact of human decisions makes sense if the human reviewer can make a difference in the decisory outcome. Sometimes, as previously discussed, a formally autonomous decision-maker might be bound by automated data analysis. Here, a narrow interpretation that restricts intervention to the end stages would make it useless, but human intervention in the design stages may be more effective by proposing alternative models of the data that take such concerns into account. There is also the possibility that human intervention actually leads to *worse*[21] results than the ones produced by automated processing,[22] a result that might ensue from actively discriminatory measures taken by the intervenor or from unintentional causes, such as the biases present in human

[16] This is possible, for example, if individual defects restrict or harm the rights or interests of data subjects in individual ways, or by concerted action, either from civil society entities or through the tools provided for class actions within a specific jurisdiction. As Edwards and Veale [12] affirm, current data protection laws usually lack specific measures for the defence of collective and diffuse rights, a situation that risks leaving demands prompted by algorithmic generalisations over aggregate data without proper legal remedy. Thus, the problem of how (potentially new) regulations can protect transindividual rights from algorithmic abuse seems to be an appropriate direction for future research.

[17] If harm or undue restrictions ensue from human intervention, the data subject might have grounds for legal remedies under existing, non-data-related, norms, such as those from tort law. Still, it might be possible to address such potential undesirable results under a *preventive* approach or at least provide data subjects with the means to learn whenever they are harmed in this way.

[18] Lehr and Ohm [22] provide a useful typology of the stages involved in the construction and use of machine learning models.

[19] Roughly speaking, taste-based discrimination considers that discrimination against groups is a result of individual preferences and aversions. In contrast, statistical discrimination refers to the idea that, under limited information scenarios, people and firms might use accessible traits, such as race or perceived gender, to extrapolate information that is currently unavailable [14].

[20] Either because of their non-representative sampling [1] or because of their geographical specificity [36].

[21] For any perspective that is relevant for the data subject, such as fairness or efficiency.

[22] As previously discussed, this is an outcome that might be expected as machine learning algorithms obtain better performance. Yet, it can already be seen in practice: Kleinberg *et al.* [20], based on a policy simulation, claim that an algorithmic approach to deciding whether or not a defendant should await trial in jail can, among other effects, reduce the share of jailed African-American and Hispanic defendants in comparison to the decisions made by human judges. Such approaches, if adequately implemented and scaled, can go some way towards mitigating racial biases, which are present in existing judicial systems [30]. For a more extensive legal discussion of machine learning as a tool for addressing biases within the criminal law system, see Sunstein [31].

thought,[23] or, in some cases, the cognitive limits of the human mind [32].

As, in these cases, the actual decision will have been made by a human, the usual liability remedies present in general legislation might be useful as means to seek redress against harms or undue restrictions introduced by human review,[24] provided that institutional opaqueness, as discussed in Section 2, does not prevent the identification of which persons should be held legally responsible for the ensuing effects. Seeking those legal means, however, leads to the costs associated not only with the review process itself but also with those related to adjudication or other approaches to dispute resolution, costs that might be avoided or mitigated by adequate measures.[25]

One possible approach for containing the risk that might result from unwarranted or inadequate human intervention is to compare the intervention outcomes with the outcomes from the original automated decision. If in a particular case, human intervenors — or even analogous decisions taken entirely by humans — are more likely to act in biased ways or even discriminate against individuals or groups, then a data subject would probably not benefit from exercising their right to human intervention,[26] or, if the intervention has already happened, they would be better off with the original automated decision.[27]

Since the decision on requesting or not human intervention ultimately resides with the data subject, it follows from the right to meaningful information discussed in Subsection 3.1 that data controllers and processors should provide access to fairness evaluations in a way that data subjects can use as a basis for informed decisions. Therefore, AI users can benefit from approaches that use counterfactual analysis to compare algorithmic bias to alternatives, such as explicit human decisions [10]. Following a similar line, Kleinberg *et al.* [20] use econometric approaches, such as counterfactuals, to measure unobservable variables can end up introducing "noise" that leads to biased decision-making by humans, something that may not only allow data subjects to be aware of possible sources of intervenor bias, but also inform human intervenors, allowing them to preventively address biases.

By establishing standards for the comparison of human-made and machine-made decisions, those lines of study provide a starting point for establishing actual compliance requirements[28] for the institutional arrangements involved in human review, without falling into a double standard [35] that subjects human decisions to weaker levels of scrutiny than the ones applied to automated decision-making.[29] Avoiding this contradictory standard, in turn, ensure that reviewed decisions are at least as fair as the original algorithm-based decision,[30] to be at least as fair as the original algorithmic decision, lest they provide data subjects with grounds for withdrawing their consent or, if the decision already has produced effects, starting a lawsuit. With that sort of standard in place, human intervention can work as a tool to preserve human dignity and self-determination.[31]

4 SAFEGUARDS AGAINST AI-SUPPORTED DECISION-MAKING

A right to human intervention is usually part of a more general legal framework for protecting natural persons against harms ensuing from automated decision-making.[32]Considering this intent, it is reasonable to interpret the actual regulations that govern human intervention — and their implementation in subsidiary norms and private regulation schemes — from a *functional* perspective [28], that is, considering the ultimate goals that this legal institute was designed to achieve: provide data subjects with the means to contest automated decisions that unduly constrain their rights or legitimate interests.

But why should data subjects have a general right to contest decisions based on automated processing? For Mendoza and Bygrave [23], the GDPR answers, at least in part, to "a concern to uphold human dignity by ensuring that humans (and not their 'data shadows') maintain the primary role in 'constituting' themselves.".[33] A concern with self-determination could provide a basis for human

[23] Biases such as a propensity to maintaining the *status quo* or confirm previously held opinions are present in human beings as heuristics that occasionally misfire [32]. The "errors" committed by following those heuristics can be useful to ensure that an Artificial General Intelligence does not stray far away from human-level cognition [32], but they may help perpetuate systemic injustices already present in society or prejudices deeply held by a human reviewer of automated decisions. Nonetheless, the introduction of those biases only makes sense when they do not produce systematic discrimination against legally protected groups.

[24] Criminal jury decisions can provide an example of how human biases might affect a decision, not only by affecting the actual result but by compromising overall faith in the juridical system, a problem that becomes particularly salient when it results in legally relevant discrimination. As Sundquist [30] describes, it may be difficult to interpret in practice what is "a clear statement of racial bias" and whether this sort of bias is a "significant motivating factor" in a juror's decision. Even after detection, addressing this bias may prove to be a more expensive task than it would be to redesign an automated system.

[25] As an example, preventing the occurrence of actual harms is a way to avoid the discussions about the liability that would ensue from the avoided effect. Since the correct attribution of liability in cases involving intelligent systems will involve complications ensuing from the legal understanding of relevant characteristics of artificial intelligence but also from more general corporate liability concerns, *ex post* redress might be a particularly expensive and complex to a problem that might have been solved at a low cost before it produced any impact.

[26] That is not to say, of course, that they should simply accept the actual decision. Instead, other means of contesting, such as lawsuits, might be more fruitful than direct intervention in the decision-making process.

[27] And might even have grounds to seek reparation, for example, through civil liability, if the comparison can show that the human decider neglected their duties to address biases and systematic discrimination issues.

[28] The preventive evaluation of review standards can, in the long run, be treated as another aspect of the general data compliance policies adopted by companies, which means that a mix of private and public authorities might play a role here, as discussed by Roig [27].

[29] Such a double standard might be accepted or even desirable for other reasons, such as fostering the use of human labor, but adopting it would effectively say that a human-caused harm to a right or legitimate interest is more acceptable than an equivalent harm caused through automated means, something that reflects a political choice and not a matter of legal technique.

[30] Actual assurance levels should be carefully evaluated; despite claims that most relevant algorithmic decisions come from deterministic computation [28], some operations still depend on randomness, either for the outcome itself (as is the case in a lottery) or due to the error factors involved.

[31] The existence of standards for comparison might even provide a basis for a scenario mentioned by Kamarinou, Millard, and Singh [19]: the possibility of automated review of human decisions.

[32] GDPR Article 22(3) explicitly mentions human intervention as a minimum requirement for system safeguards.

[33] Going even further, Brennan-Marquez and Henderson [2] sustain that some sorts of decisions, such as legal judgment, can only be seen as democratically valid if the decision-making process is *role-reversible*, that is, if the decision-maker can, at least in principle, be subject to the effects of a decision. This idea, related to the idea that nobody is above the law in a democratic society, establishes an explicit condition for the automation of democratic processes: the use of an AI would only be valid if an intelligent system can be subject to the sort of decisions that one intends to automate. On the other hand, attributing legal — or even moral — patiency to artificial intelligence systems intensifies the tensions between the instrumental role attributed

intervention even if, as discussed in Section 3, automated decision-making somehow becomes more efficient in preserving specific rights and interests than a human intervenor could be.

This new basis for the right to intervention leads to a follow-up question: how can automated decision-making diminish human dignity? One answer to this question is provided by Hildebrandt [15], who claims that some aspects of human personality are incomputable, not merely as a result of inadequate modelling, but rather as a necessary consequence of the human condition. Since computer models — and, therefore, intelligent systems as currently understood – cannot, by definition, process those incomputable aspects, any decision made by them would necessarily be based on an incomplete portrait of the natural persons affected by the outputs of the processing algorithm. A right to contesting algorithmic decisions would, in Hildebrandt [15]'s account, allow an "agonistic debate" about automated models, enabling humans to retain control over the ways their lives are presented and shaped by those algorithms.

Hildebrandt [15]'s emphasis on the "incomputable self" does not, by itself, establish grounds for the social or moral unacceptability of evaluating a natural person based on their quantifiable aspects, a practice that, as Fourcade and Healy [13] point out, dates at least to the nineteenth century in bureaucracies and markets alike. Yet, the more general idea of "agonistic machine learning" can be desirable even if one does not buy the additional premises adopted by Hildebrandt [15], as it would subject modelling and processing choices to internal and external evaluation regarding their compliance to the standards set by science, law, and democratic values.[34]

The role of data protection laws, as well as other regulations concerning automated systems, would be to ensure that automated decisions "explore and enable alternative ways of datafying and modelling the same event, person or action" [15], a requirement that includes, but goes beyond, *ex post* remedies to harms and undue constraints ensuing from personal data and its processing. Such a reading of the right to human intervention can be validated through a broad interpretation of the GDPR,[35]. but even within a more restrictive formulation, "agonistic machine learning" can be used as a frame for interpreting data protection regulations in light of the broader commitments to human rights and human dignity that data protection laws aim to preserve.

4.1 Predictive contestability

So far, we have discussed human intervention as a *post hoc* reaction to automated decision-making: a data subject can request intervention whenever they feel that a decision based solely on automated data processing harms or otherwise frustrates their rights, freedoms, or legitimate interests. Thus, intervention can be seen not only as a protection against intentionally harmful algorithms but also as a recognition of the fact that even software designed according to the best technical practices can fail in specific cases. If such failures are inevitable, then human intervention at least provides data subjects with means to seek redress without having to resort to the judicial system, reducing their costs and the time that they must wait until their situation is improved.[36]

Yet, there are legal grounds for sustaining that the concept of human intervention should not be understood solely as a reactive measure. In the European Union, GDPR Article 25(1) establishes that the protection of the rights of data subjects — a category that includes the right to human intervention — should be pursued through the adequate technical and organisational means "both at the time of the determination of the means for processing and at the time of the processing itself". Therefore, automated decision-making should not merely seek to provide *post hoc* remedies, but rather consider at every step the rights and interests of those potentially affected by the resulting decision.

But what is the meaning of a preventive approach to human intervention? In one sense, a preventive approach can be seen as an *early intervention* which allows data subjects to contest the hypotheses and design choices made at each stage of an AI's design and deployment, rather than just the decisory outcome of automated processing, through the use of approaches such as participatory design [11]. While trade secrecy or practical concerns might constrain the extent to which this approach can be feasibly adopted, early feedback might flag potential issues with an automated decision-making process before they can manifest themselves through harmful outcomes.

Another preventive approach to human intervention would be to consider, at each stage of software development and deployment, what can be done to allow a data subject to properly contest a decision that might result from the final system. Measures toward that goal might include the construction of auxiliary platforms for making intervention requests or the representation of decision rules in a human-accessible format, among other possibilities. This approach can be deployed either as a substitute or as a complement to early intervention procedures, and can substantially lower the effort that a data subject must make to exercise their rights.

A system that follows either reading (or both) can be said to be *contestable by design*, as the possibility of human contestation of the ensuing decision will be part of its acceptance criteria.[37] Contestability by design can be seen as an analogue of a more established concept: *privacy by design* (PbD),[38] that is, the idea that privacy-related concerns should be treated as software quality attributes

to intelligent systems and human moral intuitions, which leads Bryson [5] to claim that AI moral subjectivity, even if possible, is not a socially desirable outcome.

[34] A secondary benefit of multiple, differing, models for the same phenomenon can be seen in the use of *ensemble models* as solutions to artificial intelligence problems. Since in many cases it might be too complicated to build a model that covers all possible scenarios, some computer science techniques instead build a series of independent models, whose answers are then combined to provide a final output. Drawing from that, legally enforced adoption of multiple modelling perspectives might, at least in theory, lead to a better problem-solving outcome than a very good, singular, model would obtain, even if the results are not as clear-cut as usually claimed by readings of Hong and Page [18] such as the ones critiqued by Brennan [3].

[35] As anticipated by Kamarinou, Millard, and Singh [19].

[36] Drawing from the "crashworthy" doctrine in American tort law, Choi [8] sustains that damage mitigation should be considered when measuring tort liability ensuing from software. Adequate arrangements for human intervention could then reduce an AI user's liability, as they allow data subjects to contain the harmful effects of automated decisions if the grounds for intervention are detected early enough for it to have a meaningful impact.

[37] Individual developers and hobby projects usually do not have development processes that are structured enough to benefit from design-based approaches to safeguarding rights. However, most AI projects with real impact are complex, and their development usually follows, to a lesser or greater extent, software engineering practices which may be effectively shaped by the regulatory demands.

[38] Privacy by design was originally developed within the professional software engineering community [17], but GDPR Article 25 places that approach as a paradigm for data protection, as acknowledged in the article title itself.

[17] that are considered from the initial stages of development to the end of a system's operational lifetime. However, contestability by design should not be reduced to a subset of privacy by design, as those approaches seek different goals — the protection of human self-determination and control over automated systems for the former; for the latter, preservation of the private sphere — that might be pursued in independent or even conflicting ways: for example, system designers might have to sacrifice some user privacy in order to allow data subjects to have access to data that would be relevant to their decisions on whether a decision should or not be contested. Or, in the opposite direction, it might not be possible to disclose information that is relevant for an algorithm's decision-making processes without revealing personal information from other data subjects. Adequate conciliation of those standards will depend on the circumstances of each practical case, but it requires a clear understanding of what are the goals that contestability by design intends to achieve.

At a bare minimum, a system that is contestable by design should be built in ways that allow users and third parties to effectively seek human intervention in a given automated decision. The exercise of this right will usually require that the interested parties have access to relevant information, as discussed in Subsection 3.1, and those parties must also have access to adequate channels — automated or otherwise — for requesting intervention. Contestability by design also encompasses measures that, even if not explicitly demanded by the pertinent regulations, ensure that the rights and interests of data subjects are protected and, if that line of protection fails, that they can effectively exercise the right to intervention in ways that comply with the protection of the incomputable self.

4.2 Designing contestable systems

Contestability by design can draw lessons not only from the existing privacy by design literature [17] but also from experiences in applications such as psychotherapy [16], in the nature of the application already requires extensiveh interaction between the automated decision-making processes and the affected data subjects. While the extension of those best practices to new domains might present issues of their own — such as those stemming from the scale of the automated processing involved, for example, in profiling customers at a large eCommerce —, AI designers may benefit from techniques and insights originally used for interaction design or privacy protection.

One approach to deal with the right to human intervention during the software development cycle would be to directly incorporate, at each stage, feedback from relevant stakeholders, that is, the persons or roles that are directly or indirectly affected by the system [29]. Approaches such as participatory design [11] can be used to involve stakeholders in the identification of potential hazards to rights and interests that might arise from the use of the completed system, and what kinds of information would be needed to identify and address those concerns. This sort of feedback from direct and indirect stakeholders provides a form of *ex ante* intervention in the automated processing [12], as it acts over the processing system to prevent harms and undue constraints to rights and interests instead of redressing them. Since many use cases of intelligent systems may directly or indirectly affect the rights and legitimate interests of many people and organisations, direct engagement with all parties might not always be feasible. The use of adequate sampling techniques may, however, compensate for this issue, for selecting which stakeholders to involve and by engaging with entities — either public, such as data protecting authorities, or private, such as private certification agencies [12] or NGOs — that represent those collective interests that would otherwise go unheard.

Another line of action for contestability by design would be to add non-functional software requirements that ensure that data subjects will have the necessary information and tools to exercise the right to intervention. That could happen by considering that an automated processing system is only ready for production if it is possible to provide explanations [4] of its behaviour. Alternatively, a more general approach would require the use of techniques such as zero-knowledge proofs [21] to provide "opaque" assurances that a system behaves as intended,[39] providing information regarding the relations between inputs and outputs without revealing the actual algorithms used.[40] Non-functional requirements may also be used to require that automated processing systems provide adequate interfaces — either through the system itself or by processes involving humans — that allow users to exercise their right to human intervention.

Digital means might not only provide an interface for requesting a human intervention: as discussed in Subsection 3.2, it is at least theoretically possible that algorithms can avoid some, if not all, of the biases present in human decision-making. Under those circumstances, it might be interesting to automate the review process itself, for example, by using a trusted third-party algorithm to automatically review a decision. Current legislation does not allow for the complete removal of the human from the review loop, but there is no *a priori* ban against human-supervised use of automatic review tools, as long as the process does not become solely based on automated processing, as discussed in Section 2. Adequate use of existing — and future — technologies may, therefore, enable data subjects to better exercise the right to human intervention while reducing the corporate and judicial overhead.

5 CONCLUSION

Allowing data subjects to contest automated decisions that affect them is a useful tool for safeguarding the rights, freedoms, and legitimate interests of persons, natural or otherwise. Under regulations such as the GDPR, this right does not affect just those systems which entirely remove human beings from the decision loop, but it also is relevant in cases where the human decider makes choices within a space that is entirely defined by artificial intelligence.

However, allowing data subjects to seek redress after a decision is made might lead to a less-than-optimal protection, either due to the lack of information that is necessary for the proper exercise of that right or because a human intervention can result in a worse

[39] In this case, the exercise of the right to intervention would be based not on *ex post* explanations of individual decisions, but on the general rules and parameters that govern the overall functioning of the system. Based on opaque controls, users could then verify if a specific factor, such as race, affects the decision that concerns them, while at the same time preserving the secrecy of the algorithm itself [21].

[40] This sort of technique is particularly relevant when dealing with the use of systems or components that are not developed by the organisation that uses it, as is the case with many service-based solutions or when organisations use open-source solutions.

outcome than the original automated decision, either due to deliberate action or due to the biases and limitations of human thought. In this context, ensuring contestability from the early stages of software design can be seen as a way to enable human intervention, both before the automated data processing is in place and after meaningful decisions have happened. Therefore, legal and technical systems should foster contestability by design practices to ensure human control over decisions based on automated processing.

ACKNOWLEDGMENTS

The author would like to thank Victor Luís Nascimento Barroso, Enrico Roberto and Juliano Maranhão for their valuable feedback. This work was partially funded by a Lawgorithm research grant.

REFERENCES

[1] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'it's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.

[2] Kiel Brennan-Marquez and Stephen E Henderson. 2019. Artificial intelligence and role-reversible judgment. *Journal of Criminal Law and Criminology*. Forthcoming. Available at SSRN: https://dx.doi.org/10.2139/ssrn.3224549.

[3] Jason Brennan. 2017. *Against Democracy*. Princeton University Press.

[4] Maja Brkan. 2018. Do algorithms rule the world? algorithmic decision-making in the framework of the GDPR and beyond. In *Terminator or the Jetsons? The Economics and Policy Implications of Artificial Intelligence*. Technology Policy Institute. Washington, (February 2018).

[5] Joanna J Bryson. 2016. Patiency is not a virtue: AI and the design of ethical systems. In *2016 AAAI Spring Symposium Series*.

[6] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25, 3, 273–291.

[7] Jenna Burrell. 2016. How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data & Society*, 3, 1.

[8] Bryan H Choi. 2019. Crashworthy code. *Washington Law Review*, 94, 1.

[9] Márcio Cots and Ricardo Oliveira. 2018. *Lei Geral de Proteção de Dados Pessoais Comentada*. Thomson Reuters Brasil, São Paulo.

[10] Bo Cowgill and Catherine Tucker. 2017. Algorithmic Bias: A Counterfactual Perspective. Working Paper. NSF Trustworthy Algorithms.

[11] Janet Davis. 2009. Design methods for ethical persuasive computing. In *Proceedings of the 4th International Conference on Persuasive Technology*. ACM, 6.

[12] Lilian Edwards and Michael Veale. 2018. Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"? *IEEE Security & Privacy*, 16, 3, 46–54.

[13] Marion Fourcade and Kieran Healy. 2016. Seeing like a market. *Socio-Economic Review*, 15, 1, 9–29.

[14] Bryce W Goodman. 2016. Economic models of (algorithmic) discrimination. In *29th Conference on Neural Information Processing Systems*. Volume 6.

[15] Mireille Hildebrandt. 2019. Privacy as protection of the incomputable self: from agnostic to agonistic machine learning. *Theoretical Inquiries of Law*, 20, 1.

[16] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing contestability: interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, 95–99.

[17] J-H Hoepman. 2018. Making Privacy By Design Concrete. Technical report. KPN CISO Office, The Hague.

[18] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *PNAS*, 101, 46, 16385–16389.

[19] Dimitra Kamarinou, Christopher Millard, and Jatinder Singh. 2016. Machine Learning with Personal Data. Queen Mary School of Law Legal Studies Research Paper 247. Queen Mary, University of London, United Kingdom.

[20] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics*, 133, 1, 237–293.

[21] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *University of Pennsylvania Law Review.*, 165, 633.

[22] David Lehr and Paul Ohm. 2017. Playing with the data: what legal scholars should learn about machine learning. *UC Davis Law Review*, 51, 653.

[23] Isak Mendoza and Lee A Bygrave. 2017. The right not to be subject to automated decisions based on profiling. In *EU Internet Law*. Springer, 77–98.

[24] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *FAT* '19: Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA.

[25] Nathalie Nevejans. 2016. European civil law rules in robotics. Technical report. European Parliament, Brussels.

[26] Michael Polanyi. 2009. *The tacit dimension*. University of Chicago Press.

[27] Antoni Roig. 2018. Safeguards for the right not to be subject to a decision based solely on automated processing (article 22 GDPR). *European Journal of Law and Technology*, 8, 3.

[28] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law*, 7, 4, 233–242.

[29] Ian Sommerville. 2011. *Software Engineering*. (9th edition). Pearson Education Ltd, London.

[30] Christian B. Sundquist. 2019. Uncovering juror racial bias. *Denver University Law Review*, 96, 1.

[31] Cass R Sunstein. 2019. Algorithms, correcting biases. *Social Research*, 86, 2. Forthcoming. SSRN draft.

[32] Michaël Trazzi and Roman V Yampolskiy. 2018. Building safer AGI by introducing artificial stupidity. *arXiv preprint arXiv:1808.03644*.

[33] WP21. 2018. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. Technical report. Article 29 Data Protection Working Party, (February 2018).

[34] Tal Zarsky. 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.

[35] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2018. Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology*, 1–23.

[36] Baobao Zhang and Allan Dafoe. 2019. Artificial Intelligence: American Attitudes and Trends. Technical report. Center for the Governance of AI, Future of Humanity Institute, University of Oxford, (January 2019).