Transparency as design publicity: explaining and justifying inscrutable algorithms

Loi, Michele
Digital Society Initiative University of Zurich
Zurich, Switzerland
michele.loi@uzh.ch

Ferrario, Andrea
Mobiliar Lab for Analytics at ETH, and
Department of Management, Technology, and Economics, ETH Zurich
Zurich, Switzerland
aferrario@ethz.ch

Eleonora Viganò
Digital Society Initiative University of Zurich
Zurich, Switzerland
eleonora.vigano@ibme.uzh.ch

Abstract:

In this paper we argue that transparency, just as explanation, can be defined at different levels of abstraction. We criticize recent attempts to identify the explanation of black box algorithms with making their decisions (*post-hoc*) interpretable. These approaches simplify the real nature of the black boxes and risk misleading the public about the normative features of a model. We propose a new form of transparency, that consists in explaining the artifact as an intentional product, that serves a particular goal, or multiple goals (Daniel Dennet's design stance), and that provides a measure of the extent to which such goal is achieved, and evidence about the way that measure has been reached. We call such idea of transparency 'design publicity'. We argue that design publicity can be more easily linked with the justification of the use and of the design of the algorithm, and of each individual decision following from it. Finally, we argue that when models that pursue justifiable goals (which may include fairness as avoidance of bias towards specific groups) to a justifiable degree are used *consistently*, the resulting decisions are all *justified* even if some of them are (unavoidably) based on incorrect predictions. For this argument, we rely on John Rawls's idea of procedural justice applied to algorithms conceived as institutions.

1. Introduction

In this paper, we provide a new theory of algorithmic transparency, with a focus on both explanations and justifications, where we consider as 'algorithms' those human artifacts stemming from the training of machine learning models on digital data, in order to generate predictions to assist or automate decision-making. These algorithms are subject to intense scrutiny for both technical and moral reason, as their applications in product and services is constantly increasing, as well as their potential to affect everyone's lives. Examples come from credit scoring, to digital financial coaching and job assistants, automated insurance claim processing bots, smart home services, online dating platforms, autonomous driving solutions and policing as well as recidivism scoring algorithms. One current limitation of modern algorithmic-assisted decision-making is that most advanced machine learning models are considered as 'black boxes' or inscrutable (Selbst and Barocas 2018). Therefore, the last few years has seen the rise of an active debate in the scientific community around interpretability, transparency, explainability and justification of (machine learning model-based) algorithms and their outputs. Without a proper understanding of these constructs and their outcomes, any decision generated or supported by these algorithms cannot be adequately contested.

According to Lipton (2016), interpretations of machine learning models fall into two categories: model transparency and post-hoc explanations. Model transparency is 'some sense of understanding the mechanism by which the model works' (Lipton 2016, 4). Different ideas may be conveyed by demanding that a machine learning model be transparent, each focusing on different aspects of the model, its components and the training algorithm (Lipton 2016). On the other hand, post-hoc explanations focus on the outcome of the (learned) model; they include (Mittelstadt, Russell, and Wachter 2018) natural language processing explanations, visualizations, case-based and counterfactual explanations, and local approximations; they can be classified in model specific or model agnostic. Local approximation allow, in particular, to explain why a black box model produced a selected prediction by approximating it with an interpretable model (e.g. a linear regression) around the prediction at hand (Ribeiro, Singh, and Guestrin 2016). We refer to the goal of post-hoc explanations of individual decisions as 'model interpretability'.

In this paper, after providing some definitions (2), we start by highlighting some limitations of interpretable algorithms, in particular the method of counterfactual explanations, drawing support from the recent literature (Selbst and Barocas 2018; Kroll 2018; Kroll et al. 2016) (3). We then propose a new concept of algorithmic transparency, which we label 'transparency as design publicity', and argue that it provides a kind of *explanation* of their behavior: a *teleological* explanation,

or explanation by design (4). The special value of this explanation is that it links the behavior of an algorithm to their justification (5) and, when the algorithm is used consistently, to the *procedural justice* of its decisions.

2. Machine learning and algorithms: some definitions

In this section, we introduce some definitions that are relevant for the remainder of this contribution. The aim here is to provide the reader with an overview of some commonly used concepts in the most recent literature on philosophy of technology and artificial intelligence without indulging (too much) in technicalities and jargon. We start with *machine learning*, which is a multidisciplinary discipline "concerned with the question of how to construct computer programs that automatically improve with experience" (Mitchell 1997, 7). Machine learning draws on concepts from artificial intelligence, information theory, algorithmics and philosophy, among others. A machine learning problem 'can be precisely defined as the problem of improving some measure of performance P when executing some task T, through some type of training experience E" (Mitchell 1997). Training experience E is represented by (digital) input data, which are preprocessed and formatted for the machine learning problem under consideration. Performance measures P can be off-the-shelf or ad-hoc, that is engineered by the resources responsible for the solution of the corresponding machine learning problem; they provide with an estimation of the error made by the solution to the machine learning problem in executing T, using experience E.

What does it mean to solve a machine learning problem? Essentially, it consists of specifying a class H of mathematical constructs called *machine learning models*, to be trained on input data D using a set of *algorithms* implemented in computer-understandable programming languages (Mitchell 1997). Therefore, through the algorithms in the training process, the best machine learning model is trained or learned. The result of this process is an *object* in a programming language embedded in an IT infrastructure to generate predictions on new data with the goal to assist or automate decision-making; such complex, dynamic computer system becomes a "cognitive engine" at the core of products and services mentioned in (1). In the remainder of these notes, we will call this object "algorithm"; in fact, this is an algorithm - i.e. a procedure or rule to compute predictions from input (new) data points - and stemming from the training of machine learning models to solve a given machine learning problem. We will come back to the teleological nature of algorithms in (4).

3. Post-hoc Explainability: limitations of counterfactual explanations

As discussed by Selbst and Barocas "interpretability has received considerable attention in research and practice due to the widely held belief that there is a tension between how well a model will perform and how well humans will be able to interpret it" (2018, 1110). We now explain more theoretically, with reference to prior work by Selbst and Barocas, why post-hoc interpretability explanations lead to partial understanding and are apt to be misleading, and why this should be considered problematic. We shall focus on counterfactual explanations, which recently drew attention in the artificial intelligence research community (Wachter, Mittelstadt, and Russell 2017). In the same paper, the author states that counterfactuals[1] 1) are "a novel type of explanation of automated decisions that overcomes many challenges facing current work on algorithmic interpretability and accountability" (Wachter, Mittelstadt, and Russell 2017, 5), 2) "should be used as a means to provide explanations for individual decisions" (Wachter, Mittelstadt, and Russell 2017, 7), and 3) "can bridge the gap between the interests of data subjects and data controllers that otherwise acts as a barrier to a legally binding right to explanation" (Wachter, Mittelstadt, and Russell 2017, 5).

For simplicity, we do not consider the theory of counterfactuals and causality, limiting our considerations to machine learning counterfactuals only. We will argue that an individual, who understands a counterfactual explanation, without understanding the limitations of the approach, will potentially attribute to the feature mentioned in one or a limited set of counterfactuals an importance that is not objectively justified. The method of counterfactual explanation misleads by suggesting that some *ceteris paribus* clauses (among many) have an importance that they do not have. This is clearly an instance of the more general problem described by Selbst and Barocas (2018): contemporary machine learning models are designed to reflect the complexity of reality: they involve many variables, that interact in complex ways, which are hard to grasp intuitively. Counterfactual explanations achieve a reduction of such complexity, but "there will be situations where complexity cannot be avoided in a faithful representation of the scoring system, and listing factors alone will fail to accurately explain the decision"[2]. In particular, counterfactual explanations, like others, suggest that complex decision are explained by the causal role of a limited number of features. This can be problematic when, for example, there are in fact many features playing an equivalently important, or near equivalently important role (Selbst and Barocas 2018, 1115).

---

[1] In these notes, we will use the terms "counterfactual explanations" and "counterfactuals" interchangeably.

[2] In this passage, the authors refer to the eight-feature credit scoring system by Taylor and discussed in (Taylor 1980). We will consider the model later in this section, we discussing selection bias in counterfactual explanations.

Let us now discuss counterfactual explanations in some detail. To this end, let M be a machine learning model (implemented via algorithms in a computer understandable programming language and trained on a dataset D) that computes a real number p called score, for each data point x in D (or for new inputs). Data points are nothing but finite strings of alphanumerical values which are realizations of a finite set of (random) variables (V_1,...,V_n). In this setting, for each data point x, the machine learning model returns a score value p(x), which is a function of the values of the variables (V_1, ..., V_n) at the point x. The definition of counterfactual explanations is:

> statements taking the form: Score p was returned because variables V had values (v1, v2,...) associated with them. If V instead had values (v1', v2',...), and all other variables had remained constant, score p' would have been returned. (Wachter, Mittelstadt, and Russell 2018, 9)

From the very definition it follows that counterfactual explanations in the sense of Wachter, Mittelstadt, and Russell (2017) are an application of the *ceteris paribus* principle, namely they identify the explanation of a score with a factor whose change affects the score keeping all other factors equal. The world in which inference is generated following the ceteris paribus principle is represented by all the possible realizations of the random variables (V_1, ..., V_n).

The choice of which factors to keep constant and which one to vary affects the choice of counterfactual explanation to propose, which "would alter values as little as possible" (Wachter, Mittelstadt, and Russell 2018, 9) and it is claimed to metaphysically correspond to the closest possible world. An example of counterfactual explanation, offered again by Wachter, Mittelstadt, and Russell: "*You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan*" (2017, 5). Here a machine learning model classifies each individual's creditability as function of annual income and (possibly) other variables which are not mentioned in the counterfactual explanation itself. The sentence "*If you income...have been offered a loan*" represents the "counterfactual scenario", which is defined once new values for the variable "annual income" are generated, keeping all other variables, if any, as fixed (that is, *ceteris paribus*). The data point characterized by the annual income value of £45,000 is said to be synthetic. We note that in the counterfactual scenario, the outcome of the modelling exercise - i.e. the credit-worthiness level expressed by the acceptance or rejection of a loan request - changes with respect to the original data point, which is characterized by an annual income of £30,000.

Clearly, counterfactual explanations intercept only local properties of the model, where locality refers to a given data point x whose outcome is the *explanandum* under consideration. We note that, in general, a synthetic data point could be close to the original data point under

consideration, while showing different values with respect to it for all the variables in the data set. We argue that counterfactual explanations are vulnerable to at least three objections:

1) *Selection bias in ceteris paribus*: in case *ceteris paribus* is used to generate counterfactuals, one has to choose a subset of variables to generate synthetic data points. This choice reflects personal bias on interpretability of the machine learning model at hand by the individual performing the counterfactual routines. To our knowledge, no study focuses on rationale for the selection of variables to generate counterfactuals, or discusses empirical evidence from experiments on this topic. The choice of the variables to keep fixed in the *ceteris paribus* procedure is linked with the problem of informativeness of the outcome explanations. This is the case of Taylor's eight-feature credit-scoring system (Taylor 1980), where "choosing arbitrarily among equivalently valid reasons runs counter to the instruction to give specific and actionable notice" (Selbst & Barocas 2018, 1103) and "If the creditor tried to explain these rules simply, it would leave information out, but if the creditor were to explain in complete detail, it would likely overwhelm a credit applicant" (Selbst & Barocas 2018, 1103).

2) *Closest world arbitrariness and metaphysical coherence*: algorithms to generate synthetic data points use distance functions. One wants to select synthetic data maximizing the distance with respect to the original data point when considering the machine learning outcome (e.g. "loan" vs. "no loan"), but points closed to the original data point (Wachter, Mittelstadt, and Russell 2017). Clearly, distance functions are arbitrary. A throughout analysis of counterfactuals generation as function of different distance functions (while tuning their parameters) is needed, especially given what are considered open problems in the logic of counterfactuals, relative to the identification of the "nearest possible world". Several philosophers have analyzed counterfactual claims by appealing to this concept (Todd 1964; Stalnaker and Thomason 1970; Lewis 2013; Nute 1976). No proof has been given that any mathematically defined distance function corresponds to the relation "being the closest possible world" invoked in the logic of counterfactuals.

3) Lack of ontological scalability: the problem of closest world arbitrariness may be overcome by assigning to the "interpreter" of a machine learning model the task of excluding some potential explanations based on *common-sense* assumptions about what counterfactuals are irrelevant (e.g. "if your 50 m^2 apartment had 4 rooms, it would sell at $300'000"), as counterfactual possibilities too distant from the real world are not useful as explanations. If that is the way counterfactual explanations are supposed to be given, the method faces the problem of *lack of scalability.* Mathematical routines can generate synthetic data points but do not guarantee their consistency with the criterion of possible world similarity based on the common-sense intuitions about the closeness of possible worlds of the interpreter. In a big-data context, i.e. in presence of hundreds

or thousands of variables and synthetic data points, hard-coding constraints in the synthetic data generating algorithm that reflect *a priori* criteria of plausibility or possibility is an unviable strategy, due to the time needed for considering all possible scenarios. In summary, counterfactual explanations based on the generation of synthetic data points do not scale in big-data contexts and suffer of potential lack of ontological coherence.

4) Lack of normative informativeness A fourth problem affecting counterfactual explanations is the one of inferring normative properties of the model from explanations of *individual decisions*. As will be later argued, we think that the most practically important normative properties of model-based decisions emerge from repeated application of the model – they are properties of the kind of patterns, e.g. the distribution of errors, or of benefits, between groups, or of groups defined by morally salient properties, that emerge when the law of large numbers applies. A case in point is indirect discrimination or disparate impact, which can be considered morally or legally relevant in certain contexts, but cannot be determined with such methods, because it can exist even when information about a protected category plays no role at all in a decision (Selbst and Barocas 2018, 1105). Another example is the property of separation – the fact that the false positive and false negative rate is statistically dependent on membership to a protected group (Hardt et al 2016). The risk here is to falsely infer that, since sex or race play no role in the counterfactual explanation of decisions by an algorithm, the system cannot be biased against a particular sex or race. Interestingly, the opposite misunderstanding may also occur. In some cases, protected group characteristics may be necessary to avoid discrimination in data-driven decision models, for example because they are necessary to avoid omitted variable bias (Žliobaitė and Custers 2016). A counterfactual explanation may show that a decision, e.g. concerning a loan, would have been different had the individual been of a different race. This may lead the public to consider the system discriminatory, even when the information about the protected group is used to make the prediction fairer.

Thus, we conclude that *counterfactual* explanations do not ensure that there is a way to evaluate the justification of algorithmic decision-making. In what follows, we provide a model of transparency that relies on explanations that are relevant for the justification of algorithmic decisions and, thus, their public acceptability.[3] We do not maintain that transparency as design

[3] By public acceptability we do not mean public in the sense of Rawlsian public reason (Rawls 1996; Binns 2018), which involves standards of justification which can be shared by individuals with different conceptions of the good sharing a commitment to core liberal and democratic values and principles. We assume that different standards of justification will be employed in different contexts and by different publics.

publicity – the approach we propose – fulfill all the *desiderata* various authors have associated with explainable and interpretable AI. Our transparency idea serves a particular purpose: that of normative justification. It provides the kind of explanation which is useful for the public to assess if the deployment of algorithmic decision-making and the decisions following from it is justifiable.

4. Design explanation of algorithms and their property of design, value, translation, performance, and consistency transparencies

As showed by Kroll (2018), the thesis that the understanding and transparency of algorithmic-assisted decision-making is limited by the inscrutability of the machine learning models and their algorithms (i.e. the fact that they are opaque or "black boxes") is criticizable. The debate on algorithm inscrutability mostly depends on the meaning we attribute to the expression "explaining the model' and accordingly "understanding the model".

Explanation - the process and product (Ruben 2012) of making something understandable - has many meanings: definition, interpretation, individuation of the necessary conditions or sufficient conditions, of purposes, of functions, and of goals. An explanation is effective when the *x* that is explained is clear and open to people that want to understand *x*. An effective explanation renders an object understandable and its understandability contributes to the transparency of the object, i.e. the quality of being easy to see through, analyze, and assess.

The explanation of the behavior of an algorithmic system has not only different meanings but also different levels of abstraction to which it can refer. If the explanation of a model is meant in mechanistic terms, then the algorithm functioning may be difficult to understand even to computer scientists and engineers. An explanation that clarifies *the purpose* of algorithms would be understandable to the public, from the end users with low expertise to policymakers in the need of justifying the use of algorithmic-assisted decision-making, to corporate executives adopting such models, to computer scientists and engineers that design them. The explanation of the purpose of an algorithmic system is an explanation by intelligent design (or, more briefly, a *design explanation*), namely it explains an *x* by referring to that for the sake of which *x* was created. This explanation is more abstract than the mechanistic one and corresponds to Dennett's design stance, namely the intellectual strategy by which we explain the behavior of a system by referring to its purpose and intentional design (Dennett 1987). Design explanations are teleological and focus on

the final cause of a system (Aristotle, Phys. II, 3; Metaphys. V, 2).[4] Design explanation is applicable to algorithms as the latter are goal-directed, they are human artifacts produced in a specific sociotechnical context (Baker 2004). In the design explanation of a common object such as a chair, we provide the reasons for which the chair was designed as such: being stable and comfortable; these goals directed the design of the chair and explain why respectively the chair has four or three legs and has an ergonomic or flat surface in the spot in which we sit down. The design explanation of an algorithm comprises "the understanding of what the algorithm was designed to do, how it was designed to do that, and why it was designed in that particular way instead of some other way" (Kroll 2018, 4). In other words, explaining the purpose of an algorithm requires giving information on various elements: the goal that the algorithm pursues, the mathematical constructs into which the goal or its proxy is translated in order to be implemented in the algorithm, and the tests and the data with which the performance of the algorithm was verified. *Design transparency,* the providing of information about these elements, contributes to an explanation of the artifact *by design*. As design explanation is made of different elements, similarly design transparency can be split into various components: *value transparency*, *translation transparency*, and *performance transparency*, as we will now show.

The goal of an algorithm is something valuable that is achieved. Since it is something that is desired by a person or group, we can also call it a good or *value* for that person or group. Value transparency should also indicate why and for whom the goal is valuable, when this is not obvious from the context. The design *goal* (e.g. identify the most profitable clients, minimize hospital readmissions) is typically also the *goal* of the person who decides to employ the artifact in practice. Thus, it also figures in the *intentional explanation* of the action to develop, or purchase, and employ the AI, by the persons accountable for such decisions. Thus, the design explanation should indicate which is the goal – the reasons or motivations – of the computer scientists and engineers who designed the algorithm and of the persons accountable for its employment in real-world settings. These goals should be one and the same; when this is not the case the artifact does not respond to the reasons of the person who are supposed to have meaningful human control (Santoni de Sio and Van den Hoven 2018) over it. This is problematic for accountability. The goal of an algorithm is usually a practical objective, such as profit or efficient allocation of scarce resources, but can

[4] The final cause described by Aristotle can be used to explain the behavior of entities with no psychological states (desires, beliefs, conscious purposes, etc.) such as algorithmic systems, as Aristotle applies the teleological model of explanation to natural processes, which have no psychological states (Broadie 1987; Gotthelf 1976).

include moral values such as equity, beneficence, trustworthiness, and the rules that are socially accepted as pertinent for the domain in which the model is employed. In both cases, the goal introduces normativity in the model, as it represents something that there are good reasons to pursue. Hence, normative choices are made both when normative standards are explicitly invoked in the design of a model and when they are ignored. As Binns points out:

> [W]hen attempting to modify a model to remove algorithmic discrimination on the basis of race, gender, religion or other protected attributes, the data scientist will inevitably have to embed, in a mathematically explicit way, a set of moral and political constraints (Binns 2018, 547).

The goals or values that guide the design of algorithmic models should therefore be included in an explanation of such models. *Value transparency* is the result of an explanation that makes the standards, norms, and goal that were implemented in the system accessible. These normative elements should also correspond to the *reasons for which* it was deployed.

The goal of an algorithmic system needs to be translated into something that is measured: a set of rules with which the algorithm elaborates inputs and produce outputs. A machine learning algorithm requires the quantification of the goal because, in particular, the algorithm that generates the model needs to quantify the departure from the model objectives of several potential candidate models. There is no straightforward and only one translation of a goal into a mathematic construct. For this reason, making such translation a publicly verifiable criteria provides the public and scientific community with the information to assess how a given goal is operationalized in machine-language. Making this piece of information public constitutes *translation transparency*, which is part of design transparency. In applications, it is possible to have alternative translations in machine language of the same goal. For example, let us consider the problem of designing a predictive model of customer churn[5] for an airline company. The goal is to design and implement a predictive model of customer churn in order to assess future profitability of a given portfolio of customers. However, in the case of an airline company, the business concept of "churn' could be translated into different set of computer-understandable rules. In one case, one could simply define a customer as churned if no revenue is generated by the customer in a given year of interest. On the other hand, one could introduce churn as the absence of revenue in a given year of interest and the lack of flying activities (i.e. avoiding the case of zero-revenue generating customers flying

[5] To churn or to lapse is the activity of moving out a given group. In business, it refers to the activity of customers to move out of portfolios. Predictive models of customer churns are important to organizations to predict the volumes of portfolios in (future) timeframes and to assess their (future) profitability.

using promotions). Both choices lead to alternative implementations of the same business goal. Design transparency recommends to explain the definition of churn and its motivations to the public. Another example is the goal of fairness or avoiding discrimination. Different definitions of fairness for predictive models exist and it is often impossible to satisfy all of them simultaneously (Berk et al. 2017). One definition requires that the probability of a favorable prediction be statistically independent of group membership (statistical parity). Another one, proposed by Hardt (2016), requires the favorable probability to be statistically independent of group membership only for those individuals who have the actual feature that the predictive model tries to predict (equality of opportunity). Design transparency requires declaring which fairness definition has been adopted, and, if possible, to provide a justification of such choice.

Once the criteria to measure degree of goal achievement are specified, a design explanation of an algorithm should provide information on the effective achievement of such objectives in the environment for which the system was built. In fact, for instance, the mere implementation of the most advanced norms of equal treatment in a credit-granting system does not warrant that the system will be effectively impartial. The impact of the algorithms and its outcomes needs to be considered. *Performance transparency* consists in indicating the logic with which the algorithm has been tested in order to verify how much it departed from achieving the goal and in indicating the results of such logic, starting with the choice of performance measures used in both training phase and during the assessment of the model on test data.[6] These latter are data that have not been used during training and whose scope is to assess the adaptability of the model to unseen inputs. The test data are part of performance transparency as the choice and the quality of them, which can be subject to biases, influences the performance measure and thus the assessment of the algorithm.

In summary, an algorithmic system has the property of design transparency if and only if it provides the public with the goal of the algorithm (value transparency), how this goal was translated into programming language (translation transparency), how the algorithm rule achieves that goal and how the goal achievement has been assessed (performance transparency).

An important difficulty here concerns the explanation of the singular decisions by the artifact, which should be distinguished logically from the nature of the artifact itself. The algorithm's performance connects the explanation of the artifact (i.e. an algorithm, or rule) with

[6] Training and test data are often the result of a random split of an original set of data used for modelling purposes. This implies that the object resulting from training and the outcomes of which are the object of the explainability analysis is in reality a *pair* consisting of the model and a random seed, which is the integer value chosen by the analyst that governs the *randomness* in the routines leading to the training of the model itself.

the application of the rule to particular cases. The simple solution is to view each individual decision as a means through which the artifact achieves the overall goal for which it has been designed. This explanation is however problematic in the light of the fact that, when algorithmic decisions are based on statistical predictions, it will often fail to decide in a way that directly promotes the goal the model is designed to achieve. E.g. a loan is refused to someone willing and able to repay it, an inmate who will not reoffend is denied parole, a patient is prematurely released from the hospital, causing readmission. This is due to the fact that decisions based about imperfect predictions about stochastic events will typically be often wrong, but sufficiently often right to justify the use of the model in practice. In the next section we are going to show why even the statistical imperfection of a model can be justified by appealing to its design goals and the trade-offs between all values pertaining to the justification of its use.

There is a further type of transparency – *consistency transparency* – that contributes to explain individual decisions by algorithms, given the assumption that the employment of such systems should be minimally fair. Consistency transparency is showing proof that consistency is achieved, i.e that the algorithm always generates predictions by the same rules even when we cannot observe those rules in operation. Consistency is not a feature of the model but of its *deployment*. It does not contribute to explain why the model works in a certain way, but why certain decisions are made (namely, they result from applying the model consistently). Consistency can even be a property of the deployment an algorithm that applies a discriminatory rule such as filtering job candidates by their residence address. Nonetheless, as consistency transparency shows that identical cases are treated identically, it represents the first step towards fairness; it is a sort of basic requirement of fairness that, as we shall show, is necessary but not sufficient to justify it.

In some cases, models are *unidentifiable*, by which we mean that in most AI powered solutions the underlying machine learning models are updated (i.e. retrained) with frequencies that depend on the domain of applicability of the solution itself. This implies that an AI potentially generates different outcomes for the same end user, depending on the moment at which the outcome is generated: any explanation of this outcome (for the purpose of contesting or auditing it, for example) depends on time, as well. Consistency transparency requires that changes in a model be declared because, as we shall maintain, this is relevant for their justification. Consistency is a normative goal and showing that it is achieved by the model contributes to explaining why an individual decision is made – namely, by showing that it is explained by a normative consideration. Conversely, the failure to satisfy consistency implies that the decisions of the model can be challenged on a specific normative ground.

In conclusion, the design explanation of the *model* shows that an algorithmic model gives a decisional outcome because the model pursues a certain goal (value transparency), which is translated into mathematical constructs implemented in the algorithm (translation transparency), which in turn enables to verify whether the model achieves the goal (performance transparency). When, as in most cases, consistency is among the reasonable goals of model *deployment*, the explanation of the *decisions* by the model includes consistency transparency.

5. Design publicity and justification

In what follows, we argue that transparency conceived as design publicity provides explanations that are useful to assess if the model is justified. The explanations provided by design publicity relate directly to the question of explanation. Design publicity provides information about a) the goal the algorithm is designed to pursue and the moral constraints it is designed to respect (value transparency); b) the way this goal is translated into a problem that can be solved by machine learning (translation transparency), c) the performance of the algorithm in addressing problem (performance transparency), and d) a proof of the fact that decisions are taken by consistently applying the same algorithm (consistency transparency). Let us now consider how each of these elements contributes to the justification of using an algorithm and of the decisions that follow from its use.

Let us begin with the goal or goals the algorithm is designed to pursue. All algorithms are designed to pursue *primary goal* (e.g. a business objective); some more advanced algorithms are also designed to take into consideration a *plurality of different values*, such as fairness or privacy, that often can be conceptualized as moral or legal *constraints*. *Constraints* are typically in trade-off with the primary goal and affect the way and the extent to which the primary goal can be achieved. For the sake of simplicity, we will refer to both goals and constraints (as "goals"), in what follows.

The first step of the *justification* of decision taken by an algorithm, thus, requires evaluating the goals and constraints that the algorithm, respectively, achieves and respects. In a justified algorithm, they reflect those values and constraints a reasonable person may want to see promoted/respected in the context of a service.

The *primary goal* of the algorithm matters for the justification of the decisions that follow from its employment on people, even if the goal is not what people commonly perceive as a political, moral or legal value. Consider the goal of maximizing the amount debt that is repaid, when lending. This optimization can be justified prudentially: the company cannot be profitable if it lends too much money to people who cannot repay their loans. The optimization goal can also be justified, morally, legally, or in terms of legitimate authorities. Prudentially speaking, a company needs to achieve this goal to be profitable and stay in business. Morally speaking, credit contributes

to wealth that contributes to the general welfare. Financial instability, on the other hand, has negative implications on the general welfare, as recent economic events – e.g. the subprime crisis – have shown. Alternatively, one may invoke the legal freedom to conduct business, which is legally protected in many countries.

Primary goals matter to justification when they are *valuable* goals, e.g. there are good reasons to pursue such goals, which can be explained by reference to moral, political or legal values. Other goals (the "constraining" goals) typically reflect value considerations, e.g. privacy, or fairness. Different types of justification are possible, for example in terms of common, or philosophical morality, of the law, of by virtue of political principles and values, that may be universal or characteristic of the society in which the model operates.

Take, for example, "anti-discrimination measures incorporated within [a] model to prevent it from giving higher credit risk scores to members of groups which have historically been discriminated against" (Binns 2018, 548). "Anti-discrimination" is the general name of a value that society expects from a service lending money to individuals, and that contributes to define the goal of the algorithms (in this case, by affecting the way and accuracy with which it will be pursued). Such goal is considered when designing the algorithm for moral, legal, or reputational reasons. Algorithmic transparency requires that these normative goals and the reasons for considering them are clearly specified – i.e. the choice of such normative goal is not a mere arbitrary decision by the data scientists. Goal transparency contributes to the ability of the public to understand and assess the validity of a potential justification for accepting decisions taken by a model pursuing such goal. If the goals pursued by a model are not values worth pursuing, the decisions following from the model are not justified.

An algorithm pursuing such goals will achieve them to a determined degree, which is expressed by "performance transparency". The performance can only be assessed by translating the goals in question into measurable quantities. This exercise of translation is not trivial. As Pak-Hang Wong observes "the idea of […] algorithmic fairness is […] contestable […] there is a great number of definitions of what […] algorithmic fairness amounts to, and it seems unlikely for researchers […] to settle on *the* definition of fairness any time soon" (Wong 2019). Translation transparency requires that the translation be declared and reasons, if any, for such choice. If the value translation is not justified, or at least justifiable, the decisions of the model are not justified.

Notice that *transparency as design publicity* – the approach we favor – does not require that individuals that are accountable for algorithmic decisions provide full justifications. It is sufficient that they declare the elements that are needed for one, that we identify with the values they are pursuing, the translation they adopt, and the *extent* to which the quantified values are achieved

(performance transparency). Performance transparency is especially important when there are trade-offs between different values simultaneously pursued by a model. Performance metrics provide an important indication of the extent to which every value has been achieved, which is especially important for the overall justification of the system when a value can only be achieved at the expense of another value. For example, fairness can only be pursued at the expense of efficiency; in the case of the recidivism predictions, optimizing for fairness measures leads to a partial failure to release some high-risk detainees, that are mistakenly classified as low-risk, or a partial failure to release low-risk ones (Corbett-Davies et al. 2017; Wong 2019). Performance transparency provides an indication of the degree to which both values, of efficiency and (quantified) fairness have been sacrificed.

There is still a gap in the justification of individual decisions. As anticipated, the fact that prediction-based decisions will often be wrong can be justified. In the case of machine learning-driven algorithms, individual mistreatment happens because the information necessary to always make perfect predictions does not exist. And even the information required to make a model *more* accurate may be too costly to collect, or cannot be collected in morally permissible ways. It is known that value-driven design that considers privacy and non-discrimination pays a price in terms of predictive accuracy (Hajian et al. 2015) and efficiency (Corbett-Davies et al 2017). A further reason why errors are unavoidable is that some outcomes result from human free will, for example, success during parole. The same considerations (of cost, privacy, or fairness) justify statistical decisions that rely on incomplete information, even when it is theoretically possible to collect and analyze all the information that matters, in principle, if one is to treat each individual case "as a distinct individual" (Lippert-Rasmussen 2010).

An individual subjected to an unfavorable decision may accept, in principle, that the algorithm is justified as a whole, yet challenge the *necessity* of implementing the model when taking a decision *about him*. The particular individual may argue: "I understand that the algorithm achieves these goals and that it does so in a reasonable way. But why can't you make an exception for me?". This would violate consistency. For example, suppose that a software is used to randomize access to scarce life-saving resources in a hospital of a dystopian country. This software translates fairness into a basic mathematical condition, which is equal chances of getting the resource in question. This goal can be achieved by an algorithm whose outcome is completely random. Yet consistency would be violated if, when the case of head physician's son is submitted to it, the randomized model is no longer used by the person in charge, who recognizes the head surgeon son, and assigns the resource to him. In this case, the software does not satisfy consistency.

The violation of consistency *for an arbitrary reason* (e.g. the head physician son's case) is incompatible with equal respect; on the other hand, if the same exception were made for everyone who had an interest to demand it, the algorithm wouldn't achieve its design goals, which justify its use. The violation of consistency is incompatible with formal justice "the impartial and consistent administration of laws and institutions" (Rawls 1999, 50–51), applied to the algorithm, considered as a law, or as an institution. This is why – we maintain – algorithms that change their identity as they are used are normatively problematic in high-stake decisions. In such cases, any change due to retraining should be at least publicized, and justified, by pointing out a considerable improvement in performance, which overrides consistency concerns.

When the design of an algorithm is justified, then, if the algorithm is also used consistently, we obtain a *procedural justification* of all the decisions that follow from it. To explain this kind of justification, we draw from Rawls's idea of the justification of individual shares of the goods produced by cooperation (1999, p 76-77). Rawls rejects the idea of allocative justice, namely, he rejects describing justice as a property of the end-state of process of the distribution of goods, a property independent from how that distribution came about. For example, an end-state distribution is just, according to a resource egalitarian account, if and only if resources are equally distributed, according to a meritocratic account, if and only if resources are proportional to each person's contribution to society, and according to a utilitarian one, if and only if the distribution maximizes utility. As Nozick (1974) observes, these allocative end-states are undermined by processes, like markets, that are not fully deterministic, because they are perturbed by human free decisions. In Nozick's slogan, liberty upsets patterns. Importantly, this applies to many cases of algorithmic decision-making, where the outcomes that justify the decision are future events that depend from the free will of an individual. According to non-compatibilist libertarianism, believing that an inmate success on parole could be predicted with perfect precision is tantamount to denying that the inmate has free will.

This gives us a moral reason to consider Rawls's procedural alternative to end-state conceptions. In this case, distributive shares are just if they result from just institutions. But unlike Nozick, Rawls relates the justification of institutions to the outcomes they tend to bring about, their general statistical tendencies, considered from a suitably general perspective. As in Hume it is the "general scheme or system of action, which is advantageous" (Hume 2000, secs. III-iii–1), not every single decision considered individually. The outcomes which justify the institutions are characterized by Rawlsian principles of justice. Rawls (1999), for examples, requires economic institutions as a whole (including taxation) to maximize the expectations of the worst off groups

in society. If institutions are justified, and if they are consistently and impartially applied, then the outcomes of free human decisions constrained by institutions are just, whatever they are.

For algorithmic decision-making, the principles of justice correspond to its design goals. The design goals of the algorithm are *that which justifies an algorithm which amounts to specific rules* (including inscrutable complex ones). To assess if inscrutable algorithms satisfy their "principles of justice" we consider their *performance*. If they do, the consistent and impartial application of the algorithm to individual cases corresponds to the consistent and impartial administration of just institutions. Summing up in one word: we are bound *by procedural justice* to accept as just *all decisions* that result from the consistent application of an algorithm that is justified by design.

Notice that, in the institutional case, the fact that institutions are administered consistently and impartially is a *public* fact. This publicity is achieved thanks to special procedures. E.g. the consistent fulfillment of the legal obligations emerging from civil law can be tested by going to court. In the algorithmic case, the consistent application of *inscrutably complex rules* appears to lack transparency. The solution to this is to provide a technical solution that delivers a proof that the rules are followed – that is, consistency – even when the rules themselves are not transparent to anyone because the algorithm is a black box; it appears that this is indeed technically feasible (Kroll et al. 2016, 665–71).

6. Conclusion

In this paper, we discuss what it means to achieve *transparency* for algorithms, i.e. the provision of explanations to see through, analyze, and assess artifacts trained on data *via* machine learning methods and generating predictions to assist or automate decision-making. We propose a form of transparency that consists in the publicizing the *design* of an artifact (including *value*, *translation* and *performance*) as well as its *consistent* application. We maintain that this kind of transparency provides 1) an *explanation* of the artifact, namely, an explanation "by design"; 2) an *intentional* explanation of its deployment; 3) a justification of its use; 4) when used consistently, a *procedural* justification of the individual decisions it takes.

The proposed approach to algorithmic transparency deviates from the existing body of literature on explainable artificial intelligence (xAI), where the concept of transparency focuses on the explanation of the inner workings of algorithms or the interpretability of their individual outcomes (Lipton 2016; Ribeiro, Singh, and Guestrin 2016). We do not claim here that transparency as design publicity achieves the goals that these approaches are said to achieve. Rather, we stress that transparency as design publicity achieves a distinct goal, namely, providing

the public with the essential elements that are needed in order to assess the justification (and, when consistency is satisfied, procedural justice) of the decisions that follow from its deployment.

7. References

Aristotle. 2018. *Physics*. Indianapolis: Hackett.

Aristotle. 2016. *Metaphysics*. Indianapolis: Hackett.

Baker, Lynne Rudder. 2004. "The Ontology of Artifacts." *Philosophical Explorations* 7 (2): 99–111. https://doi.org/10.1080/13869790410001694462.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *ArXiv:1703.09207 [Stat]*, March. http://arxiv.org/abs/1703.09207.

Binns, Reuben. 2018. "Algorithmic Accountability and Public Reason." *Philosophy & Technology* 31 (4): 543–56. https://doi.org/10.1007/s13347-017-0263-5.

Broadie, Sarah. 1987. "Nature, Craft and Phronesis in Aristotle." *Philosophical Topics* 15 (2): 35–50.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." *ArXiv:1701.08230 [Cs, Stat]*, January. https://doi.org/10.1145/3097983.309809.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, Mass: MIT Press.

Gotthelf, Allan. 1976. "Aristotle's Conception of Final Causality." *The Review of Metaphysics* 30 (2): 226–254.

Hajian, Sara, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2015. "Discrimination- and Privacy-Aware Patterns." *Data Mining and Knowledge Discovery* 29 (6): 1733–82. https://doi.org/10.1007/s10618-014-0393-7.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *ArXiv:1610.02413 [Cs]*, October. http://arxiv.org/abs/1610.02413.

Kroll, Joshua A. 2018. "The Fallacy of Inscrutability." *Phil. Trans. R. Soc. A* 376 (2133): 20180084. https://doi.org/10.1098/rsta.2018.0084.

Kroll, Joshua A., Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. "Accountable Algorithms." *University of Pennsylvania Law Review* 165: 633. https://heinonline.org/HOL/Page?handle=hein.journals/pnlr165&id=648&div=&collection=.

Lewis, David. 2013. *Counterfactuals*. John Wiley & Sons.

Lippert-Rasmussen, Kasper. 2010. "'We Are All Different': Statistical Discrimination and the Right to Be Treated as an Individual." *The Journal of Ethics*, 1–13. http://dx.doi.org/10.1007/s10892-010-9095-6.

Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." *ArXiv:1606.03490 [Cs, Stat]*,

June. http://arxiv.org/abs/1606.03490.

Mitchell, Tom M. 1997. *Machine Learning*. 1 edition. McGraw-Hill Education.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. 2018. "Explaining Explanations in AI." SSRN Scholarly Paper ID 3278331. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3278331.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Nute, Donald. 1976. "Counterfactuals and the Similarity of Words." *The Journal of Philosophy* 72 (21): 773–778.

Rawls, John. 1996. *Political Liberalism*. Expanded ed. New York: Columbia University Press.

———. 1999. *A Theory of Justice*. 2nd ed. Cambridge, MA: Harvard University Press.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *ArXiv:1602.04938 [Cs, Stat]*, February. http://arxiv.org/abs/1602.04938.

Ruben, David-Hillel. 2012. *Explaining Explanation*. Updated and expanded 2nd ed. Boulder: Paradigm Publishers.

Santoni de Sio, Filippo, and Jeroen Van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5. https://doi.org/10.3389/frobt.2018.00015.

Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." *Fordham L. Rev.* 87: 1085.

Stalnaker, Robert C., and Richmond H. Thomason. 1970. "A Semantic Analysis of Conditional Logic 1." *Theoria* 36 (1): 23–42.

Taylor, Winnie F. 1980. "Meeting the Equal Credit Opportunity Act's Specificity Requirement: Judgmental and Statistical Scoring Systems." *Buff. L. Rev.* 29: 73.

Todd, William. 1964. "Counterfactual Conditionals and the Presuppositions of Induction." *Philosophy of Science* 31 (2): 101–110.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *SSRN ELibrary*, November. https://ssrn.com/abstract=3063289.

Wong, Pak-Hang. 2019. "Democratizing Algorithmic Fairness." *Philosophy & Technology*, June. https://doi.org/10.1007/s13347-019-00355-w.

Žliobaitė, Indrė, and Bart Custers. 2016. "Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models." *Artificial Intelligence and Law* 24 (2): 183–201.