The philosophical basis of algorithmic recourse*

Suresh Venkatasubramanian†
suresh@cs.utah.edu
University of Utah

Mark Alfano†
mark.alfano@gmail.com
Delft University of Technology
Macquarie University

ABSTRACT

Philosophers have established that certain ethically important values are modally robust in the sense that they systematically deliver correlative benefits across a range of counterfactual scenarios. In this paper, we contend that recourse – the systematic process of reversing unfavorable decisions by algorithms and bureaucracies across a range of counterfactual scenarios – is such a modally robust good. In particular, we argue that two essential components of a good life – temporally extended agency and trust – are underwritten by recourse.

We critique existing approaches to the conceptualization, operationalization and implementation of recourse. Based on these criticisms, we suggest a revised approach to recourse and give examples of how it might be implemented – especially for those who are least well off[1].

KEYWORDS

recourse, algorithmic decision making, precarity, robust goods

ACM Reference Format:
Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3351095.3372876

1 INTRODUCTION

Human agents are distinctive among animals in the amount of long-term planning they engage in. We make plans that may come to fruition days, weeks, years, or even decades in the future. In some cases we even plan for events that will occur only after our own deaths. Such planning is remarkable not just for the amount of time involved but also for the level of recursive means-end reasoning involved. If your ultimate aim is, for instance, to vacation next year in Hawaii, you might go about it by saving money in order to be able to purchase a ticket. And you might go about saving money by getting a second job. And you might go about getting a second job by receiving certification to work that job. And you might go about receiving certification by taking vocational training courses. In this scenario, you take training in order to receive certification in order to get a second job in order to save money in order to purchase airfare in order to go to Hawaii. Such plans are only likely to succeed in a sufficiently well-ordered system, in which the reasons things might go wrong are foreseeable and understandable, and in which errors can be identified and rectified. If you could not trust that the vocational training would be sufficient to get certified, it would not make sense to plan in this way. Likewise, if you could not trust that hyperinflation would not destroy your savings, it would not make sense to plan in this way. The kind of agency that we both expect to be able to exercise and in many cases actually do exercise presupposes that our society is organized in a sufficiently regular, understandable, and corrigible way, which makes it possible to trust that the elaborate, *temporally-extended* planning we engage in is likely to be successful.

Among the things people typically make long-term plans for are various essential primary goods, such as housing. Computer scientists interested in algorithmic fairness have tended to focus on the distribution of such goods. In this paper, we are also concerned with their nature. In particular, we are interested in the fallback mechanisms and dispositions that people may be able to take advantage of when they lack an important primary good. In recent years, social scientists have begun to study the growing instability surrounding access to various primary goods. Researchers sometimes speak of the problem of *precarity*[5, 11, 21], which broadly speaking can be characterized as a state of precarious existence (or precarious access to resources like employment, housing, health care and so on) in which small "shocks" can remove access to such critical resources. Someone who suffers a precarious existence lacks financial and social security, which impinges on their ability to engage in temporally extended agency. They may have housing and a steady job today, but if anything were to go wrong in their life (e.g., a chronic illness, an unexpected financial burden, a parking ticket), they would lose their housing or job. A recent study in the state of New Jersey found that loss of driving privileges due to license suspension (which is often used as a punishment for reasons unrelated to driver safety) led to severe collateral impacts: 42% of the people whose license had been suspended reported reported losing their jobs [6]. This is unsurprising in a state with inadequate public transport: if you can't get to work, it's hard to hold down a job. Furthermore, of those who reported losing their job, 45% reported being unable to find a new job, and 88% of those who did find new employment reported a decrease in income.

*This research was partially funded by an Australian Research Council Discovery Project (code: DP190101507) and by the National Science Foundation under grant IIS-1633724

†Both authors contributed equally to this research.

[1]We focus on the least well off because this is arguably the most defensible principle from an ethical and political point of view.[27]

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372876

In a similar vein, a recent survey by bankrate.com found that only 41% of American adults would be able to cover an unexpected cost of $500 from their existing finances[2]. Likewise, the Federal Reserve Board [25] found that 46% of adults in the United States could not cover an emergency expense of $400 without having to sell something or borrow money. Someone who suffers from precarity in this way will find it hard to engage in temporally extended agency and to trust that their plans will come to fruition. They face a life of constant worry and stress, and such worry and stress can have knock-on effects that feed back into the precariousness of their existence. For example, stress and anxiety may lead someone to snap at their boss, which could get them fired. Elizabeth Anderson [2, pg. 63] estimates that approximately 80% of American workers – essentially, all those who are "neither securely self-employed nor upper-level managers" – are just "one arbitrary and oppressive managerial decision away" from being fired, demoted, or otherwise mistreated by the pervasive "authoritarian governance in our work and off-hours lives."

1.1 The modally robust good of recourse

The examples described in the previous section and others like them suggest that people will often need some way to reverse unfavorable decisions that would otherwise impair their ability not only to accomplish one particular goal but also to accomplish all of the other goals that it is a means to. For example, someone who is counting on a loan in order to purchase a car in order to be able to drive to a well-paying job in order to take care of their family might be denied that loan. In such a case, the denial affects not just their immediate financial situation but their whole life plan. If someone cannot trust that they will have some way of overcoming challenges that thwart the crucial means to their long-term ends, they will have little reason to try to engage in the temporally-extended agency characteristic of mature adults.

We live in a world in which many decisions that significantly affect our ability to exercise temporally-extended agency are made by algorithms and bureaucracies[3]. These algorithms and bureaucracies establish a system of incentives and disincentives that apply to both the ends that people might pursue for their own sake and the means to those ends. If you want to enter a profession, you typically need to receive some sort of certification. If you want to make a large purchase, you may need to take out a loan (that you can pay back at a reasonable interest rate in a reasonable amount of time). If you want to travel internationally, you need to get a passport and potentially also a visa. Across a vast range of sectors, decisions that fundamentally affect people's lives and their ability to engage in long-term planning are made by algorithms and bureaucracies. Sometimes, those decisions are unfavorable. When they are, the subject of the decision can only reasonably plan their subsequent course of action if they know what it would take to receive a more favorable decision. After all, a desired or hoped-for end can only become the target of a plan if the agent is able to select a means to that end. Moreover, this need to be able to plan applies not just to one-off cases, but generally over the course of one's life. As such, someone can be positioned in such a way that they know or reasonably expect that, were things to go wrong, they would be able to set them right again. Such positioning refers not only to the way things currently are but also to how they might be across a range of counterfactual scenarios.

As such, we need some way to ensure that people both have some way of getting unfavorable decisions reversed and know, in general, that they will have a way of getting unfavorable decisions reversed. Let us define the *enjoyment of recourse* as being in such a position. Recourse systematically delivers the benefit of reversing harmful decisions by algorithms and bureaucracies across a range of counterfactual scenarios[4]. If someone enjoys recourse, then not only are they able to get a single decision reversed, but they also enjoy the power to reverse decisions across a range of counterfactual scenarios. As such, someone who enjoys recourse need not passively suffer the slings and arrows of outrageous fortune, but is positioned to take up arms against a sea of troubles. They do not suffer from what Condorcet [3, 11:161, 191] considered one of the most debilitating aspects of poverty: "the idea of being counted for nothing, of being delivered up, without defense, to all vexations and all outrages."

It is illuminating in this context to refer to recent work by Philip Pettit [26], who has argued that a wide range of ethically important values are *modally robust*[5]. For a good to be modally robust in Pettit's sense, it must systematically deliver some other benefit in a range of counterfactual scenarios. For instance, according to Pettit, people value the non-robust good of *favor*, and therefore also value the robust good of *friendship*, which delivers favor in a range of counterfactual scenarios. If someone is your friend, not only do they favor you now, but also they would be disposed to favor you in a range of nearby possible worlds. Friends are disposed to put one another back on course rather than simply abandoning each other when the going gets tough [1], and there are derogatory natural language expressions (e.g., 'fair-weather friend') for people whose favor cannot be counted on in a broad enough range of

[2] https://www.bankrate.com/finance/consumer-index/money-pulse-0117.aspx. Accessed 5 May 2019.

[3] For a critical history of this phenomenon, see [19]

[4] Thus, we distinguish between particular token acts of exercising recourse (reversing a single harmful decision) and the general state of enjoying systematic access to the power to reverse harmful decisions (knowing that if a harmful decision were to be made, one would be able to get it reversed).

[5] In philosophy, robustness of this sort is understood in terms of counterfactual conditionals. Unlike the material conditional, "If p, then q," the truth conditions for the counterfactual conditional, "if p were the case, then q would be the case" refer not only to the world as it actually is but also to various ways the world could be. There are multiple, competing analyses, but the most prominent hold that the conditional is true just in case, in the most "nearby" possible world(s) in which p is true, q is also true (Stalnaker 1968, Lewis 1973). A world counts as "nearby" if it differs only slightly from the actual world. In statistics, robustness is a property of an estimator (a quantity computed from a sample that purports to be an estimate of a population-level property). For example, we might ask what the mean salary of all residents of the United States is. To estimate this quantity, we might sample 1000 people at random and average their salaries. This sample mean is an estimator of the population mean, and we can determine the extent to which the sample mean is a good proxy for the population mean. An estimator can be very sensitive to corruption in the data. For example, the sample mean could deviate arbitrarily from the population mean if even one point of the sample has a corrupted salary entry that is arbitrarily large. A robust estimator is an estimator that is resilient to small amounts of data corruption. For example, if we instead computed the median of the sample rather than its mean, this is a robust estimator for the population median because the median of a collection of numbers does not change significantly if even a fraction of the points are corrupted. Referring back to the idea of a robust good, we can think of a robust estimator as one that is valid in "nearby" worlds where only small amounts of data corruption exist (note that the notion of "near" refers to the number of points that are corrupted, rather than the amount of corruption).

counterfactual scenarios. Beyond friendship, Pettit argues, people value a variety of other robust goods. The virtue of *honesty* is a robust good that delivers the non-robust benefit of truth-telling in a range of counterfactual scenarios. If someone is honest, you can trust them to tell you the truth when they have no incentive to lie, but also to tell you the truth were lying to be to their benefit. Likewise, the robust good of respect delivers the non-robust benefit of non-interference in a range of counterfactual scenarios.

According to Pettit, robust goods are valuable because they are "resilient enough to survive situational shifts" (p. 24), and thus deliver their correlative non-robust goods both "as things actually are" and "as they would be under certain variations" (p. 46). For this reason, when we are assured that someone embodies a robust good, we can live free from anxiety and fear that the correlative non-robust benefit (and everything that depends on it) will suddenly be snatched away without notice or warning. Robust goods thus systematically deliver, as a side-effect, peace of mind and warrant for trust.

While Pettit's account focuses primarily on robust goods as they are embodied in individual humans, it is also possible for a social group or an institution to embody a robust good. For example, a fail-safe nuclear reactor is a complex socio-technical system in which multiple layers of safeguards are put in place. When such a reactor is working as designed, it delivers two non-robust benefits in a range of counterfactual scenarios: namely, electrical power and safety from radiation. If something were to go wrong – either mechanically or via human error – in a fail-safe reactor, multiple alerts and protective actions would be triggered that would (at least if it works as designed) set the reactor on a course towards equilibrium.

In addition, whereas Pettit's account focuses only on robust *goods*, it is possible in similar fashion to define modally robust *ills* as ills that deliver non-robust harms in a range of counterfactual scenarios. For example, malevolence towards someone is a robust ill because it delivers harm in a range of counterfactual scenarios. If someone harbors malevolence towards you, then not only are they going to harm you in the actual world when it is easy for them, but also they will go out of their way to harm you in nearby possible worlds where they face obstacles to harming you. And, just as robust goods can be embodied by both individuals and institutions, so robust ills can be embodied by both individuals and institutions. For example, Kate Manne [23] argues that misogyny is a set of institutionalized social norms and expectations (and related behaviors) that function to enforce patriarchal oppression. Women who deviate from patriarchal norms and expectations are punished by misogynistic actions and emotional reactions, whereas women who conform to such norms and expectations are rewarded.

1.2 A deontological argument for the value of recourse

Pettit's account of robust goods presupposes a consequentialist normative ethics. For those more sympathetic to deontological or Kantian ethics, a closely related argument may be more appealing. In particular, deontologists tend to place great value (indeed, in many cases, supreme value) on human *dignity*. But what is dignity?

For Kant (e.g., *Groundwork* 4:431) and his interpreters (e.g., [35], [20], [15]), dignity is grounded in humanity, which in turn is typically understood as involving two closely-connected capacities. First, humanity involves the capacity to select ends or goals. Second, humanity involves the capacity to be autonomous. It should be clear from the discussion in the previous section that the capacity to select goals is deeply dependent on the sort of long-term, temporally-extended agency that recourse makes possible. What about autonomy? While there are of course disagreements among interpreters about how best to understand Kantian autonomy, according to the *Stanford Encyclopedia of Philosophy*, it involves "the capacity to freely direct, shape, and determine the meaning of one's own life" [7]. Again, this seems to be exactly the sort of long-term planning and self-direction that recourse as we have formulated it helps to secure.

Other philosophers working in the deontological tradition without hewing quite as closely to the orthodox Kantian line offer similar analyses of the meaning and importance of dignity. For example, Griffin [12] contends that the source of human dignity is our capacity to form, revise, and pursue what we take to be worthwhile lives. Likewise, Raz [28] says that "Respecting human dignity entails treating humans as persons capable of planning and plotting their future." And Fuller [10] emphasizes that constraints on human agency thus understood should be understandable to the people on whom they are imposed. One very clear way in which to make the constraints under which someone operates understandable to them is to explain what it would take for the rules (including algorithmic rules) to spit out a different result.

These reflections suggest that – regardless of whether you favor a consequentialist or a deontological normative ethical framework – recourse will turn out to be a fundamental good for anyone who lives in the sort of society that many people currently inhabit. Unless we think that algorithmic and bureaucratic systems – which exist in domains as diverse as criminal justice, credit scoring, hiring, insurance, and voter certification – are or could soon be made infallible, we must place a high value on recourse, on the ability to know what it would take to get a different outcome from high-stakes decisions arrived at by algorithms and bureaucracies.

2 RECOURSE IN COMPUTER SCIENCE

The study of recourse in automated decision making is a relatively recent phenomenon that has till now proceeded without being firmly embedded in a philosophical framework such as Pettit's. The first paper that explicitly addresses it is the work of Ustun et al.[31]. They define recourse as "the ability of a person to change the [harmful] decision of the model through actionable input variables" and then present an algorithm that generates candidate changes of variables that would reverse an algorithm's decision (these sets of variables are called flipsets in their work)[6]. Two aspects of their definition are noteworthy. Firstly, the word 'actionable' is an important part of their definition. By this, they mean that recourse is defined in terms of features that a) can be changed by the individual and b) are (or *should be*[7]) relevant to the decision that they are

[6] Thus, they focus on token acts of exercising recourse, but they are also concerned with the general state of enjoying systematic access to the power to reverse harmful decisions.

[7] More on this ambiguity below.

trying to achieve. For the first aspect, an example is that expecting an individual to change their age or race in order to get a job is not actionable, and therefore would not be considered part of any flipset. The second aspect draws a contrast with other literature that considers the problem of strategic manipulation of features to achieve a desired goal[14, 16, 24]. In such cases, the individual might change features that are (or should be) irrelevant to the decision but that the individual has learned might be important to manipulating the decision-making algorithm. For example, if a tool that analyzes video interviews scores a smiling face highly, an individual might artificially smile in order to get a positive outcome from the interview. Or if a tool that screens the dossiers of job applicants rewards resumes of people named 'Jared', an applicant might misrepresent or legally change their name to get a positive outcome[8].

Recourse is quantified as a "distance from the decision boundary." In the case of the linear classifiers studied by Ustun et al., this is expressed as a sum of (weighted) distances along each dimension (i.e., a weighted l1 distance) or as a sum of quantiles along each dimension (feature). For example, consider a hypothetical example of a linear classifier used to decide whether an applicant should be admitted to a university based on two features: their SAT score and their high-school GPA. GPA has a range from 0-4, and assume that it is considered equally difficult to make an improvement of 0.1 in the GPA as it is to make a 10 point improvement in the SAT score. Then, in order to measure recourse, we would associate a "weight" of 100 with the GPA and a "weight" of 1 with the SAT score. Suppose now that the linear classifier is willing to trade off GPA with the SAT score, and the automated system yields a positive outcome if the following condition is true:

$$0.5 \times \frac{\text{SAT}}{1500} + 0.5 \times \frac{\text{GPA}}{3} \geq 1$$

Informally, this rule captures the idea that an SAT score of 1500 and a GPA of at least 3.0 is sufficient for the system to recommend admission. Consider now a student with a GPA of 3.0 and an SAT score of 1200, as illustrated in Figure 1. A quick calculation reveals that the effort needed to change their SAT score to 1500 (and therefore qualify) is 1500-1200=300, whereas the effort needed to update their GPA to 3.6 (and therefore qualify) is 100(3.6-3.0)=600. Thus, the cheaper way to reverse the algorithm's decision would be for the student retake the SAT and get a score of 1500 or greater. Notice that this calculation incorporates the effort of studying for the SAT again, taking the test another time, and so on into the weights of 100 versus 1. If in fact it was more onerous to retake the SAT, then these weights would need to be modified, possibly indicating a different optimal flipset.

Challenges left open by [31] include computing "distance to the boundary" when the space is not linear (and might be a manifold) as well as dealing with black-box classifiers for which it might be difficult to measure distance to the decision boundary. Recent work by Joshi et al[17] addresses these issues directly, as well integrating causal frameworks into the estimation of recourse. Gupta et al[13] extend recourse calculations to support vector machines as well as looking at the problem of *equalizing* the average recourse with respect to different demographic subgroups.

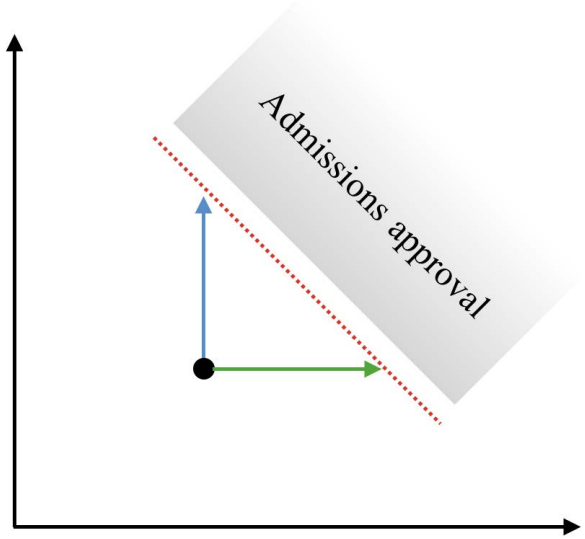

Figure 1: Two flipsets for a student denied admission to university based on their GPA and SAT. The student could either improve their GPA by 0.6 (represented by the vertical blue vector) or improve their SAT by 300 (represented by the horizontal green vector.

Work on recourse builds on earlier research on explanations in machine learning via counterfactuals by Wachter et al.[32]. In that work, the authors argue that, for a number of reasons linked to considerations provided by the European General Data Protection Regulation (GDPR), it would be beneficial to have *counterfactual* explanations of decisions. A counterfactual explanation is of the form "If variable V took different values, the decision would have been different." [9] Wachter et al. argue that such explanations furnish the information necessary to contest decisions as well as exercise recourse, and they suggest ways in which one might produce such a counterfactual from a given model.

Counterfactual explanations have proven to be a popular framework for generating explanations[18, 29]. However, an *explanation* in and of itself need not provide recourse if it is merely providing insight into the decision process of the system. In this paper we focus on recourse directly because of its connection to trust and temporally-extended agency, but many of our critiques and interventions can be applied when appropriate to other forms of explanations. For a detailed exposition of the nature of explanations in decision-making, see the work of Selbst and Barocas [30]. Barocas, Selbst and Raghavan[4] provide a more focused critique of counterfactual explanations.

[8]This example is based on a real case (https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased/). The other main factor that the algorithm rewarded was having played lacrosse in high school.

[9]Philosophers will be familiar with this approach from Kenneth Waters's "difference maker" account of causality[33].

3 CRITIQUING EXTANT APPROACHES TO RECOURSE

In this section, we present a number of criticisms of the existing work on recourse. We begin with more philosophical and conceptual criticisms. As we progress, the criticisms become more technical. Since it is envisioned that recourse should be implemented algorithmically in real socio-technical systems, it is important to address both types of criticisms.

3.1 Conceptually distinguishing appeal from recourse narrowly conceived

Large datasets that algorithms use to make decisions almost always contain errors – sometimes many errors. This means that an algorithmic decision might be unfavorable for two very different reasons. First, the adverse decision could be based on correct and comprehensive data. For instance, someone may have defaulted on multiple loans and therefore be genuinely ineligible for another loan unless they improve their financial situation. This is the type of decision that Ustun et al. seem to have in mind throughout their paper. Second, the adverse decision could be based on faulty or incomplete data. For instance, someone may be recorded as having defaulted on a loan when in fact they repaid it on time and in full, or they be recorded as having no credit history when in fact they have taken out and repaid several loans. In the case of a correct-but-unfavorable decision, the recommended flipset would enumerate the actions that the agent would need to undertake to receive a more favorable decision from the algorithm. Call this *the exercise of recourse narrowly conceived*. In the latter case, the recommended flipset would enumerate rectifications that need to be made to the dataset so that the person will receive the correct (favorable) decision. Call this *appeal*. Recourse narrowly conceived and appeal both promote and protect peace of mind and warrant for trust. Someone who inhabits a socio-technical system rife with errors and no prospect of appeal is just as badly off as (and perhaps worse off than) someone who inhabits an error-free socio-technical system that offers no understandable ways to have unfavorable decisions reversed. Both lack a fundamental robust good and are thus liable to all the stress and anxiety and inability to trust that lack of recourse entails. Thus, while the distinction between appeal and the exercise of recourse narrowly conceived is external to the workings of the mathematical model, it is important to bear in mind when considering how to implement recourse broadly conceived through policy. We flag examples of the distinction below.

3.2 Who decides what's actionable?

As we mentioned in section 2 above, a key feature of recourse is that it is *actionable*, meaning that the recommendations in flipsets should only include the sorts of actions that the individual who receives them might be able to enact of their own volition, and which are relevant (in some unspecified sense) to the decision at hand. For example, in the context of credit scoring, one might plausibly count the number of credit cards someone has as an actionable variable. It is reasonable to expect that someone could increase or decrease the number of credit cards they have (though, of course, not below zero or to infinity). It is also reasonable to allow for variables to be actionable or vary only in one direction like a person's age (which can only increase) and possession of a Ph.D. (which can go from FALSE to TRUE but not conversely). But deciding which features are actionable and which are not can be problematic. For example, consider an individual's current debt load (presented by [31] as an actionable variable). Naturally, it is easy to increase one's debt load. But consider the case of an American with a gigantic student loan. If the interest rate on their loan is high enough, they may never be able to pay it down. Moreover, in the United States it is **illegal** to discharge student debt (alone among all other types of debt) in bankruptcy except in very unusual circumstances[10]. It is therefore unclear whether debt load should be considered fully actionable.

This is just one example, but they can be proliferated. Two further variables considered actionable by [31] are whether someone has a savings account and whether they have a retirement account. Consider the case of an undocumented immigrant. Is such a person in a position to flip the values of these variables from FALSE to TRUE? If not, they are not truly actionable variables, only actionable for some individuals. There are even more complex features, such as race and even gender, where the notion of what is actionable leads to deeper questions about identity and social constructs that are well beyond the scope of this paper. While we do not want to take a stand on these fraught questions here, we assert that it should not be left to algorithm designers to (perhaps unconsciously) build into their models their implicit assumptions about areas far beyond the scope of their training and expertise, especially given that their models might have policy implications. The point of embedding recourse into a socio-technical system is to give people peace of mind, the ability to exercise temporally-extended agency, and warrant for trust. If the design of the system does not involve input from people who fully appreciate the costs imposed by unfavorable decisions and the agency of those to whom flipsets are to be recommended, it will not live up to these essential desiderata.

3.3 Who exercises recourse on behalf of whom?

The framework of recourse is highly individualistic – indeed, we contend, unreasonably so. In this framework, each individual is characterized by a vector of features $\mathbf{x}$ and a binary label y, which takes the value -1 when disfavorable and 1 when favorable. The features are the values assigned to the individual by each column of the dataset that characterizes them. The label indicates, for example, whether they are creditworthy, hirable, etc. Recourse for the individual characterized by $(\mathbf{x}, y)$ is then determined by whether there is an action $\mathbf{a}$ such that $f(\mathbf{x}) = -1$ but $f(\mathbf{x}+\mathbf{a}) = 1$. This mathematical formulation of recourse is silent on a key issue: whether the set $A(\mathbf{x})$ of available actions from which $\mathbf{a}$ is to be selected a set of actions that only the individual herself can perform. In other words, it is assumed that recourse *for* a person is necessarily recourse *by* them. In many low-stakes circumstances, this is a plausible assumption. If the height-scanning app at the amusement park says that I am too short to ride the roller-coaster, I am well-poised to object by insisting that a tape-measure be used to check my actual height. Who better to advocate for me than me? (This would be a case of appeal rather than an exercise of recourse narrowly conceived.)

[10] For a brief discussion of this sorry state of affairs, see https://www.forbes.com/sites/zackfriedman/2019/01/09/student-loans-bankruptcy-discharge/#4fe7c4416d56

That said, many cases in which recourse is essential do not fit this paradigm. Consider a case in which a health insurance company declines to reimburse a medical intervention because it is "non-routine." The patient is in the hospital with a morphine drip that makes them incapable of concentrating for more than a few hours each day. Their family and friends understand that if the "non-routine" label is not reversed to "routine," the patient may face bills that will bankrupt them (another example of appeal rather than of the exercise of recourse narrowly conceived). We suggest that the only normatively acceptable conception of recourse in this scenario is one in which the family and friends are also able to take actions on behalf of the patient. In an individualistic system, someone might reasonably refuse to undergo medical treatment because they are worried that they will be unable to exercise recourse on their own behalf during recovery. Moreover, we note that this sort of case is going to be especially common among children, the elderly, people with disabilities, people who are undocumented, people who are less well-educated, and so on. In order to ensure that the most disadvantaged members of society can sincerely and reasonably expect to be able to get unfavorable decisions, recourse must be conceived of less individualistically and more communally.

3.4 Strategic manipulation versus undoing unfair outcomes

As mentioned earlier, An implicit critique of recourse appears in papers that concern themselves with strategic manipulation of a classifier to reverse an (otherwise-justified) outcome. Examples include the idea that if we revealed the algorithm behind search ranking, entities will manipulate the algorithm to place themselves higher on the list (often called 'search engine optimization' or SEO), or that if we revealed the algorithm behind a credit scoring system, then an individual could inspect the algorithm and improve their attributes strategically to get a good score.

The debate over whether an individual is seeking to exercise recourse or merely strategically manipulating a classifier rests on the answer to an ambiguous question: are the set of attributes being changed *relevant* to success at the task or not? If we construe relevance descriptively, then this question just asks whether changing a particular set of attributes would result in a different decision by the algorithm. By contrast, if we construe relevance normatively, the question asks what attributes *should* make a difference. For example, a risk assessment algorithm trained on historical data is likely to treat the attribute of race as relevant. However, we might agree that race *should not* be taken into account when deciding whether to grant bail. In the ideal case, an algorithm would take into account all and only the attributes that should make a difference. Of course, this is rarely if ever the case. Instead, algorithms almost always use proxies for attributes of interest, and the gap between the attribute of interest and its proxy opens up the possibility of strategic manipulation. If this gap is too large or too easily exploited, we may question whether the algorithm itself is valid. Consider the example of an African-American job-seeker changing their name in order for their CV to be taken seriously, or a female academic job seeker muting any indicators of her gender in order to get a favorable assessment by a hiring committee[11]. One can easily argue that the attributes being changed (e.g., the name) are not relevant to the task (success at the job). However, the classifier treats these attributes as relevant. Because the classifier appears to be blatantly unfair, such manipulation may be justified all-things-considered. In any case, this determination rests on which attributes are considered relevant or irrelevant, which itself can be controversial and touches on deeper arguments about the difference between the world-as-it-is and the world-as-it-should-be.

3.5 Changes in the classifier over time

Classifiers change. As more and more data is incorporated into a model, the rule for classification might change its dependence on input attributes. Suppose at some time t an unfavorable decision is made about an individual. They request and are provided with various flipset options, each of which might involve some investment of time and effort. Later, at time $t' > t$, the individual returns and expects a favorable decision to be made. But now the classifier has changed, and the attributes that would have given the individual success in the original classification will no longer do so. For example, there might be a GPA cutoff to gain entry into a computer science major at a university. A student learns what the cutoff is, determines how much they need to improve their GPA to gain admittance (the flipset), and retakes a few classes to get an improved grade. However, by the time their coursework is complete (say, a year later), the threshold has increased again (for example, due to increased demand) and their new GPA is still not sufficient for entry into the major. Since the goal of recourse is to ensure that people can plan and can reasonably trust the socio-technical system that makes decisions about them, this temporal dimension needs to be taken into account.

3.6 Qualification

Sometimes, perhaps often, the cheapest or easiest way for someone to ensure a more favorable decision from an algorithm is just too expensive or onerous. For example, consider a prospective pilot whose eyesight is not sufficiently acute to be qualified to fly a plane. If the only recourse for such a person is either to undergo a somewhat risky LASIK surgery or eye transplant, that might be too costly/risky to recommend. The recommended flipset for such a person would involve more effort than they are willing or able to put out. Is it wise – or humane – to offer a flipset in a case like this? Prompting someone with an overly burdensome way of reversing an adverse decision suggests that if they do not follow the suggestion, they have only themselves to blame for their situation. We should be reluctant to make people feel guilty for failing to rise to standards that are beyond what could be reasonably expected of them, as this runs directly contrary to the trust that recourse is meant to foster. As Case & Deaton (2015, 2017) have shown, the despair associated with such self-blame has had a measurable impact on both mortality and morbidity in middle-aged, white non-Hispanic Americans, leading to increases in drug overdoses, suicides, and alcoholism. Casually building such standards in to a recourse algorithm is therefore to be avoided.

[11] There are numerous examples of race and gender bias in the academic hiring process. The most recent one is [8] about bias in the physics postdoc process.

3.7 Features are not always jointly actionable

An implicit assumption in formulations of recourse is that if attribute X is actionable and attribute Y is actionable, then X and Y are jointly actionable. We contend that this is not necessarily so. Consider for example, the attributes of spending, debt-payment patterns, and educational attainment. It might be the case that someone who single-mindedly focused on one of these variables would find it actionable, but acting on all three simultaneously may prove challenging or even impossible. For instance, someone could quit their job, take on student loans, and work their way through a master's degree, thus increasing educational attainment at the cost of their finances. Alternatively, the same person could take on overtime hours, work at paying down their debt, and neglect their education. If the flipset recommended to such a person included only combinations of actions in which they increased their educational attainment, decreased their spending, and increased the amount of debt they paid down per month, it might seem actionable without actually being so. As before, the point of recourse is to ensure that people can reasonably expect that, were they to receive an unfavorable decision, they would have some way to get that decision reversed. If the algorithm that recommends flipsets serves up non-compossible combinations of actions, they cannot reasonably expect this.

On a related note, recall that the bureaucratic and algorithmic decisions for which we might seek recourse are ubiquitous. Thus, we cannot treat recourse as something that one might seek in isolation from all the other things happening in one's life. An example from the academic job market should illustrate this. In many disciplines, there is a stark contrast between the desirable CV of someone applying for a job at a research university and the desirable CV of someone applying for a job at a small liberal arts college. Consider the case of someone who completely strikes out on the job market. When they ask interviewing committees from research universities what they could do to improve their chances (i.e., when they ask for a flipset), they are told to publish more and in more prestigious venues. When they ask interviewing committees from small liberal arts colleges what they could do to improve their chances, they are told not to publish any more and instead to build up a longer and stronger track record of undergraduate teaching. Pursuing one flipset makes the other no longer actionable. We expect that this sort of dilemma is liable to crop up not just on the academic job market but in many domains. Treating recourse atomistically obscures and could even exacerbate such problems.

3.8 Diversity of cost functions

A key modeling element of recourse is the cost function governing one's ability to modify one's attributes. There is nothing intrinsic to the definition of recourse that requires that everyone must have use the same cost function – however in practice it is important that the cost function be known, and so it is more practical to assume a fixed cost function. To return to our university admissions example, we might assume that the effort required to improve one's GPA is the same for everyone, and that the effort required to improve one's SAT score is also the same for everyone. This is manifestly untrue, however. Some students may find it much easier to improve their score on a one-off standardized test than others. Likewise, some students may find it easier to improve their GPA than others. For those who find it easier to improve their SAT score, retaking the test should be in their recommended flipset. By contrast, for those who find it easier to improve their GPA, retaking courses should be in their flipset. However, in many – perhaps most – actual cases, the algorithm is not going to know which student is which. In [31], this problem is addressed by allowing the system to return multiple "flipsets" – ways to take action to remedy an unfavorable outcome – as a recognition of this potential lack of knowledge of the true cost function. However, even this approach does not truly engage with the modeling uncertainty in the cost function.

One way to respond to this problem is to recommend the average or the modal shortest-path across the decision threshold or even express costs relatively in terms of percentiles based on the training data (as Ustun et al also do in their work). However, either of those approaches privileges the majority or the plurality at the expense of minorities and outliers. If, as we suggested above, we want to be especially careful not to disadvantage those who are least well off, then these approaches are problematic.

4 REVISING AND IMPLEMENTING RECOURSE

From the perspective of recourse as a modally robust good, the specific algorithmic recourse proposals fall shy in the ways described above. In this section, we articulate a revised approach to recourse that responds to the criticisms canvassed in the previous section. Each of the subsections here addresses its correlative subsection from Section 3 above.

4.1 Disjunctive instructions for flipsets

Because datasets contain errors and omissions, any flipset recommended to an individual might specify a set of attributes that they already satisfy. In such a situation, the individual would need to appeal for rectification of the dataset rather than exercise recourse narrowly conceived. For this reason, flipsets should come with disjunctive instructions. For example:

> We are sorry that you did not receive a favorable outcome. Our model indicates that if your profile were changed in the following way [attributes $X, Y, Z \ldots$], you would receive a favorable outcome. If you believe that your profile should already characterize you in this way, you can appeal the decision by contacting [relevant authority]. Otherwise, you have until [date] to make these changes and then seek to exercise recourse.

This way of framing the flipset recognizes that the dataset may contain errors or omissions.

4.2 Stakeholder and expert panels to establish acceptable action sets.

Recall that it is not appropriate for computer scientists to build their own implicit or explicit assumptions about what is and what is not actionable into a recourse algorithm. Likewise, it is generally not advisable for computer scientists to build their implicit or explicit

assumptions into the construction of training datasets. In both cases, the best practice, which is also more defensible and relieves computer scientists of burdens they are not typically trained to handle, is to engage in systematic consultation with stakeholders and relevant domain experts in the humanities and social sciences. Just as training data needs to be generated carefully and in a way that respects various ethical and epistemic constraints, so action sets for use in a recourse algorithm need to be generated carefully and in a way that respects various ethical and epistemic constraints. Of course, this does not guarantee that the resulting training datasets or action sets are guaranteed to be infallible, but it does respect the rights and local expertise of those who are best positioned to say which attributes are genuinely actionable.

4.3 Fiduciaries

As we explained above, in instances in which someone may need recourse (either via appeal or via the exercise of recourse narrowly construed), they may not be well-positioned to act on their own behalf. This is especially the case when it comes to children, the elderly, people with disabilities, people who are undocumented, people who are less well-educated, and others at the margins of society. To ensure that these people have adequate access to recourse, we recommend that any socio-technical system in which recourse is embedded make a role for fiduciaries who are charged to act on behalf of those they represent. A fiduciary may be a family member, someone with the power of attorney, or a representative appointed by a court or other body. In this way, we relax the extreme individualism implicitly assumed by Ustun et al. so that recourse for x can be recourse exercised by y or z or w on behalf of x. Recall that their framework for providing recourse includes a cost for updating any individual feature, and that this is specific to the individual requesting recourse. Setting aside the issue of how to set the costs for a given individual, there is no reason the system cannot use costs associated with the fiduciary rather than with the harmed party.

4.4 Auditing the gap between normative and proxy attribute sets

As we explained above, there are two senses in which an attribute or set of attributes might be relevant to an algorithmic decision. Descriptively, an attribute is relevant when it makes a difference to the classification. Normatively, an attribute is relevant when it is the sort of thing that should make a difference to the classification. Because normative attributes are typically difficult or impossible to measure directly, datasets tend to use descriptive proxies. The gap between a normative attribute and its proxy can lead to misclassification in both directions: favorable when the correct classification is unfavorable, and unfavorable when the correct classification is favorable. There is no automatic way to handle this problem. Instead, proxies need to be chosen with care, and the gap [9] between normative and proxy attribute sets must be audited on a regular basis. The same panel of stakeholders and domain experts envisioned in section 5.1 would probably be well-positioned to do such auditing.

4.5 Handling change over time via *ex post facto* and *lex mitior*

Recall that classifiers change. In the time between when a flipset is recommended and when its criteria are fulfilled, a classifier may have changed in such a way that the individual who received the recommendation no longer qualifies. This is frustrating and makes it difficult to plan. Furthermore, such scenarios are likely to crop up more and more as algorithms become embedded in a wide range of mundane decision making processes. If recourse is to be employed in such contexts, it needs to suggest flipsets that have indefinite temporal stability or clear expiration dates. One option would be to follow the precedent set in United States law, according to which *ex post facto* laws[12] that disadvantage the accused are not binding[22]. Translated to the context of algorithms and recourse, this would mean that if someone receives a flipset recommendation, it should be valid forever or come with an expiration date. Then, if the person satisfies the criteria in the flipset (before the expiration date), they automatically gain whatever benefit they were pursuing even if the rules for others have changed in the meantime.

An alternative option would be to follow the precedent set in European law: namely, *lex mitior* (or "milder law") [34], according to which the milder rule is the one that applies. In the case of the GPA cutoff sketched above, if the admission criteria become more strict, then anyone who was recommended a flipset prior to the change should be eligible under the old rules. By contrast, if the admission criteria become less strict, then the new standards should be used for everyone. In cases where there is no well-defined ordering of strictness, individuals should be able to choose for themselves whether they are judged by the old or new criteria. Implementing either of these options assumes that the flipset is recommended by those with enough institutional power to ensure that either the *ex post facto* or *lex mitior* rule is applied. If, instead, the flipset is recommended by a third party, such as a consulting firm or independent coach, the flipset should at least be accompanied by a very clear warning that the rules may change.

4.6 Introducing an upper bound on costs

As we pointed out above, sometimes even the cheapest or least onerous way to reverse an unfavorable decision is to expensive or burdensome. To handle this problem, we suggest introducing an upper bound to the cost threshold when calculating a flipset. If even the cheapest option exceeds this threshold, the individual would receive a recommendation like the following:

> Sorry, but in your case recourse does not seem to be a feasible option because even the least burdensome avenue to reversing the decision is probably too onerous for you to pursue. In particular, the easiest way for you to receive a favorable decision is <*details of recourse*>. You are of course welcome to pursue this option, but please do bear in mind that it may be excessively costly.

In this way, the individual would still see their best flipset (which, if all they need to do is appeal rather than exercise recourse narrowly

[12] That is, laws that take effect retroactively.

conceived, may in fact be feasible), but the flipset would not be provided as a recommendation so much as an explanation of why it may not be worth pursuing recourse.

One interesting advantage of using upper bounds on costs is that we can count the number (or proportion) of times this upper bound on cost is reached. A large value would indicate systemic society-wide problems in the application of recourse and indicate a need for deeper structural reforms[13]

4.7 Changing the geometry of the intervention space.

The problem with features that might vary jointly is that the assumption of independence – that features can be modified separately and without incurring extra cost – is broken. A purely mathematical approach to addressing this is to modify the underlying geometry in which the distance associated with recourse is being calculated. In particular, dependency among attributes corresponding to moving along a submanifold of the underlying vector space – informally, a curved surface in the space, rather than the entirety of the space itself. From a purely technical perspective, the challenge is then to compute (shortest) distances in this submanifold to the boundary, which while difficult to do within the existing frameworks is still amenable to analysis via methods for manipulating manifold geometry that are common in machine learning. Indeed, recent articles seeks to do precisely this via estimating the submanifold via data sampling[17].

The bigger challenge is how to learn and encode this submanifold. Doing so requires a deeper knowledge of the ways in which different features interact, as well as the relative costs associated with modifying dependent features. This places a greater burden on the modelers of the system, reflecting a tradeoff between how well the recourse calculation reflects the real world and the ease of use of the system. Note that an alternate modeling tradeoff is to merely *assert* that features are actionable separately in order to prioritize explainability of the method of recourse over effectiveness. Again, we point out that [17] seeks to address this via the explicit use of causal models.

4.8 Handling diversity of cost functions via personalization

Because different people have different cost functions associated with the attributes on which a classifier operates, a one-size-fits-all approach to constructing flipsets may not be advisable. In cases in which a great deal is known about the individual for whom a recommendation is to be made, using their personal cost function should make it possible to offer a bespoke flipset. Such personalization may be difficult to achieve, however, or may infringe on privacy rights. If the recommendation is not personalized, as current notions of recourse are, then it might be better to recommend several different options.

5 CONCLUSIONS

In this paper, we used the fact that temporally-extended agency and trust are fundamental to human flourishing to argue for the importance of recourse. We understand recourse as the modally robust good that delivers the correlative non-robust good of reversing unfavorable decisions across a range of counterfactual scenarios. Someone who enjoys recourse and knows that they enjoy it is better positioned to engage in long-term planning and to trust that, if something were to go wrong on the way towards their long-term goals, they would not fall victim to precarity but rather be able to set themselves right again. We then pointed out that many of the decisions for which people might want recourse are made by algorithms and bureaucracies. This in turn suggests that these socio-technical systems should be implemented in a way that automatically delivers suggested flipsets to people about who unfavorable decisions are made. In the remaining two sections of the paper, we raised some problems for the existing literature on algorithmic recourse, then addressed these problems to the extent possible.

In closing, we want to address an objection that we expect many readers may have: namely, that recourse is a red herring that will only distract us from the more fundamental problems of precarity. After all, one might think that if precarity were (much) less severe than it currently is, there would be little need for recourse. Moreover, one might worry that focusing on recourse places yet an additional burden on individuals as they navigate complex and challenging socio-technical systems. Wouldn't it be better to focus on systemic solutions that don't require individual attention and intervention? Are we suggesting that if you had an opportunity to take advantage of recourse and neglected to do so, you only have yourself to blame? What if (as seems likely) disadvantaged people are also those who most frequently need to resort to recourse, yet have the least bandwidth – in terms of time, energy, money, and social connections – to do so?

We take these concerns seriously. However, we do not see why the matter should be conceived as a binary choice: either alleviate underlying precarity or treat its symptoms via recourse. We would like to think that academics concerned about social justice can, as it were, walk and chew gum at the same time.

But there is a deeper point to be made in this connection. Recourse is an already-extant good. We are not proposing to introduce it to a system where it is absent. Instead, we are proposing to ameliorate its distribution in a system where it is very much present. Indeed, the availability and effectiveness of recourse in contemporary societies is itself unequally distributed in much the same way that primary goods are. If you are wealthy, you can pay a lawyer to get a parking ticket expunged. If you are powerful, you can use backchannels to get your failson accepted to a prestigious university on a sham athletic scholarship. If you are well-connected, you can call in favors to land yourself a job that would otherwise have gone to someone else. These are all cases in which recourse is already being employed. Our aim is to establish a more adequate conceptualization and operationalization of the phenomenon, not to introduce it into a system where it does not already exist.

[13] We thank the anonymous reviewer for this suggestion.

REFERENCES

[1] Mark Alfano and Joshua August Skorburg. 2016. The embedded and extended character hypotheses. In *The Routledge Handbook of Philosophy of the Social Mind.* Routledge, 481–494.
[2] Elizabeth Anderson. 2019. *Private government: How employers rule our lives (and why we don't talk about it).* Vol. 44. Princeton University Press.
[3] François Arago et al. 1847. *Oeuvres de Condorcet.* Vol. 4. Firmin Didot frères.

[4] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. In *The Conference on Fairness, Accountability and Transparency (FAT*)*. ACM. https://doi.org/10.1145/3351095.3372830.

[5] Judith Butler. 2006. *Precarious life: The powers of mourning and violence.* verso.

[6] Jon A Carnegie. 2007. *DriverâĂŹs license suspensions, Impacts and fairness study.* Technical Report.

[7] Robin Dillon. 2018. Respect. *Stanford Encyclopedia of Philosophy* (2018).

[8] Asia A Eaton, Jessica F Saunders, Ryan K Jacobson, and Keon West. 2019. How Gender and Race Stereotypes Impact the Advancement of Scholars in STEM: ProfessorsâĂŹ Biased Evaluations of Physics and Biology Post-Doctoral Candidates. *Sex Roles* (2019), 1–15.

[9] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[10] Lon Fuller. 1964. *The morality of law.* Yale University Press.

[11] Duncan Gallie, Serge Paugam, and Union européenne. Direction générale Emploi et affaires sociales. 2003. *Social precarity and social integration.* Vol. 56. Office for official publications of the European communities Luxembourg.

[12] James Griffin. 2008. *On human rights.* University of Oxford Press.

[13] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing Recourse across Groups. *arXiv preprint arXiv:1909.03166* (2019).

[14] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. ACM, 111–122.

[15] T.E. Hill. 2000. *Respect, pluralism, and justice: Kantian perspectives.* Oxford University Press.

[16] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 259–268.

[17] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *arXiv preprint arXiv:1907.09615* (2019).

[18] Amir-Hossein Karimi, Gilles Barthe, Borja Belle, and Isabel Valera. 2019. Model-Agnostic Counterfactual Explanations for Consequential Decisions. *arXiv preprint arXiv:1905.11190* (2019).

[19] Colin Koopman. 2019. *How we became our data: A genealogy of the informational person.* University of Chicago Press.

[20] Christine Korsgaard. 1996. *Creating the kingdom of ends.* Cambridge University Press.

[21] Wayne Lewchuk, Michelynn Lafleche, Diane Dyson, Luin Goldring, Alan Meisner, Stephanie Procyk, Dan Rosen, John Shields, Peter Viducis, and Sam Vrankulj. 2013. ItâĂŹs more than poverty: Employment precarity and household well-being. *Toronto, ON: Poverty and Employment Precarity in Southern Ontario* (2013).

[22] Wayne A Logan. 1997. The ex post facto clause and the jurisprudence of punishment. *Am. Crim. L. Rev.* 35 (1997), 1261.

[23] Kate Manne. 2017. *Down girl: The logic of misogyny.* Oxford University Press.

[24] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. The Social Cost of Strategic Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 230–239. https://doi.org/10.1145/3287560.3287576

[25] Board of Governors of the Federal Reserve System. 2016. Report on the economic well-being of U.S. households in 2015. https://www.federalreserve.gov/econresdata/2016-economic-well-being-of-us-households-in-2015-executive-summary.htm. Accessed 5 July 2019.

[26] Philip Pettit. 2015. *The robust demands of the good: Ethics with attachment, virtue, and respect.* OUP Oxford.

[27] John Rawls. 2005. *Political liberalism.* Columbia University Press.

[28] Joseph Raz. 1977. *The authority of law.* University of Oxford Press.

[29] Chris Russell. 2019. Efficient search for diverse coherent explanations. *arXiv preprint arXiv:1901.04909* (2019).

[30] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.

[31] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 10–19.

[32] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.* 31 (2017), 841.

[33] C Kenneth Waters. 2007. Causes that make a difference. *The Journal of Philosophy* 104, 11 (2007), 551–579.

[34] Peter Westen. 2015. Lex Mitior: Converse of ex post facto and window into criminal desert. *New Criminal Law Review: In International and Interdisciplinary Journal* 18, 2 (2015), 167–213.

[35] A.W. Wood. 1999. *Kant's ethical thought.* Cambridge University Press.