From Explanation to Recommendation

From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse

Emily Sullivan*
e.e.sullivan@tue.nl
Eindhoven Artificial Intelligence Systems Institute
Eindhoven University of Technology
Eindhoven, The Netherlands

Philippe Verreault-Julien*
p.verreault-julien@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

ABSTRACT

People are increasingly subject to algorithmic decisions, and it is generally agreed that end-users should be provided an explanation or rationale for these decisions. There are different purposes that explanations can have, such as increasing user trust in the system or allowing users to contest the decision. One specific purpose that is gaining more traction is *algorithmic recourse*. We first propose that recourse should be viewed as a recommendation problem, not an explanation problem. Then, we argue that the capability approach provides plausible and fruitful ethical standards for recourse. We illustrate by considering the case of diversity constraints on algorithmic recourse. Finally, we discuss the significance and implications of adopting the capability approach for algorithmic recourse research.

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics** → *User characteristics*; • **Human-centered computing** → *Social recommendation.*

KEYWORDS

algorithmic recourse, recommendations, capability approach, diversity, explainable AI, counterfactuals

ACM Reference Format:
Emily Sullivan and Philippe Verreault-Julien. 2022. From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3514094.3534185

1 INTRODUCTION

There is widespread agreement that providing explanations for model decisions is important, especially for end-users. Such explanations can help users gain trust in an otherwise opaque system. Explanations can also spur user engagement on product-based platforms. However, there is no one-size-fits-all box for successful explanations. Explanatory norms differ depending on the stakeholder, the domain, and the specific goals a user has [49, 61]. One specific explanatory norm that is gaining more and more traction is *algorithmic recourse* [e.g. 20, 22, 23, 25, 27, 35, 52, 55].

Algorithmic recourse was borne out of counterfactual explanation methods. Wachter et al. [58] highlight three uses for counterfactual explanation: i) answer why a certain decision was reached, ii) provide the user with grounds to contest the decision, and iii) provide the user with actionable changes to reverse the decision. While Wachter et al. argue that counterfactual explanation can satisfy all three, recent work suggests otherwise [e.g. 40]. Models can make decisions based on immutable features, which may satisfy (i) and (ii), while failing to satisfy (iii). Since algorithmic recourse is concerned with the specific project of providing users with an *actionable* counterfactual explanation, immutable features prevent users from getting feasible and actionable advice on what changes they could implement to get a new decision.

There are clear benefits from the user's perspective for recourse and some have argued for its ethical value [55]. Recourse seems especially important in domains where algorithmic systems are part of decision pipelines that greatly affect people's lives, such as granting a loan, sentencing decisions in a judicial system context, college admissions and more. Nevertheless, as Venkatasubramanian and Alfano [55] discuss, algorithmic recourse faces pitfalls. The important work on fairly defining cost, distance, etc. is necessary. However, shared (ethical) standards for constraining recourse counterfactuals in particular directions are conspicuously absent, with papers approaching the problem in different ways. Some focus on the desiderata of proximity [e.g. 58], while others highlight the need for sparsity [e.g. 15] or for user input for specific feature constraints [e.g. 55], and others emphasize the need for diversity [e.g. 31].

While we do not provide an all things considered ethical argument that algorithmic recourse is the best way to approach the problems of opaque systems that make highly impactful decisions, we seek to make progress on how to best constrain algorithmic recourse—assuming recourse is desirable—by providing an ethical framework that helps design recourse recommendations. Accordingly, proposing ethical standards for recourse does not imply letting designers and suppliers of artificial intelligence systems off the hook. Algorithmic decisions do not become exempt of other ethical standards because of the presence of recourse. This work makes three contributions:

(1) Recasting algorithmic recourse as a *recommendation* problem, not an explanation problem. Taking recourse seriously as a recommendation problem allows us to utilize insights from research programs on recommendation systems, which are

*Both authors contributed equally to this research.

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534185

largely siloed from questions in explainable AI. Moreover, it separates two distinct desiderata for algorithmic recourse: methods of generating or extracting counterfactuals and how to *explain* counterfactual information to users. Once we solve which recommendations are necessary for recourse, then we can ask the explanatory question about how to best explain these recommendations to users. It may turn out through user studies that providing recourse recommendations is more successful through a different explanatory framework besides counterfactuals.

(2) Providing ethical standards (via the capability approach) that can guide research on how best to constrain algorithmic recourse toward feasibility and the well-being of users.

(3) As a case study, we use the capability approach as grounding the value of diversity for recourse recommendations. We highlight gaps in current research and suggest paths forward by taking inspiration from the role of diversity in recommendation systems.

We hope that this work contributes to establishing plausible and fruitful ethical standards for recourse recommendations.

Section 2 argues that recourse should be viewed as a recommendation problem, not an explanation problem. In section 3 we introduce the capability approach and make the case for its descriptive and normative adequacy. Section 4 looks at diversity constraints on recommendations to illustrate the usefulness of the capability approach and viewing recourse as a recommendation problem. We discuss several topics of potential significance for recourse research in section 5.

2 ALGORITHMIC RECOURSE: FROM EXPLANATION TO RECOMMENDATIONS

2.1 Recourse as an explanation problem

People are increasingly subject to algorithmic decisions, with an increased use of 'black-box' models. This presents a challenge and need for explainability. Explainable AI can increase users' trust in the system, aid developers in building more robust and reliable models, and more. Moreover, regulations like the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AIA) discuss the importance of end-users receiving an explanation or rationale for decisions involved in algorithmic processing. This has spurred a flurry of development of different methods and approaches to explaining black-box models.

One explanatory approach that has gained significant traction is counterfactual explanation (CE). CEs provide answers to *what-if-things-had-been-different* questions. The claim is that understanding the modal space of a model can serve as a way to explain and provide understanding of the model's decision boundary. One of the benefits of CE is that building a proxy model, that is necessary for other feature importance methods, need not be necessary [31]. Instead, CEs probe the black-box model by changing various inputs to see what changes would lead to a change in the output.

As we have seen, Wachter et al. [58] highlight three uses for CEs. Ethicists and those interested in algorithmic fairness have especially latched onto (iii)—how CEs can provide users with actionable advice to reverse the outcome—known now as *algorithmic recourse*. Ustun et al. [52, p. 10, emphasis in original] define algorithmic recourse "as the ability of a person to change the decision of the model through *actionable* input variables [...]".

Since recourse was borne out of CE, recourse itself has been understood as a type of explanation method, especially salient in domains where algorithmic systems are part of decision pipelines that greatly affect people's lives. In these contexts, when users are given a negative or unfavorable decision, advice on how to get a different result in the future is top of someone's mind. Thus, a recourse explanation seems most suitable.

While explanations can serve a number of different goals, like transparency and trust [28, 49, 51], explanation first and foremost has epistemic aims, like filling knowledge gaps and enabling understanding [12, 16]. As such, most works look at recourse through the lens of an explanation problem, where the evaluative goals center around the epistemic goals of explanation, such as understanding the model and its decision boundary [58]. For example, Ustun et al. [52] describe recourse as a type of actionable CE. Mothilal et al. [31] evaluate their method of generating recourse counterfactuals with other XAI methods, specifically LIME [36], to show that recourse explanations can provide users with understanding of the decision boundary.

However, we propose that conceptualizing recourse as an explanation problem is ill-suited. As we explain in the next section, the goals of explanation are distinct from the goals of providing users with actionable information. While in some cases the same counterfactual can explain *and* provide actionable information to reverse a decision, it is not by virtue of the counterfactual's *explainability* that it provides actionable information. Instead, we propose that algorithmic recourse is best understood as a *recommendation* problem and that doing so has the promise of improving metrics and methods for algorithmic recourse.

2.2 Recourse as a recommendation problem

CE methods generate counterfactuals by making small changes to input variables that result in a different decision. Counterfactual generation serves as an explanation method because finding the smallest changes that would flip a decision tells us important information regarding how a model made its decision [48]. However, sometimes counterfactuals involve changing features that are immutable, or mutable but non-actionable [22]. Immutable features are those that *cannot* change, for instance someone's race. Mutable features can change, but not because of a direct intervention on them. Someone's credit score may change as a result of debt repayments, but it is not possible for someone to intervene on her credit score. For this and other reasons, the goals of explanation *simpliciter* can come apart from the goals of actionable information important for algorithmic recourse. In this section, we discuss that explanation is possible without recourse and that recourse is possible without explanation, indicating that recourse is better understood as a recommendation problem.

2.2.1 Explanation without recourse. The first reason why algorithmic recourse is ill-suited to be an explanation problem is that CE is possible without recourse [52, 55]. Consider the difference between the following counterfactual explanations for a loan decision discussed above: "If you had less debt, then the loan would have been approved," versus "if you were younger, then the loan would have

been approved." The former CE gives the end-user recourse, while the latter does not. It is not actionable advice for someone to become younger, though it is actionable advice for someone to pay off some of their debt. Moreover, in criminal justice cases, using a simplified model based on COMPAS data [4, 11], CE methods found that race is often one of the more common features that would reverse a risk categorization [31]. But again, since race is immutable, it cannot be a recourse explanation but *is* an explanation of the model's decision. Along these lines, Karimi et al. [21] make a distinction between contrastive explanations and consequential recommendations, the latter being a subset of the former. The idea is that recommendation requires information on the causal relationship between inputs, while explanation just requires information regarding the relationship between the model and its inputs. If recourse requires a consequential recommendation—which Karimi et al. [21] argue is the case—then again explanation is possible without recourse, especially since the causal relationship between inputs involves a heavier burden to satisfy (more on causation in section 5).

2.2.2 Recourse without explanation. Even though most works discuss that a CE need not entail recourse, recourse can still be first and foremost an explanation problem. Recourse could be understood as a specific *type of explanation* that is actionable [21, 52]. However, a less appreciated distinction is that it is possible to have a recourse counterfactual that fails to be an *explanation*.

Barocas et al. [5] highlight a notable difference between principle-reasons explanations and recourse explanations. The former provide the data-subject with information regarding which features serve as a justification or rationale against the decision, while recourse explanations provide helpful advice *without* the decision subject learning about the features that were "crucial marks against" them. Recourse serves a practical purpose of giving decision subjects guidance for the future. Thus, having the most salient explanation that can answer *why* a model made its decision—or the rational for the decision—can come apart from providing users with *recommendations* on how to reverse the decision. Consider again the example of a recidivism classifier or loan decision algorithm as discussed above. It very well might be that the immutable factors were the more discerning factor for the decision. In this case, a recourse 'explanation' focusing on actionable factors becomes epistemically misleading since the most discerning reason for the model's decision is hidden. The user does not have access to the central difference-makers of the model's decision, and thus would fail to really understand the model.

Conceptualizing recourse as a type of explanation can also mask bias. Explanation methods are used for auditing the fairness of models [28], with one central source of bias resulting from models using immutable features in a problematic way. Since recourse disregards counterfactuals that involve immutable features, recourse has the potential to mask bias and be epistemically misleading.

2.2.3 Recourse as recommendation. The chief goals of model explanation center around providing users with understanding the rationale of the model's decisions. Recommendation systems, on the other hand, have a different primary goal. They seek to help users with selecting a subset of items that are among an ever-growing list of possible items by creating user profiles that are continuously updated to aid in filtering the most relevant items for users. As such, recommendation systems explore a specific relationship between a user and the model that is not mirrored in more traditional explainability questions regarding why a black-box model made a decision.

The difference between recommendations and explanations can be subtle in some contexts. Often recommendation systems also provide explanations to users as to *why* they are seeing the recommendations that they do. However, the recommendations and the explanations of recommendations are distinct. Our proposal is that algorithmic recourse stands to benefit from such a distinction. The purpose of generating the list of actionable advice is distinct from explaining this advice and explaining the model's decision boundary.

The relationship between recourse and recommendations has not gone unnoticed. There has been work that takes insights from algorithmic recourse to improve recommendation systems [10]. And those working on recourse make the explicit connection that recourse is similar to recommendation systems [31]. However, Mothilal et al. [31] stop short of casting the goals of recourse to be recommendation goals, since they evaluate their recourse model as if it was an explanation problem, as discussed above. Karimi et al. [21] distinguish between two types of questions for recourse. (Q1) explanatory questions, like "why was I rejected for the loan?", and (Q2) recommendation questions, like "What can I do to get the loan in the future?", where answers to Q2 questions provide "consequential recommendations." However, this terminology aims to point out a difference in causal presuppositions needed for counterfactual generation. They do not explicitly reconceptualize recourse as dealing with the class of problems found in the recommendation systems literature.

Our contribution is to explicitly conceptualize recourse as a *recommendation* problem akin to those problems facing recommendation systems and not as an explanation problem. The unique feature of algorithmic recourse is not explanation, but rather giving advice and finding a subset list of actions from a large possible subset of actions (i.e. recommending). It is our contention that shifting the dialectic away from algorithmic recourse as an explanation problem to a recommendation problem will improve recourse recommendations as well as help to make sure that algorithmic recourse is not used in ethically or epistemically misleading ways. It shifts the focus away from explainability to a more user-modelling perspective regarding the interplay between user-preferences and capabilities and the model.

Once we solve which recommendations users should have such that recourse is possible, then we can ask the question how best to explain or convey this information to users. This may be through counterfactuals, or it may turn out through user studies that providing recourse recommendations is more successful through a different explanatory framework. An added benefit of considering recourse as a recommendation problem is that it allows us to utilize insights from a rich research program in recommendation systems that is still largely siloed from questions in XAI. Moreover, conceptualizing recourse as a recommendation problem allows us to utilize particular ethical tools—like the capability approach—to guide research in filtering counterfactuals that respond well to users' capabilities even if they are far removed from the model's decision boundary.

3 ETHICAL STANDARDS FOR RECOMMENDATIONS: THE CAPABILITY APPROACH

3.1 The ethical standards of recommendations

In theory, recourse has ethical appeal through purportedly promoting agency and autonomy. Venkatasubramanian and Alfano [55] provide some general ethical standards for algorithmic recourse by arguing that it is a modally robust good [see 34]. Robust goods deliver benefits in a range of actual and counterfactual circumstances. For example, the robust good of honesty provides the benefit of truth-telling not only on one specific occasion, but on many occasions. According to this view, we value robust goods because they deliver benefits in various circumstances.

Venkatasubramanian and Alfano hold that someone who has recourse enjoys a capacity to obtain decisions across a range of circumstances and not in a coincidental or piece-meal fashion. That person can reasonably expect that she will be able to obtain a decision and will not be subject to other people's discretionary power or to changing situations. This is crucial for exercising what Venkatasubramanian and Alfano call 'temporally-extended agency', namely the capacity to pursue long-term plans. This sort of agency is important because algorithmic decisions are often a means among a chain. A person seeking a loan to buy a car, they say, may do so in order to take a well-paying job which itself is a means to care for her family. The implications of being denied a loan are thus more far-reaching than simply not being able to obtain the immediate goods or services the loan is for.

While Venkatasubramanian and Alfano provide both consequential (Pettit's framework) and deontological (based on human dignity) reasons to value recourse, how these foundations relate to specific constraints on recommendations and how they may help comparing them remains unclear. They discuss a variety of issues, for instance changes to classifiers over time, and importantly convey that these issues need to be resolved for algorithmic recourse to live up to its ethical promise. Other works on recourse have differed in their approach to the evaluation of constraints, picking and choosing which are necessary or interesting for their specific study, with some of the above concerns in mind.[1] However, no principled ethical framework is currently guiding the design of recourse recommendations. In order to make progress on algorithmic recourse, we need to make progress on delineating which reasons may justify adopting some constraints over others. We need ethical standards that can do this work. We propose that that the capability approach provides such plausible and fruitful standards. First, we introduce the capability approach and then illustrate its relevance by considering one particular constraint: diversity (section 4). In section 5, we discuss the more general significance of the capability approach for recourse research.

3.2 The capability approach

The capability approach, initially developed by Amartya Sen [43–45; see also 32, 38], is a normative framework which characterizes the normative space of evaluation in terms of *functionings* and *capabilities*. According to the capability approach, we should make interpersonal comparisons or assess states of affairs on the basis of these two core concepts. Functionings are 'beings'—ways of being, like being healthy or educated—and 'doings'—activities, like coding or cycling —people may be or undertake. Having an appropriate set of functionings is "constitutive of human life" [38, p. 39]; what makes up and gives value to human life are the 'beings' and 'doings' people achieve. Capabilities are the real freedoms, or opportunities, people have to achieve functionings. Here, 'real' underlines that having a capability goes beyond having a merely formal possibility. It requires having the resources (broadly construed, e.g. income, credentials, social network, etc.) to effectively achieve chosen functionings. Another important claim of the capability approach is that the capabilities people have depend on *conversion factors*, namely the differential capacity to convert resources into functionings. With equal resources, different people will not always have the same capabilities. Other things being equal, a person who suffers from depression will need more resources to achieve the same level of motivation as someone without depression. Conversion factors can be personal (e.g. a disability), social (e.g. being discriminated), or environmental (e.g. the climate) and can be intertwined. Acknowledging conversion factors is important for ethical evaluation because it urges caution in equating resources with well-being.

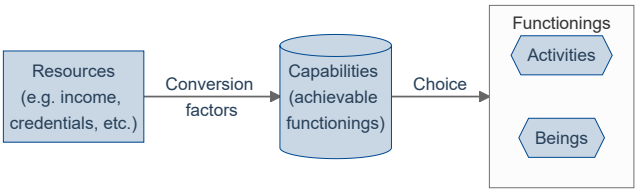

Figure 1: Schematic overview of the capability approach. Resources are converted into capabilities and people choose which functionings to realize from their set of capabilities.

The notion of capability aims to distinguish between what is *actually realized* (functionings) versus what *could effectively* be realized (capabilities) if people wanted to. As figure 1 illustrates, resources are converted into capabilities, effectively possible but unrealized functionings. From that capability set, a person then chooses which functionings to actually achieve. For instance, someone may have the capability to cycle, yet never do it. That person may opt for moving about using public transportation. Again, what matters is the real freedom people have to achieve a combination of functionings.

A capability set is the set of alternative functionings people can achieve. For instance, let us consider the capabilities to be healthy, educated, mobile, sheltered, and participate in politics (see figure 2). Different people may have different capability sets, due e.g. to conversion factors, and thus have a differential real freedom to achieve the related functionings. For instance, Person A might have a greater capability for health than B, but B might be advantaged in terms of education, perhaps because of the social environment. The capability approach holds that interpersonal comparisons should be made in terms of capabilities and functionings.

While figure 2 represents a 'static' capability set, in reality there are often trade-offs between capabilities. As figure 3 shows, having

[1] For a survey, see Karimi et al. [21].

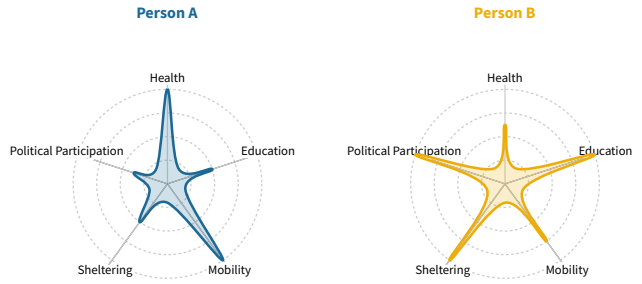

Figure 2: Interpersonal comparison of capabilities. The capability approach holds that we should compare people's advantage in terms of the capabilities and functionings they have. Figure made with Flourish.

more of one capability may sometimes have positive or negative effects on the capability set. Using more resources in order to gain an increased capability in terms of education might have a negative effect on the capability for health, which in turn might reduce one's mobility. More education, however, might contribute positively to political participation. People face similar trade-offs all the time when considering the real opportunities they have. Some could become a scientist or a rock musician, but achieving both is not always effectively possible.[2]

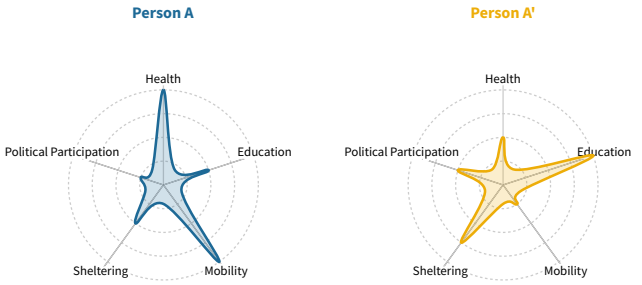

Figure 3: Trade-offs between capabilities. Resource allocation for one capability can have an influence on other capabilities. Figure made with Flourish.

Capabilities help capture the idea that the freedom to achieve certain beings and doings is of utmost moral value. A person's well-being is constituted by what is ultimately good for that person. As Sen [46, 231] notes, any ethical or political theory must select an 'informational basis', viz. features of the world that help to assess well-being and injustice. The capability approach contrasts with alternative theoretical frameworks by submitting that these features are the capabilities people have reason to value instead of, for instance, pleasure or resources. This broadens the informational basis insofar as information about resources or rights can be legitimately used to compare well-being. How to determine the relevant capabilities for the purpose of normative assessment is context-dependent. It can be used for assessing individual well-being, evaluating social states of affairs, or policy-making [39]. It is an influential framework that has been used in fields such as human development [13], poverty [3], mental health [47], technology [33, 62], or education [59]. One famous use of the capability approach is within the United Nations Development Programme's *Human Development Reports*, in particular the Human Development Index.[3] For the purpose of assessing and comparing human development between countries, using indicators such as life expectancy or the level of education may target adequate capabilities. But for assessing whether older people have mobility through public transport, looking at residential density and physical functional capacity would be more relevant [41].

3.3 Recommendations and the capability approach

The capability approach provides plausible and fruitful ethical standards for recourse recommendations because it is descriptively and normatively adequate.

3.3.1 Descriptive adequacy. The capability approach is descriptively adequate because it captures the relevant features of recourse recommendations. Current formulations of recourse have natural analogues within the capability approach. Recourse can readily be understood as a functioning; it is the *activity* of obtaining a decision from a model. When someone obtains a decision, that person achieves the functioning of recourse. But recourse is also viewed as an 'ability' or as something that a person has the 'capacity' to do irrespective of whether they actually achieve it or not. As such, recourse is also a capability; it amounts to the real freedom to obtain a decision from a model. When someone has recourse, that person *would be able* to obtain a decision would she choose to do so. Viewing recourse as a capability also explains the widespread emphasis on actionability. Recommendations are those that users could in principle, but not necessarily, achieve.

Although the notion of capability captures usage of recourse in the computer science literature, it also stresses one underrated feature of recourse, namely its connection to *freedom*. Capabilities are a type of freedom, in particular *option-freedom* [see 38, pp. 102ff.]. Options are what an agent can achieve or realize. The freedom of options depends on two aspects: 1) the agent's access to the options and 2) the options themselves. Some people may face more obstacles (e.g. different conversion factors) than others to realize certain options, resulting in different access to options (1). Option-freedom also depends on the number or quality of options available (2). A person with more options has more option-freedom than a person with fewer options.

For the purpose of recourse, recommendations (should) aim to give option-freedom. In fact, viewing recommendations as seeking to promote option-freedom helps understand the aims of different recourse methods. Some emphasize the importance of causal possibility [e.g. 22] and thus that people should have the proper access to options (see sec 5.1 below for a critique). Others draw attention to the options themselves by generating a large quantity of options users can choose from [e.g. 31]. Adopting the capability approach

[2] One notable exception is Brian May, guitarist of the famous band Queen, who received a PhD in astrophysics in 2007.

[3] http://hdr.undp.org/en/content/human-development-index-hdi

thus provides a rich description of what recourse is, explain its usage and the motivations behind specific recourse methods.

3.3.2 Normative adequacy. The capability approach is normatively adequate because it picks out relevant normative features for designing and assessing recommendations.

First, it picks out an important moral feature of recourse recommendations, viz. that people who can obtain a decision from an algorithm are in a better position than those who are not. Recommendations that provide recourse qua capability give them the real freedom to obtain decisions. Insofar as we accept that one key metric of well-being is people's capabilities, it follows that promoting the capability of recourse will also promote people's well-being. Second, the capability approach provides a substantive, but flexible, evaluative framework to design and compare recommendations. In particular, it provides the key metric recommendations should optimize for, namely capabilities. Consequently, good recommendations will be ones that fall within a person's capability set. If a person does not have the capability to achieve the recommendation, then that recommendation is not actionable and, crucially, that person does not have recourse. When assessing recommendations, we should thus pay special attention to whether people have the capability to achieve them.

As we noted earlier, there are various reasons why we would consider recourse to be valuable, for example because of its role in agency and autonomy. We do not deny that those may ground the value of recommendations. In fact, our goal is more modest: assuming we want recourse, what are fruitful ethical standards for designing and assessing recommendations? One key advantage of the capability approach over alternative evaluative frameworks is that it broadens the informational basis. For instance, it takes into account people's preferences, but also incorporates information about their conversion factors and the (real) freedom people have to achieve functionings. As a result, recommendations that aim to promote capabilities can come apart from recommendations that aim to solely satisfy preferences.

To illustrate, suppose someone would like to receive a recommendation for obtaining a loan. Recommendations that aim to promote the satisfaction of preferences face several challenges. One of them is that it is not always possible to act on one's preferences. Someone born in Canada might have a preference for becoming President of the United States, but it is impossible to satisfy that preference. Only natural-born-citizens may become President. Likewise, giving users recommendations that they prefer, but are not actionable to them, will not contribute to their well-being. Another challenge is that since the recommendation process may itself contribute to shaping preferences, then the users' preferences become a moving target. We could assume that a user seeking a recommendation for a loan would prefer to obtain it and that, accordingly, the recommendation should help the person satisfy that preference. However, a recommendation may show that obtaining the loan could only be done through a difficult process. Even though the person would have the capability of achieving the recommendation, she might choose, or prefer, to not do so. The capability approach emphasizes that giving users the *freedom* to realize a preference, not its actual satisfaction, is what matters for recourse.

A last challenge is that since the preference-satisfaction framework is fundamentally individualistic, it fails to take into account structural constraints from the social environment. On the contrary, the capability approach can incorporate larger social complexities via conversion factors and by broadening the informational basis [38, see, e.g., secs. 2.7.5 and 4.10]. This then allows to take into account differences between groups (see section 5 below).

One specific constraint that falls out of the capability approach is that recourse explanations should be diverse. In other words, in order for users to increase their capabilities requires that they are given more than one recommendation, and that these recommendations are in an important sense distinct. In the next section, we look closely at the constraint of diversity and the value it has for algorithmic recourse.

4 THE VALUE OF DIVERSE RECOMMENDATIONS

In what follows, we show the fruitfulness of conceptualizing algorithmic recourse as a recommendation problem and the fruitfulness of the capability approach by taking a close look at the constraint of *diversity* on algorithmic recourse. Wachter et al. [58] discuss the importance of providing diverse recourse recommendations, with many others agreeing [31, 40]. However, detail about *why* diversity matters and how diversity constraints specifically can overcome some of the problems facing recourse is lacking. Moreover, diversity constraints are largely undervalued in current research on algorithmic recourse. Only 16 of the 60 recourse algorithms found in a recent survey Karimi et al. [21] include diversity as a constraint. And of the works that include diversity, several lack sufficient detail motivating their choice of diversity metric. Meanwhile, the value of diversity in recommendation systems is well documented with several research lines investigating the best suited diversity metrics for specific use cases [26], as well as user perceptions and reactions to diversity [8, 19, 50]. Vrijenhoek et al. [57], in their work on diverse news recommendation, develop diversity metrics that reflect normative democratic values. In a similar vein, the capability approach can serve as a motivation for specific diversity metrics for algorithmic recourse.

4.1 Diversity for recourse recommendations

Providing users with a diverse set of recourse recommendations is currently motivated because of prevailing uncertainty in user preferences [23]. This problem has analogs to the cold start problem in recommendation systems, where recommendations are provided even when the system has little data regarding the user or their behavior [42]. Providing users with a diverse set of recommendations is one way to overcome the cold start problem [26]. However, there are additional reasons for valuing diversity besides uncertainty in user preferences. For example, in news recommendation diversity can help with combating filter-bubbles [29]. Importantly, depending on the overall purpose of diversity, different diversity metrics are more or less suitable [49, 57]. Thus, the fact that diversity in algorithmic recourse only seeks to address uncertainty in user preferences narrowly constrains the choice of diversity metrics. If diversity in recourse recommendations is valuable for other

purposes—e.g. broadening one's capability set—then the choice of suitable diversity metrics will be notably different.

The majority of works in algorithmic recourse understands diversity as a type of distance or similarity metric between counterfactuals [21]. While this approach may very well yield diverse counterfactuals that help to overcome uncertainty in user preferences, there are drawbacks. First, the similarity or distance function is operative both in generating the list of possible counterfactuals and also in selecting the diverse set. This can retain biases that result in determining distance or similarity in the first place. However, the value of diversity metrics is that they have the potential to counteract this bias by considering other trade-offs. For example, Dandl et al. [9] discuss diversity in relation to trade-offs between different objectives, such as the number of feature changes, closeness to the nearest observed data points, and plausibility according to a probability distribution. They argue that exploring trade-offs improves understandability and the number of options for the user compared to other approaches that build in *a priori* a weighted sum. Mothilal et al. [31] also describe different trade-offs. They identify proximity diversity and sparsity diversity. The former concerns the distance and the latter the number of features that need to be changed to reverse the decision.

Moreover, most current works on algorithmic recourse diversify recommendations post-hoc (i.e. after initial counterfactual generation). However, as learned from work in recommendation systems, post-hoc diversity methods face a problem that if the initial generated list is not diverse, the diversity metrics do little to help [26]. Making progress on the effectiveness of diversifying recourse recommendations starts with conceptualizing recourse as a recommendation problem and then learning from the various methods of diversity discussed in recommendation systems.

4.2 Capability approach and diverse recourse recommendations

The capability approach not only tells us why diverse recourse recommendations are valuable—because they increase the likelihood that a user actually has the capability to have recourse—it provides a way of thinking about ethical standards for diversity metrics. First, recommendations are usually evaluated based on how accurate recommendations are for fulfilling user preferences. However, the capability approach tells us that it is not preferences that should make up the evaluative space, but a user's capabilities. This would entail that evaluating whether a recourse recommendation is successful should not be geared toward preference-satisfaction, but promoting capabilities. Second, following the method of Vrijenhoek et al. [57], we identify two key normative themes that motivate how to diversify recourse recommendations. While it is possible that the capability approach could motivate more considerations of diversity, we highlight two that are currently missing from recourse diversity metrics.

4.2.1 Temporality. The capability approach highlights that capabilities have the potential to be realized involving various trade-offs and time frames, with Venkatasubramanian and Alfano [55] discussing the value of recourse as a type of temporally extended agency. Recourse recommendations can account for this temporal dimension by diversifying the time frame for realizing a capability. For example, getting an additional educational degree may take more time compared to other activities. Another aspect of temporality is the time it might take before particular capabilities become possible. For example, someone may have several capabilities that are only realizable after their children become a certain age.

The diversity metric of temporality diversifies recourse recommendations based on differences in user capability time frames. Current recourse techniques account for aspects of temporality through a brute cost function, with cost generally understood as a probability distribution for a given feature compared to others. Diversifying over temporality focuses on another kind of cost: time. It gives the user the ability to see for themselves the options for a shorter versus longer time frame potentials.

4.2.2 Resource conversion. The capability approach highlights that different people have different conversion factors (i.e. the differential capacity to convert resources into functionings). Equality of resources does not imply equality of capabilities. Resource conversion diversifies over a range of more or less resource intensive actions. While resource conversion shares many similarities with current cost metrics, the capability approach urges us to understand cost differently from the probability distribution method that is currently popular among recourse algorithms. The probability method of cost assumes that everyone has the same conversion factors. However, this is not the case. The capability approach motivates diversifying cost to reflect the differences in users' conversion factors. Gaining knowledge about a user's specific conversion factors could improve the accuracy of recommendations, but diversifying on resource conversion is still valuable according to the capability approach to facilitate option-freedom.

4.2.3 Limits of diversity. Maximizing diversity and including a never-ending list of diverse recommendations will not be successful for providing users with actionable choices. There are a variety of trade-offs that we need to consider when devising specific recourse recommendations. For example, people can face 'option overload' when there are too many live options to choose from. As a result, adding yet another diverse recommendation may actually reduce one's capability set since it makes it harder to convert a recommendation into an achievable functioning. Thus, it is important to engage in user-study research concerning the number of recommendations that is optimal. The length of the list could differ between users, with some users achieving their goals with two options, while for others, five options may be optimal. The capability approach may help in navigating how to handle such trade-offs. Specifically, user-studies should be designed that seek to validate the extent to which one's capability set is captured, instead of the feeling of trust the user has in the system. Additional options include getting user input regarding which diversity metrics they are interested in seeing for recourse recommendations.

5 SIGNIFICANCE FOR RECOURSE RESEARCH

The capability approach provides a conceptual and normative framework against which we can assess and compare different constraints and proposals for recommendations. Naturally, it does not (and will not) settle all disputes, but no theoretical framework can do that. But it is important to at least agree on *what terms* disputes should

be settled. These terms are that recommendations should promote people's capabilities. As a result, we believe that the capability approach may help define adequate optimization procedures besides diversity. In this section, we present several implications that adopting the capability approach has on current themes in recourse research.

5.1 Causality

Some recent work [e.g. 22] emphasize the importance of building causal models to provide actionable recommendations. One benefit of causal models is that they can assess which features are immutable or non-actionable in the sense of not being causally possible. Counterfactual explanations may not provide actionable recommendations if there is no causal path between the features the user would have to intervene on and the decision. For instance, it is not causally possible to increase one's level of education while reducing or keeping one's age constant. This why Karimi et al. propose a method for generating "recourse through minimal interventions". Minimal interventions aim to minimize the cost of implementing a set of actions that would change the decision.

Although causal possibility is certainly an important dimension of actionability, even if we assume away the problem of having perfect causal knowledge [see 23], the capability approach allows us to see that we arguably need to broaden the causal lens. Capabilities (or lack thereof) do not always neatly fall within the 'causal' category. Recall that capabilities are best understood as option-freedoms and that they are a function of the character of the options themselves and their access. One's route to achieving recourse may be more difficult and less accessible. One particularly pressing problem is that there might be a self-selection bias when people opt for some recommendations over others because of incorrect beliefs about what they can possibly do or not. Or, perhaps even more worrying, people might self-select because of normative beliefs about what they *should (not)* do. A woman might not consider a recommendation as actionable because it involves increasing her level of education, which would be frowned upon in her community. Other recommendations might be so burdensome as not falling within one's capability set, yet still being causally possible.

Another issue is whether conversion factors (personal, social, or environmental) can always be represented in causal terms. For instance, power relations and social norms may all affect one's ability to convert resources in capabilities. Moreover, it is contentious that social categories such as gender or race can be viewed as a cause [6, 14, 18, 24, 30, 60]. But even if factors such as those could be represented as having a positive or negative causal influence, our point is simply that accurate causal models need to address problems of possible causal break-down and the complexities surrounding the way conversion factors can be causally efficacious.

5.2 Proxies

One way of understanding the role of constraints for recommendation algorithms is that they are *proxies* for actionability. Reducing the distance between the factual and the counterfactual instance that crosses the decision boundary is one typical constraint. Other common constraints include 'plausibility' (i.e. likely to be actually instantiated) or 'sparsity' (i.e. recommending changes to as few variables as possible). Distance, plausibility, or sparsity are all proxies for actionability. Furthermore, as discussed above, since it is in practice difficult to build complete and accurate causal models [23], current causal models are also a proxy for actionability. Although not directly determining actionability, all the above constraints are often taken to constitute good approximations for actionable recommendations.

The capability approach provides a normative framework for assessing which proxies might better optimize the relevant notion of recommendation, viz. recommendations that people have the real freedom to achieve. For instance, the Human Development Index considers that income per capita, education level, and life expectancy are good indicators of human development along with the capabilities people have in different countries. From this, we could infer that people with more income, education, or life expectancy will have a greater capability to implement recommendations. The likelihood of providing a truly actionable recommendation for people who score high on these indicators should be greater. This is just one example of how recourse qua capability could be inferred, albeit imperfectly, from proxies. Fortunately, there is a significant literature on measuring capabilities in education, health, etc. [1, 3, 13, 33, 47, 54, 59, 62].[4] Designers of recommendations systems could find from other fields relevant proxies for providing recourse for various applications and contexts.

One key advantage of using the capability approach is that it helps answer ex ante and ex post questions about recommendations. The first is: What are the best proxies of people's *current* capabilities? This is directly related to actionability insofar as we want to provide recommendations that people have the real freedom to achieve. Following the capability approach, the answer to that question is that the recommendation should fall within one's capability set. Providing diverse recommendations is one important means to achieve that goal. But the second, often underrated, question is: What recommendations would most *improve* people's lives? The capability approach would suggest that recommendations that improve more people's capabilities are the better ones. Consider again the case of the proxies for human development (income, education, health). On that basis, we might conclude that recommendations that would privilege acting on income, education, and health may have the greater impact on people's capabilities. Ceteris paribus, people with more income, education, or health are typically freer to achieve functionings. This would suggest to favor recommendations that have the greater ex post impact.

5.3 Tough recommendations

Some recommendations may be actionable yet be 'far-fetched' in the sense of too difficult or burdensome to achieve. Venkatasubramanian and Alfano [55, sec. 4.6] argue that we should refrain from giving such recommendations. Although we agree that such recommendations may not be relevant in many cases, the capability approach suggests caution before a priori deciding that a recommendation is too difficult or burdensome. First, classifying a recommendation as too costly implies that we have sufficient

[4] See, e.g., [2, 37, 53] for discussions of challenges to measuring and operationalizing the capability approach.

information about users' current capabilities. In many cases, this assumption does not hold, which is also why recommendations should be diverse. Second, this may unduly interfere with people's capabilities. Nudging or not providing recommendations may affect the access to options as well as the options themselves. For instance, people may come to believe that acting on a recommendation is too hard for them, which might not really be the case. Or, excluding recommendations may restrict the quantity and quality of options people believe they have access to. In any case, we should be very wary of allowing recommendations systems to limit the availability of recommendations.

5.4 Strategic manipulation

One concern of recourse research is that users may try to strategically manipulate algorithms. From the perspective of the capability approach, it is unclear why 'gaming the system' is a problem for users. If we want to promote people's capabilities, giving people recommendations that they may use for achieving functionings that they value would indeed promote their capabilities. This may seem like a bug, but it is a feature. Indeed, if our concern is to provide ethical standards for assessing and designing recommendations for *users*, then our foundations should not exclude trading-off the good of the users for the good of other stakeholders. We may have reasons to not design recommendations systems that users can game, but these reasons are external to actionability and user well-being.

5.5 Fairness

One important motivation for making sure that recommendations are actionable is that some recommendations may be actionable for one person and not for another. However, mere actionability may not capture all the features we want from good recommendations. A recommendation may be actionable for two different people yet differ in their cost. This raises issues of fairness, especially if the grounds for the cost are unjust. Recommendations that are more costly for particular groups or communities may signal that there is discrimination. For example, just recommendations to acquire more work experience may ignore various work and care responsibilities that differ between groups. If we want recourse to be fair, we thus need an account of recourse fairness.

Gupta et al. [17] propose to measure recourse fairness in terms of the average group distance to the decision boundary. However, as von Kügelgen et al. [56] note, distance-based notions do not take into account the real causal effects—and thus costs—of intervening on variables. Accordingly, they suggest an individual and group-level *causal* notion of recourse fairness. Although arguably a step in the right direction, a causal approach faces several obstacles. One is that thinking of discrimination in causal terms is contentious (see sec. 5.1 above).[5] Another more serious issue is that causal reasoning will not tell, by itself, what causes *should* count. For instance, some theories of justice consider that burdens and benefits should be distributed according to desert [7]. A recommendation might be costly for a person, but she might *deserve* to be in that position. One might argue that the proverbial surfer failing to save should perhaps not obtain a loan so easily.

Although the capability approach does not solve by itself all issues related to algorithmic fairness, it provides a theoretical framework within which to conceptualize these problems. Someone more interested in the fairness of outcomes could try to optimize for recommendations that provide fair functionings; others more interested in opportunities may instead consider that capabilities should be the key metric of justice. And the notion of 'conversion factors' provides a language to formulate various issues related to fairness. Social conversion factors can be social norms that discriminate and personal conversion factors such as having a disability may justify compensating people seeking recourse.

6 CONCLUSION

Designers of algorithmic systems are often interested in providing recourse to users, viz. the ability to obtain or reverse a decision from a model. Recourse has often been associated with providing counterfactual explanations. We first proposed to reframe recourse not as an explanation problem, but as a recommendation problem. The aim of recourse is not necessarily to understand why the model made the decision, but rather simply to allow users to achieve results they value. Not all explanations provide recourse and not all recommendations provide understanding. One benefit of viewing recourse as a recommendation problem is that it leverages the existing literature on recommendation systems. But it also creates a challenge for designers of these systems: What are good recommendations?

We argued that the capability approach provides plausible and fruitful ethical standards for the design of recommendation systems whose goal is to give recourse to users. The capability approach is both descriptively and normatively adequate; it captures the relevant features of recourse and provides an ethical justification for why some recommendations are better than others. In particular, we submitted that good recommendations will be those that promote people's *capabilities*. To illustrate the relevance of the framework, we discussed one particular constraint to recourse, diversity. We closed by discussing several implications of adopting the capability approach for recourse research beyond diversity.

To conclude, we would like to emphasize that the capability approach is not the only framework which can be used to conceptualize the ethical constraints to recourse. Although there might be other suitable alternatives in some contexts, we simply hold that the capability approach is a worthy contender. That being said, one important message we hope our discussion conveyed is that if recourse is to live up to its ethical promise, then we cannot dispense with examining the ethical assumptions underlying what we take good recommendations to be.

ACKNOWLEDGMENTS

This work is supported by the Netherlands Organization for Scientific Research (NWO grant number VI.Veni.201F.051), and part of the research programme Ethics of Socially Disruptive Technologies, which is funded by the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

[5] They also propose to improve recourse fairness through "societal interventions". However, these interventions are not easily available to individuals seeking recourse and it is thus unclear why they should qualify as recommendations in our sense.

The authors discussed this work with Maastricht University's xAI research group, the ESDiT Society Line, at TU Dortmund, ACFAS 2022, and the ECPAI ML Opacity Circle. We thank the participants for comments on previous versions of the manuscript.

REFERENCES

[1] Haya Al-Ajlani, Luc Van Ootegem, and Elsy Verhofstadt. 2020. Does Well-Being Vary with an Individual-Specific Weighting Scheme? *Applied Research in Quality of Life* 15, 5 (Nov. 2020), 1285–1302. https://doi.org/10.1007/s11482-019-09733-0

[2] Sabina Alkire. 2013. Choosing Dimensions: The Capability Approach and Multidimensional Poverty. In *The Many Dimensions of Poverty*, Nanak Kakwani and Jacques Silber (Eds.). Palgrave Macmillan, London, 89–119. https://doi.org/10.1057/9780230592407_6

[3] Sabina Alkire, José Manuel Roche, Paola Ballon, James Foster, Maria Emma Santos, and Suman Seth. 2015. *Multidimensional Poverty Measurement and Analysis*. Oxford University Press, Oxford.

[4] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, and Probublica. 2016. Machine Bias. There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[5] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 80–89. https://doi.org/10.1145/3351095.3372830

[6] Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. 2015. Causally Interpreting Intersectionality Theory. *Philosophy of Science* 83, 1 (2015), 60–81. https://doi.org/10.1086/684173

[7] Huub Brouwer and Thomas Mulligan. 2019. Why Not Be a Desertist? *Philosophical Studies* 176, 9 (2019), 2271–2288. https://doi.org/10.1007/s11098-018-1125-4

[8] Sylvain Castagnos, Armelle Brun, and Anne Boyer. 2013. When Diversity Is Needed... But Not Expected!. In *International Conference on Advances in Information Mining and Management*. IARIA XPS Press, 44.

[9] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. In *Parallel Problem Solving from Nature – PPSN XVI (Lecture Notes in Computer Science)*, Thomas Bäck, Mike Preuss, André Deutz, Hao Wang, Carola Doerr, Michael Emmerich, and Heike Trautmann (Eds.). Springer International Publishing, Cham, 448–469. https://doi.org/10.1007/978-3-030-58112-1_31

[10] Sarah Dean, Sarah Rich, and Benjamin Recht. 2020. Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 436–445. https://doi.org/10.1145/3351095.3372866

[11] Julia Dressel and Hany Farid. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580

[12] Michael Friedman. 1974. Explanation and Scientific Understanding. *Journal of Philosophy* 71, 1 (1974), 5–19.

[13] Wulf Gaertner and Yongsheng Xu. 2006. Capability Sets as the Basis of a New Measure of Human Development. *Journal of Human Development* 7, 3 (Nov. 2006), 311–321. https://doi.org/10.1080/14649880600815891

[14] Clark Glymour and Madelyn R. Glymour. 2014. Commentary: Race and Sex Are Causes. *Epidemiology* 25, 4 (2014), 488–490.

[15] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable Credit Application Predictions With Counterfactual Explanations. *arXiv:1811.05245 [cs]* (Nov. 2018). arXiv:1811.05245 [cs]

[16] Stephen R. Grimm. 2010. The Goal of Explanation. *Studies In History and Philosophy of Science Part A* 41, 4 (2010), 337–344. https://doi.org/10.1016/j.shpsa.2010.10.006

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

[18] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[19] Rong Hu and Pearl Pu. 2011. Helping Users Perceive Recommendation Diversity. In *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011)*. Chicago, Illinois, USA, 43–50.

[20] 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:1907.09615 [cs, stat]* (July 2019). arXiv:1907.09615 [cs, stat]

[21] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2021. A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects. *arXiv:2010.04050 [cs, stat]* (March 2021). arXiv:2010.04050 [cs, stat]

[22] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3442188.3445899

[23] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic Recourse under Imperfect Causal Knowledge: A Probabilistic Approach. *arXiv:2006.06831 [cs, stat]* (Oct. 2020). arXiv:2006.06831 [cs, stat]

[24] Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 228–236. https://doi.org/10.1145/3442188.3445886

[25] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. 2021. A Causal Perspective on Meaningful and Robust Algorithmic Recourse. *arXiv:2107.07853 [cs, stat]* (July 2021). arXiv:2107.07853 [cs, stat]

[26] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in Recommender Systems – A Survey. *Knowledge-Based Systems* 123 (2017), 154–162. https://doi.org/10.1016/j.knosys.2017.02.009

[27] Himabindu Lakkaraju. 2021. Towards Reliable and Practicable Algorithmic Recourse. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 4.

[28] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery. *Queue* 16, 3 (2018), 31–57. https://doi.org/10.1145/3236386.3241340

[29] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. 2020. The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism* 8, 5 (May 2020), 605–642. https://doi.org/10.1080/21670811.2020.1764374

[30] Alexandre Marcellesi. 2013. Is Race a Cause? *Philosophy of Science* 80, 5 (2013), 650–659. https://doi.org/10.1086/673721

[31] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. https://doi.org/10.1145/3351095.3372850

[32] Martha C. Nussbaum. 2000. *Women and Human Development: The Capabilities Approach*. Cambridge University Press, Cambridge.

[33] Ilse Oosterlaken. 2009. Design for Development: A Capability Approach. *Design Issues* 25, 4 (2009), 91–102. https://doi.org/10.1162/desi.2009.25.4.91

[34] Philip Pettit. 2015. *The Robust Demands of the Good: Ethics with Attachment, Virtue, and Respect*. Oxford University Press.

[35] Peyman Rasouli and Ingrid Chieh Yu. 2021. CARE: Coherent Actionable Recourse Based on Sound Counterfactual Explanations. *arXiv:2108.08197 [cs]* (Aug. 2021). arXiv:2108.08197 [cs]

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[37] Ingrid Robeyns. 2003. Sen's Capability Approach and Gender Inequality: Selecting Relevant Capabilities. *Feminist Economics* 9, 2-3 (2003), 61–92. https://doi.org/10.1080/1354570022000078024

[38] Ingrid Robeyns. 2017. *Wellbeing, Freedom and Social Justice: The Capability Approach Re-examined*. Open Book Publishers.

[39] Ingrid Robeyns and Morten Fibieger Byskov. 2021. The Capability Approach. In *The Stanford Encyclopedia of Philosophy* (winter 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[40] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 20–28. https://doi.org/10.1145/3287560.3287569

[41] Jean Ryan, Anders Wretstrand, and Steven M. Schmidt. 2015. Exploring Public Transport as an Element of Older Persons' Mobility: A Capability Approach Perspective. *Journal of Transport Geography* 48 (2015), 105–114. https://doi.org/10.1016/j.jtrangeo.2015.08.016

[42] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and Metrics for Cold-Start Recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 253–260. https://doi.org/10.1145/564376.564421

[43] Amartya Sen. 1979. Issues in the Measurement of Poverty. *The Scandinavian Journal of Economics* 81, 2 (1979), 285–307. https://doi.org/10.2307/3439966

[44] Amartya Sen. 1980. Equality of What? In *The Tanner Lecture on Human Values*. Vol. 1. Cambridge University Press, Cambridge, 197–220.

[45] Amartya Sen. 1992. *Inequality Reexamined*. Harvard University Press, Cambridge, MA.

[46] Amartya Sen. 2009. *The Idea of Justice*. The Belknap Press of Harvard University Press, Cambridge, MA.

[47] Judit Simon, Paul Anand, Alastair Gray, Jorun Rugkåsa, Ksenija Yeeles, and Tom Burns. 2013. Operationalising the Capability Approach for Outcome Measurement in Mental Health Research. *Social Science & Medicine* 98 (2013), 187–196. https://doi.org/10.1016/j.socscimed.2013.09.019

[48] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

[49] Emily Sullivan, Dimitrios Bountouridis, Jaron Harambam, Shabnam Najafian, Felicia Loecherbach, Mykola Makhortykh, Domokos Kelen, Daricia Wilkinson, David Graus, and Nava Tintarev. 2019. Reading News with a Purpose: Explaining User Profiles for Self-Actualization. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. ACM, Larnaca Cyprus, 241–245. https://doi.org/10.1145/3314183.3323456

[50] Nava Tintarev, Matt Dennis, and Judith Masthoff. 2013. Adapting Recommendation Diversity to Openness to Experience: A Study of Human Behaviour. In *User Modeling, Adaptation, and Personalization (Lecture Notes in Computer Science)*, Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro (Eds.). Springer, Berlin, Heidelberg, 190–202. https://doi.org/10.1007/978-3-642-38844-6_16

[51] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*. 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[52] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. https://doi.org/10.1145/3287560.3287566

[53] Willem J. A. van der Deijl. 2020. A Challenge for Capability Measures of Well-being. *Social Theory and Practice* 46, 3 (2020), 605–631. https://doi.org/10.5840/soctheorpract202071799

[54] Luc Van Ootegem and Elsy Verhofstadt. 2012. Using Capabilities as an Alternative Indicator for Well-being. *Social Indicators Research* 106, 1 (2012), 133–152. https://doi.org/10.1007/s11205-011-9799-4

[55] Suresh Venkatasubramanian and Mark Alfano. 2020. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 284–293. https://doi.org/10.1145/3351095.3372876

[56] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2021. On the Fairness of Causal Algorithmic Recourse. *arXiv:2010.06529 [cs, stat]* (June 2021). arXiv:2010.06529 [cs, stat]

[57] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. Association for Computing Machinery, New York, NY, USA, 173–183.

[58] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–887.

[59] Melanie Walker and Elaine Unterhalter. 2007. The Capability Approach: Its Potential for Work in Education. In *Amartya Sen's Capability Approach and Social Justice in Education*, Melanie Walker and Elaine Unterhalter (Eds.). Palgrave Macmillan, New York, 1–18. https://doi.org/10.1057/9780230604810_1

[60] Naftali Weinberger. 2021. Signal Manipulation and the Causal Status of Race. https://doi.org/10/1/Signal_Manipulation_and_the_Causal_Status_of_Race-7.pdf

[61] Carlos Zednik. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology* 34, 2 (2021), 265–288. https://doi.org/10.1007/s13347-019-00382-7

[62] Yingqin Zheng. 2009. Different Spaces for E-Development: What Can We Learn from the Capability Approach? *Information Technology for Development* 15, 2 (2009), 66–82. https://doi.org/10.1002/itdj.20115