Grounding and Validation of Algorithmic Recourse in Real-World Contexts: A Systematized Literature Review

Anonymous Author(s)
Affiliation
Address
email

Abstract

The aim of algorithmic recourse (AR) is generally understood to be the provision of "actionable" recommendations to individuals affected by algorithmic decision-making systems, in an attempt to offer the capacity for taking actions that may lead to more desirable outcomes in the future. Over the past few years, AR literature has largely focused on theoretical frameworks to generate "actionable" counterfactual explanations that further satisfy various desiderata, such as diversity or robustness. We believe that algorithmic recourse, by its nature, should be seen as a practical problem: real-world socio-technical decision-making systems are complex dynamic entities involving various actors (end users, domain experts, civil servants, system owners, etc.) engaged in social and technical processes. Thus, research needs to account for the specificities of systems where it would be applied. To evaluate how authors envision AR "in the wild", we carry out a systematized review of 127 publications pertaining to the problem and identify the real-world considerations that motivate them. Among others, we look at the ways to make recourse (individually) actionable, the involved stakeholders, the perceived challenges, and the availability of practitioner-friendly open-source codebases. We find that there is a strong disconnect between the existing research and the practical requirements for AR. Most importantly, the grounding and validation of algorithmic recourse in real-world contexts remain underexplored. As an attempt to bridge this gap, we provide other authors with five recommendations to make future solutions easier to adapt to their potential real-world applications.

1 Introduction

Algorithmic decision-making (ADM) tools are frequently seen as a way to improve decision processes in a variety of high-stakes domains such as public administration [47, 146] or healthcare [45, 87]. Deep learning models have attracted much attention due to their perceived high performance, but the predictions of such models cannot be interpreted by humans, hence end users – both individuals subjected to algorithmic decisions and decision-makers operating on them – are placed in a position where they are unable to understand the grounds of a prediction, act on it, or trust it [159].

To help address this problem, a variety of explanation methods has been proposed. Of particular interest for this paper are counterfactual explanations (CEs) that attempt to explain the predictions for individual instances of data, taking the form of conditional statements such as *"if the value of feature x was a instead of b, the model would have predicted class y instead of z"*. They are perceived to be an attractive approach to explanation that does not require "opening the black box" [151] and have been argued to align with the ways that humans naturally reason about events [84].

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

CEs are also seen as the go-to method for algorithmic recourse (AR), or the generation of actionable recommendations that provide people with the knowledge needed to achieve more desirable predictions in ADM systems. Recourse is distinct from the "explanation" or "justification" of algorithmic decisions, and more closely related to the notion of contestability of Artificial Intelligence [7] in that it aims not only to improve the *trust* in the algorithm, but also embrace human *agency* [142].

Algorithmic recourse is an inherently practical problem in that it resembles a bureaucratic complaint process: an individual unhappy with some decision engages with a representative of the issuing organization, in an attempt to overturn it. Yet, we observe that much of the existing work is highly theoretical, with little consideration of whether it could be applied in organizational settings [see also 18]. Deploying AR in realistic systems without analyzing its mechanics in a broader context and without knowing what types of dynamics are expected to arise is bound to lead to unanticipated outcomes. Many of them will be undesirable and even potentially unsafe, and impossible to validate with respect to a set of requirements because the requirements for AR are *necessarily* socio-technical.

Societal and institutional components of algorithmic recourse are the focal point of our work, as we look beyond the typical technical considerations to assess the practical aspects of the problem.

To that end, we contribute a *systematized review* of 127 publications that address the goals of algorithmic recourse and we evaluate to what extent they incorporate such practical considerations. We characterize our approach as *systematized* because we follow a fully systematic approach to the collection of publications, but their selection is not necessarily exhaustive [46] as many impactful ideas in computer science are published only in the form of pre-prints. Based on our analysis, we also provide other authors with five recommendations on how to improve the practicality of AR research.

The rest of the manuscript is structured as follows. In Section 2 we elaborate on the background of our work. Then, in Section 3 we describe our approach to this review. Next, Section 4 introduces our findings. Section 5 provides a discussion of our results, introduces our recommendations, and addresses the limitations of the current work. Finally, Section 6 forms the conclusion to this paper.

2 Background

2.1 On algorithmic recourse

Algorithmic – or actionable, individual – recourse was introduced in [138] as *"the ability of a person to change the decision of the model through actionable input variables"*, building on the earlier work of [151] who argued that CEs are a psychologically-grounded way to (1) help decision-subjects understand an algorithmic decision, (2) provide them with information needed to contest it, and (3) inform about actions that could be taken to overturn it. For instance, consider a person who has unsuccessfully applied for a loan; they may then receive AR such as *"if you requested $5000 less, you would qualify for this loan"*. The key consideration for AR is "actionability", which entails that the recipient of the recommendation should be capable of implementing it. If they had been informed *"if you were 10 years younger, you would qualify for the loan"*, they would have still received a valid CE, *but not* recourse. More recently [69] has recast the problem as reasoning about minimal interventions on the structural causal model. This formulation (at least theoretically) addresses an important shortcoming of "correlational" recourse. Without accounting for the downstream causal effects of actions, an individual may exert more effort than necessary and still fail to achieve the target outcome. Indeed, counterfactuals are an inherently causal concept [103].

We note that problems similar to AR have been studied under a variety of different names: *actionable knowledge discovery* [e.g., 2], *action rules mining* [e.g., 110], *inverse classification* [e.g., 5], *why not questions* [e.g., 58], or *actionable feature tweaking* [134]. These alternative formulations have generally focused on "business" knowledge, rather than individual recommendations, but ultimately the goal of all these approaches is to extract information from a (black-box) model that allows the user – an individual or a decision-maker – to act. We highlight them to emphasize that AR does not have to be achieved through the means of CEs. Rather CEs should be seen as *one of the means* to achieve AR, particularly promising in that they do not require expert-level understanding of the model to be useful. Nonetheless, we decide to distinguish between the literature on AR (commonly equated with actionable CEs), and these alternative formulations in our work.

Existing research has generally considered AR in simplistic settings that are far removed from real-world socio-technical decision-making systems, where it would be implemented as a process.

For example, such systems are dynamic [113, 137], must support the implementation of AR at scale [9, 94], and involve various stakeholders beyond the end users [17, 151]. Moreover, if the intended goal of AR is to help individuals subjected to algorithmic decisions in an effective manner, research must entail a rich understanding of "actionability" to account for the differences between them [142].

2.2 On the position of our review

Several groups of authors have previously surveyed the landscape of counterfactual explanations in general, and algorithmic recourse specifically. Perhaps the most relevant to our work is [71], which discusses five deficits of research on CEs, with a special focus on the (lack of) psychological grounding. Another pertinent publication is [70], which attempts to unify the definitions and formulations of AR in existing literature, but the work primarily focuses on technical aspects. Next, [143] develops a rubric to compare counterfactual explainers (equated with AR) and identifies 21 research challenges. While these also remain mostly technical, several of them are relevant to our work, for instance, CEs *"as an interactive service to the applicants"* or reinforcing *"the ties between machine learning and regulatory communities"*. More recently, [48] reviewed and benchmarked a number of CE generators, but AR is only a secondary consideration in the work. We also highlight [130], which is the only systematic review of counterfactual and contrastive approaches to date. The authors understand CEs as a way to justify model predictions (i.e., they are different from AR). We agree with this distinction in that CEs can be useful for reasons other than recourse, such as model debugging [e.g., 1, 122]. Finally, although not reviews, [13] and [142] are particularly relevant to our work, offering critical perspectives on AR and addressing multiple shortcomings of recourse literature.

3 Methods

In this section, we briefly discuss our approach to the literature review following the SALSA – Search, Appraisal, Synthesis, Analysis – framework introduced in [46]. We also provide a more detailed description to allow for the reproduction of our process in the supplementary materials. Figure 1 presents our process in the form of a PRISMA flow diagram [97].

3.1 Search

We make use of three search engines to collect the initial set of studies: ACM Digital Library, IEEE Xplore, and SCOPUS. Given the previously mentioned blurry distinction between AR and CEs, we consider the papers discussing either problem. In a small scoping review, we identify several keywords common to publications on recourse, as well as several equivalent terms to build the query. We search in titles, abstracts, and keywords, arriving at 3092 records after de-duplication. To facilitate the screening process, we employ the open-source ASReview tool, which makes use of an active learning approach to re-order the set of publications, such that the most relevant ones are always "at the top of the stack" [139]. The researchers behind the tool suggest employing a stopping rule measured in the number of consecutive irrelevant records, which we set to 30, or 1% of the entire dataset. We accept all papers that focus on algorithmic recourse and counterfactual explanations, completing the screening after evaluating 1040 abstracts, leading to 499 relevant records.

We observe that some important publications may be missing from our results. For instance, [151] was published in a legal journal that is not indexed by computer science search engines. Thus, we decide to augment the set of records by applying snowballing, which has been shown as a good alternative to databases in systematic reviews in software engineering [162]. We collect the references for the top 50 (10%) "most impactful" publications, measured by the number of citations. While this introduces several pre-prints into our result set [52, 61, 91, 113, 143, 150], we decide not to exclude them. Our review remains primarily concerned with peer-reviewed work. After adding the snowballed references to our dataset, we are left with 2018 records for the second screening with ASReview. This time, we look for publications that specifically refer to the problem of AR, "actionable" CEs, or modifying outcomes of automated decision-making systems. We employ a stricter stopping rule to minimize the risk of false negatives, completing the screening after 60 consecutive irrelevant records with 203 records considered for full-text appraisal. To allow for complete reproducibility of the search process, we provide an extended discussion (including queries) in the technical Appendix A.

3.2 Appraisal

We were able to retrieve all of the remaining 203 documents. For each document, we require that the authors explicitly cite recourse as the center of interest, or look at **(1)** explanations **(2)** provided for individual instances **(3)** with the goal of acting upon them **(4)** in an attempt to modify the predictions **(5)** of a classification model. We exclude 51 publications as they are not on topic, primarily because they focus on CEs for the sake of explanation. Four works in this category look at (what they call) recourse but extend the problem to settings beyond the scope of this review: recommender systems [31, 43, 145], text classification [37], and anomaly detection [27]. Further 15 publications are duplicates, typically pre-prints of other documents that were included in the review. Next, 8 documents were published before [151] that sparked the research on AR, and thus we exclude them as well. These look at the alternative formulations discussed earlier in Section 2.1. Finally, 2 documents are not publications: one is an abstract of a talk, and the other is a student poster. For each document, we answer a number of questions relating to the practical considerations introduced by the authors.

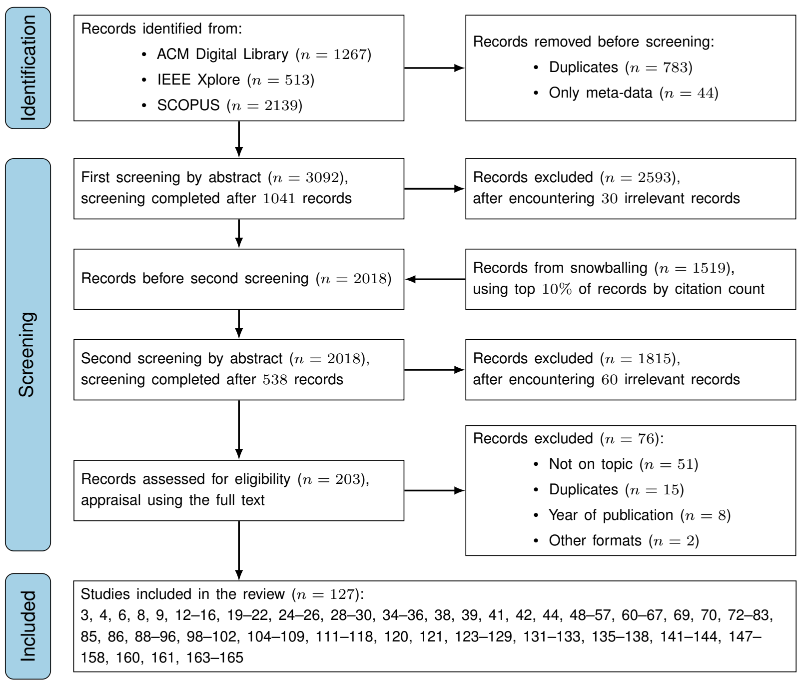

Figure 1: Identification of studies via databases and snowballing

3.3 Synthesis

To compile the results we carry out a standard thematic content analysis following the approach presented in [40]. First, we explore the data extracted from the set of publications relevant to each question to find the commonalities, which serves as the grounds for creating the initial set of codes. We evaluate the documents against these codes and keep track of any other considerations. If such considerations appear in multiple documents, we create new codes for them. Afterward, we re-evaluate all documents against the new code. As the coding exercise is carried out by one author, they do a third pass over all documents to double-check for potential errors. Finally, where relevant, we cluster the codes into larger themes. In this analysis we only look at the explicit statements provided by the authors, we do not attempt to infer their understanding of the problem. Thus, the numbers provided in Section 4 should be understood as describing how algorithmic recourse is *discussed* in the literature. For brevity, we focus our discussion on the main themes, but we still highlight specific publications if we observe that the authors introduce novel, highly relevant considerations that do not fit into other themes. Finally, even though we also evaluated the technical aspects of the proposed solutions – requirements for methods and datasets used in evaluations – they are not covered in this review. Instead, we point the interested readers to [48, 70, 143].

4 Results

The following nine sections introduce the results of the thematic analysis. For each question, we explain why it is relevant to the analysis and examine the main themes. We also highlight highly important but underexplored themes. We start with the general points such as contributions and definitions in Sections 4.1 to 4.3. Then, in Sections 4.4 to 4.7 we investigate the societal components of AR research. Finally, in Sections 4.8 and 4.9 we look at the aspects relevant to practitioners.

4.1 What types of contributions do the authors choose to make to the AR research?

We start by looking at the main goals of the collected publications to validate our assumption that AR literature is primarily concerned with technical solutions. We annotate each entry with at most two codes based on the form of contributions. By far the largest group is *propose methods*, which applies to 88 (69.3%) out of the 127 publications. These are primarily generators for individual CEs, but we also find 18 (14.2%) documents that propose other methods. Next, 20 (15.7%) publications *develop theoretical frameworks*, for instance by grounding AR in user studies or providing critical perspectives on the problem. Further, 15 (11.8%) focus on *empirical or theoretical analyses* of the properties of AR and another 15 publications *apply* it in a variety of domains. We did not identify any applications evaluated with humans in the loop. Then, 5 (3.9%) publications *benchmark* existing methods, while 3 (2.4%) *review* them. We make our annotations available in technical Appendix B.

4.2 What are the criteria covered in the authors' definitions of AR?

We also evaluate what is understood as the problem to be addressed by AR mechanisms. In particular, what are the criteria to satisfy authors' definitions of recourse. A similar question was posed by [70] who combined six definitions into *"recourse can be achieved by an affected individual if they can understand and accordingly act to alleviate an unfavorable situation, thus exercising temporally-extended agency"*, but this approach was far from systematic. Instead, we are interested in the underlying concepts. 74 (58.3%) publications explicitly define AR, 16 (12.6%) mention it but do not include a definition, while 37 (29.1%) do not mention AR, even though they align with its (overall) goals. The most common theme is *overturning undesirable decisions*, present in 47 definitions (63.5% of all definitions), but specifically *overturning algorithmic decisions* is mentioned only 43 (58.1%) times. It is generally understood that *AR is provided to affected individuals* (44, or 59.5%) but 4 (5.4%) definitions *consider stakeholders* more broadly. *Actionability* as a requirement for recourse is noted in only 39 (52.7%) definitions. Then, 20 (27.0%) publications specifically mention counterfactual explanations as means to AR, while 26 (35.1%) include various other technical considerations in the definitions, such as "changes to actionable input variables" or "desired classes".

We also point to several themes that are, interestingly, underrepresented. Only 18 (24.3%) documents mention *explanation, justification, or understanding of a decision* as the pre-requisite for AR. Next, 10 (13.5%) highlight *future-orientation or other temporal aspects* of the provided recommendations. Although *"consequential settings"*, typically bank lending, are given as examples in nine (12.2%) definitions, they are never explicitly mentioned as the scenarios where recourse ought to be provided, which may be akin to the "enjoyment of recourse" as defined by [142] where people are aware that there exists a way to reverse undesirable decisions.[1] 8 publications (10.8%) promote *AR as an ability*. Finally, only 2 (2.7%) publications require that recourse accounts for the *preferences* of its recipients.

4.3 What are the criteria covered in the authors' definitions of actionability?

As we observe, "actionability" is a concept that underpins AR but we discover that, in general, its understanding is limited. 91 (71.6%) publications attempt to define what it means (for a CE) to be actionable. Most commonly, in 48 (52.7%) out of 91 definitions, it is understood as *acting only on directly-mutable features*, 6 (6.6%) distinguish that *features may be indirectly-mutable* but still not actionable, while 22 (24.2%) also highlight that *feature values may need to be constrained*. Next, 19 (20.9%) definitions rely on a tautology that actionability means *people can take actions*, 11 (12.1%) emphasize that these *actions must be successful or lead to change*, and 3 (3.3%) further require that they are *aligned with people's real-world objectives*. Only 14 (15.4%) definitions put users

[1]Financial domain dominates the evaluations as well, with 90 of 116 evaluations on non-synthetic data making use of at least one finance-related dataset, most commonly `German Credit Data` [59] with 51 uses.

at the center stage, indicating that actionability *depends on the user or their preferences*, while 2 (2.2%) highlight the *importance of the context* [144, 156], for instance, that the ability to act on a recommendation may change over time. Importantly, ethical considerations are never mentioned as the pre-requisite for actionability, but we find some broader discussions about this [e.g., 142].

4.4 What is the role of end users? What other stakeholders are envisioned in the AR process?

Given that AR is to be implemented in socio-technical systems that include a variety of actors, we are interested in the types of stakeholders acknowledged in the literature. A total of 105 publications provide explicit consideration of this type. In general, end users subject to algorithmic decisions are envisioned to be the recipients of AR, but this is not always the case: it may also be provided to experts [e.g., 21, 22, 76] or organizations [e.g., 65, 72, 147], which highlights that in some cases AR may be carried out on behalf of the affected individuals. In any case, 47 (44.8%) publications in the subset agree that end users should inform actionability, but it is rarely clear *how* these preferences should be specified. User-friendly (interactive) interfaces are a consideration in only 14 (13.3%) documents. A total of 29 (27.6%) publications envision domain experts as someone who inform the recourse process. They are either expected to inform actionability in the AR system or provide other forms of knowledge, typically in the form of a causal structure. Besides the experts, authors of 35 (33.3%) papers have discussed a variety of stakeholders. Most commonly system owners [e.g., 20, 34, 38, 89], but also auditors [e.g., 138, 158], data scientists [e.g., 28, 82], developers [e.g., 22, 131], practitioners [e.g., 100, 156], regulators [e.g., 28, 120], or even potential attackers [102].

4.5 What types of real-world considerations motivate existing research?

With the multitude of challenges that stand ahead of real-world AR, we are interested in the considerations that motivate existing work. The main theme we find is *ensuring proper individual actionability*, which is addressed in 46 (37.4%) of 123 publications relevant to this question. This is typically achieved with the encoding of user preferences as constraints, but other means include providing diverse CEs. In fact, *tackling specific desiderata for AR* (beyond actionability) is the second largest area of research with 28 (22.8%) publications. Various *other technical challenges* are considered in 24 (19.5%) documents, for example, integrating background knowledge [e.g., 16, 62, 64, 98], or incorporating feature importance [e.g., 4, 6, 96, 116]. We also find 19 (15.4%) publications that discuss the problem of *communicating recourse to the end users*. 16 (13.0%) focus on the *dynamics of real-world systems*, typically addressing the robustness of AR [e.g., 75, 91, 93, 137], while 14 (11.4%) look at recourse in *multi-agent systems*. This also relates to *performance considerations* emphasized in 15 (12.2%) of documents. *Causality* drives research in 14 (11.4%) cases. We also find several themes that are under-emphasized: only 9 (7.3%) publications are directly *motivated by research in psychology*, while *ethics of AR* are emphasized in only 7 (5.7%) documents.

4.6 What types of real-world considerations are seen as challenges for future work?

While the previous section looked at the considerations that drive existing research, in this section we distill the recommendations for *future* research going beyond the improvement of own work, which are provided in 74 documents. *Causality* is highlighted as a challenge in 22 (29.7%) of them, while *other technical considerations* are given in 20 (27.0%) cases. These range from robustness [e.g., 51, 117, 137], support for categorical features [e.g., 36, 157], or distinguishing between valid CEs and adversarial examples [101]. Next, 19 (25.6%) documents highlight the importance of *ensuring proper individual actionability*, which also relates to *communicating recourse to the end users* (9, or 12.2%) and *supporting realistic cost functions* (8, or 10.8%). *Ethics of AR* are highlighted in 11 (14.9%) publications, for example, that AR research may detract from other obligations of model owners [77, 133]. The same number of publications emphasize the need to (1) *ground research in user studies*, and (2) accommodate for the *dynamics of real-world systems*. *Privacy or security* is highlighted in 10 (13.5%) documents, while the *abuse of recourse*, such as strategic behaviors, surfaces in 7 (9.4%) papers. Other challenges include improving *performance* (8, or 10.8%), considering *multi-agent systems* (4, or 5.4%), and developing *legal frameworks* (4, or 5.4%) for recourse. We also highlight several challenges particularly relevant to our work: (the usefulness of) recourse is perceived as difficult to evaluate in practice [41, 60, 115], it must account for individual, contextual, societal, and even cultural factors [123], which further means that engagement with recourse mechanisms and the likelihood of its implementation are context-dependent [e.g., 6, 42, 128].

4.7 What types of (emergent) group-level dynamics are addressed in the existing research?

Real-world systems entail the implementation of recourse by more than one agent, which may introduce group-level dynamics. Nonetheless, out of 119 documents relevant to this question, 93 (78.2%) seem to understand recourse as a purely individual phenomenon. Among the remaining 26 documents we find considerations for several different group-level effects. Various perspectives on the problem of fair AR, covering both individual and group formulations are addressed by [12, 36, 52, 120, 121, 131, 149, 154]. Next, [9] shows that the implementation of AR on a large scale may lead to domain and model shifts, which introduce unexpected costs for the stakeholders.[2] In [42] the authors focus on another negative consequence of AR at scale, showing that it may reinforce social segregation. The impact of the "right to be forgotten", where data deletion requests trigger model retraining that may invalidate existing recourses is addressed in [75]. Then, [94] develop a game-theoretic framework for AR in multi-agent settings, attempting to optimize for "social welfare" rather than the profits of individual agents. We find two further similar perspectives on recourse: [38] proposes auditing and subsidies to minimize the risks of strategic behaviors in a multi-agent setting, while [136] attempts to incentivize actual improvement for a population of agents. Finally, [65] provides a framework that generates transparent and consistent recourses for a sub-population. We also note two other lines of research that account for the remaining documents with group-level considerations. First, in a causal setting [e.g., 68, 73] subpopulations are necessary to estimate the interventional effects on individuals. Second, several works highlight the importance of global insights into the data [22, 41, 44, 78, 108, 112, 152], such as recourse summaries [78, 112].

4.8 What are the approaches to the realistic evaluation of proposed methods?

We now explore the different forms of "real-world" evaluations, going beyond quantitative experiments, which are present in 51 publications. Most commonly, in 28 (54.9%) of those, the authors make use of *case studies* presenting the methods in an end-to-end manner. Among those, the application of recourse in the `Hired.com` marketplace goes furthest in simulating real-world conditions for AR [89], but the recommendations are still not evaluated with humans in the loop. Further, 9 (17.6%) documents include other forms of *short walk-through examples*. We also identify 14 (27.5%) papers that evaluate the methods with *user experiments*, 10 of which involve non-expert users and 4 involve expert users. While we do not observe any interviews with non-expert users, we find 1 (2.0%) publication where *experts are interviewed* [22]. *Other involvement of non-experts* applies to [116], where they inform the development of methods. *Other involvement of experts* is featured in two documents where they evaluated the outputs of methods [25, 132]. Altogether, end users were involved in 17 publications, which is only 13.3% of all publications covered in our study, even more striking than the 21% of CE methods evaluated with user studies as reported in [71].

4.9 What are the open source and documentation practices in AR research?

Finally, we note that the lack of availability of well-documented open-source code may be an important obstacle to the application of AR in real-world systems. For all 116 publications that involve some form of computational experiments, we verify whether the source code is publicly available. If the authors do not explicitly link to their code in the paper, we attempt to find it independently. Ultimately, we collect open-source implementations for 64 (55.2%) publications. Then, for each of them, we evaluate the quality of documentation. The *instructions on the general usage* (such as installation and workflow) are provided with 27 (41.5%) repositories, while *instructions on the reproduction of results* in 23 (35.4%). In 19 (29.2%) cases we find *walk-through tutorials*, typically in the form of Jupyter Notebooks, although we note that they differ in quality. For instance, 5 repositories include code-only notebooks with no further textual explanation that could guide the practitioner. Implementations for 4 papers include more "professionalized" *documentation* [9, 86, 100, 156]. The latter sets a golden standard as it further includes a tutorial video and a live demo. We do not find *any* additional materials for practitioners for 13 (20.0%) of the available implementations.

[2] Such "endogenous dynamics" were postulated earlier in the first version of [113] dated December 22$^{\text{nd}}$ 2020, but this discussion has been completely removed from the subsequent versions of the pre-print.

5 Discussion

Regardless of whether AR can be normatively expected or not [77], many systems can genuinely benefit from recourse mechanisms, especially when the interests of the system owner and the end users are aligned [72], such as in the healthcare system to improve the well-being of patients [76, 96, 155], or on the online platforms that attempt to improve the experience of their users [89, 134]. Nonetheless, the values and norms underlying recourse – trust, agency, fairness, safety, and so on – are emergent properties of systems where recourse mechanisms would be introduced. Such norms can only be understood and evaluated when accounting for the technical, social, and institutional components of the system [32], but the latter two remain largely unexplored in the recourse literature.

Recourse is not inherently safe or unsafe, *but* its (incorrect) implementation may lead to the emergence of unsafe dynamics, such as the unexpected costs to stakeholders as discussed by [9] or the reinforcement of social segregation addressed in [42]. While it may be too challenging to provide accurate system-level evaluations at this stage of research, authors can still expand the boundaries of their analyses to account for global effects or look at the position of recourse mechanisms in the broader context of a complete socio-technical AI system [33]. As AR is a "reality-centric AI" problem [140] by its nature, working towards its integration into existing systems will require a design-oriented approach, potentially with *specific* systems in mind. The "Abstraction Traps" discussed by [119] in the context of research on fair machine learning apply here: that technical solutions designed for one social context cannot be directly repurposed for another application, that values to which they are expected to adhere to cannot be captured with mathematical formulas, that their insertion into an existing process will impact its behavior, or that the best solutions may not necessarily be technical.

It is perhaps most telling that only 12% of surveyed publications attempt to apply recourse in realistic settings. We will discuss two of these settings to highlight the stark differences in system properties. Most of the applications included in our review focus on the provision of actionable individual recommendations to students [3, 4, 24, 109, 126, 135, 160]. In this relatively low-stakes domain almost any recourse will be actionable in that following a personalized set of learning activities does not require any resources other than time. Even then, the system involves multiple actors – students, teachers, parents – whose interactions will impact the process, for example, because students may fail to benefit from certain learning activities without additional support. Conversely, we find several publications where authors attempt to provide recourse in the high-stakes medical domain [76, 96, 155]. Here, recommendations must be tailored to the preferences, resources, or lifestyles of patients in order to have a chance of being actionable. Moreover, certain aspects of their implementation fully rely on other actors, such as a clinician prescribing the medications. Finally, it may happen that recourse does not exist at all when the outcomes of a patient cannot be improved.

5.1 Recommendations for future research

We distill our findings into five key recommendations. First, in Sections 4.2, 4.3 we observed that *operational* definitions for recourse are still unavailable. Second, Sections 4.4 and 4.8 underlined little consideration for people involved in recourse processes. Third, Sections 4.5, 4.6 highlighted the overwhelmingly technical approaches to recourse. Fourth, Section 4.7 stressed the lack of group-level analyses. Fifth, from Sections 4.8, and 4.9 we learned about the missing consideration of practitioners.

1. Broadening the scope of research. AR is generally seen as a service for affected individuals, but this formalization may be unnecessarily limiting. In fact, in many systems, these individuals may be unable to directly act on recommendations [see also 142]. Instead, we propose to operationalize the aim of AR as the provision of recommendations *aligned with the preferences* of *non-expert users* in an attempt *to help them improve outcomes* in an *ADM setting*, which emphasizes that providing *easy to understand* and *individually actionable* recommendations remains the key research problem.

2. Engaging end users, affected individuals, and communities. AR solutions are rarely evaluated with humans. Instead, they attempt to satisfy a variety of desiderata formulated by authors and assessed in an automated manner. Sparsity, proximity, or mutability of features are far from perfect proxies for individual actionability. For AR to be truly useful, it must be able to satisfy the preferences of its end users. Research is also necessary to learn about the needs of the affected individuals concerning recourse, and to validate its potential contributions and inherent limitations. Authors may also benefit from the rich literature on human-computer interaction [e.g., 11, 23] or psychology.

3. Accepting a socio-technical perspective. A pervasive assumption in the literature is that all challenges of AR require purely technical solutions. For instance, many authors emphasize the importance of causal modeling to guarantee recourse, but the models that aim to be explained are themselves *not* causal. Similarly, to improve the performance of CE generators many authors turn to deep generative models [35, 42, 61, 67, 81, 90, 99]. Not only do they explain the data rather than the model [10], but more importantly they shift the problem from improving the trust in non-interpretable models, to attempting to trust non-interpretable explainers. Although a socio-technical perspective on AR brings its own challenges, such as accounting for the roles of stakeholders involved in the provision of recourse, it creates important opportunities. For example, developing "recourse contracts" [34, 39] or designing feedback processes to account for imperfect robustness.

4. Accounting for emergent effects. Decision-making systems involve multiple individuals who may be interested in receiving recourse and may have competing interests. Research on AR should, from the onset, explore group-level effects such as external costs or fairness. While this may require expanding the boundaries of analysis, it is necessary to anticipate the emergent outcomes of recourse. These may even occur due to the multi-system dynamics of AR: recommendations implemented by an individual to improve their outcomes in one system will affect them in other contexts [see also 13].

5. Attending to other operational aspects. Finally, the artifacts of AR research should be practitioner-friendly. On the one hand, this requires being explicit about the position of the proposed methods in a broader system, for example, in the form of end-to-end case studies that allow practitioners to better understand the benefits of the proposed solutions. On the other hand, this suggests that authors should attempt to move away from merely providing scripts for experiments, and focus on developing well-documented frameworks that can be adapted to different ADM systems.

5.2 Limitations of our work

Our review is not without shortcomings. Most importantly, for each paper the extraction and coding of data was performed by a single author, which means that the quantitative results may be imperfect. We account for this by focusing the analysis on the *overarching themes* represented in existing publications, thus, even if another researcher would have carried out the coding in a somewhat different manner, they should arrive at similar results and our analysis remains valid. Additionally, as our review ultimately looks at the authors' perception of recourse, we do not want to misconstrue their views. Thus, we do not infer any considerations unless they are provided explicitly. Our reading may be more strict than intended by the authors and the numbers reported in our results may be underestimated. At the same time, we believe that if certain considerations are deemed important by the researchers, they would choose to be explicit about them. Finally, although we followed a systematic process, we cannot claim that we collected AR literature in an exhaustive manner due to the specificities of computer science publishing. Thus, we acknowledge that there may exist some insightful publications addressing recourse that have not been covered in this literature review.

6 Conclusions

Algorithmic recourse concerns the provision of recommendations aligned with the preferences of non-expert users of algorithmic decision-making systems to help them achieve more desirable outcomes in the future. Existing research on the topic is predominantly theoretical, even though recourse, in expectation, is a real-world problem with strong practical implications. To that end, we conducted a systematized literature review of 127 publications that focus on algorithmic recourse, and more generally on actionable counterfactual explanations. We evaluated the practical considerations provided by the authors. Our findings indicate that, indeed, AR tends to be perceived as a (predominantly) technical problem. Although we think highly of fundamental research, we note that for algorithmic recourse to leave computer science labs, it must be more strongly grounded and validated in the real world, and consider the requirements for systems that include not only technical but also social and institutional components. To help bridge this gap, we synthesize a list of five recommendations for other authors that aim to reinforce recourse as a practical problem. We believe that AR should not be seen as only a simple ad-hoc solution to improve the acceptance of black-box models in consequential domains, but rather as a full-fledged socio-technical mechanism that can benefit many systems and improve the agency of affected individuals and decision-makers across a variety of settings.

References

[1] Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully Debugging Model Mistakes using Conceptual Counterfactual Explanations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 66–88. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/abid22a.html`.

[2] Gediminas Adomavicius and Alexander Tuzhilin. Discovery of Actionable Patterns in Databases: The Action Hierarchy Approach. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, KDD'97, page 111–114. AAAI Press, 1997.

[3] Farzana Afrin, Margaret Hamilton, and Charles Thevathyan. Exploring Counterfactual Explanations for Predicting Student Success. In *Computational Science – ICCS 2023*, volume 14074 LNCS, pages 413–420. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-36021-3_44.

[4] Muhammad Afzaal, Jalal Nouri, Aayesha Zia, Panagiotis Papapetrou, Uno Fors, Xiu Wu, Yongchaoand Li, and Rebecka Weegar. Automatic and Intelligent Recommendations to Support Students' Self-Regulation. In *2021 International Conference on Advanced Learning Technologies (ICALT)*, pages 336–338, July 2021. ISBN 2161-377X. doi: 10.1109/ICALT52272.2021.00107.

[5] Charu C. Aggarwal, Chen Chen, and Jiawei Han. The Inverse Classification Problem. *Journal of Computer Science and Technology*, 25:458–468, 2010.

[6] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. Counterfactual Shapley Additive Explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1054–1070, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533168. URL `https://doi.org/10.1145/3531146.3533168`.

[7] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. Contestable AI by design: Towards a framework. *Minds and Machines*, pages 1–27, 2022.

[8] Hissah Alotaibi and Ronal Singh. Metrics for Evaluating Actionability in Explainable AI. In *PRICAI 2023: Trends in Artificial Intelligence*, pages 481–487. Springer Nature Singapore, 2023. ISBN 978-981-99-7022-3.

[9] Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia C. S. Liem. Endogenous Macrodynamics in Algorithmic Recourse. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 418–431, 2023. doi: 10.1109/SaTML54575.2023.00036.

[10] Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10829–10837, 2024.

[11] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, 2014.

[12] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating Robustness of Counterfactual Explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–09, December 2021. doi: 10.1109/SSCI50451.2021.9660058.

[13] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 80–89, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372830. URL `https://doi.org/10.1145/3351095.3372830`.

[14] Hosein Barzekar and Susan McRoy. Achievable Minimally-Contrastive Counterfactual Explanations. *Machine Learning and Knowledge Extraction*, 5(3):922–936, 2023. doi: 10.3390/make5030048.

[15] Sander Beckers. Causal Explanations and XAI. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 90–109. PMLR, 11–13 Apr 2022. URL `https://proceedings.mlr.press/v177/beckers22a.html`.

[16] Alexander Berman, Ellen Breitholtz, Christine Howes, and Jean-Philippe Bernardy. Explaining Predictions with Enthymematic Counterfactuals. In *CEUR Workshop Proceedings*, volume 3319, pages 95–100. CEUR-WS, 2022.

[17] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 648–657, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3375624. URL `https://doi.org/10.1145/3351095.3375624`.

[18] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533083. URL `https://doi.org/10.1145/3531146.3533083`.

[19] Miguel Á. Carreira-Perpiñán and Suryabhan Singh Hada. Counterfactual Explanations for Oblique Decision Trees: Exact, Efficient Algorithms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:6903–6911, May 2021. doi: 10.1609/aaai.v35i8.16851. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16851`.

[20] Yatong Chen, Jialu Wang, and Yang Liu. Strategic Recourse in Linear Classification. In *Workshop on Consequential Decision Making in Dynamic Environments*, 2020.

[21] Ziheng Chen, Fabrizio Silvestri, Gabriele Tolomei, Jia Wang, He Zhu, and Hongshik Ahn. Explain the Explainer: Interpreting Model-Agnostic Counterfactual Explanations of a Deep Reinforcement Learning Agent. *IEEE Transactions on Artificial Intelligence*, 5(4):1443–1457, 2024. doi: 10.1109/TAI.2022.3223892.

[22] Furui Cheng, Yao Ming, and Huamin Qu. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization & Computer Graphics*, 27(02):1438–1447, feb 2021. ISSN 1941-0506. doi: 10.1109/TVCG.2020.3030342.

[23] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300789. URL `https://doi.org/10.1145/3290605.3300789`.

[24] Lea Cohausz. Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 361–367, Durham, United Kingdom, July 2022. International Educational Data Mining Society. ISBN 978-1-7336736-3-1. doi: 10.5281/zenodo.6853069.

[25] Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, and Beatriz San Miguel Gonzalez. Counterfactual Explanations as Interventions in Latent Space. *Data Mining and Knowledge Discovery*, 2022. doi: 10.1007/s10618-022-00889-2.

[26] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-Objective Counterfactual Explanations. In *Parallel Problem Solving from Nature – PPSN XVI*, pages 448–469, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58112-1. doi: 10.1007/978-3-030-58112-1_3.

[27] Debanjan Datta, Feng Chen, and Naren Ramakrishnan. Framing Algorithmic Recourse for Anomaly Detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 283–293, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539344. URL `https://doi.org/10.1145/3534678.3539344`.

[28] Randall Davis, Andrew W. Lo, Sudhanshu Mishra, Arash Nourian, Manish Singh, Nicholas Wu, and Ruixun Zhang. Explainable Machine Learning Models of Consumer Credit Risk. *Journal of Financial Data Science*, 5(4):9–39, 2022. doi: 10.3905/jfds.2023.1.141.

[29] Marcelo de Sousa Balbino, Luis Enrique Zárate Gálvez, and Cristiane Neri Nobre. CSSE - An agnostic method of counterfactual, selected, and social explanations for classification models. *Expert Systems with Applications*, 228:120373, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.120373. URL `https://www.sciencedirect.com/science/article/pii/S0957417423008758`.

[30] Giovanni De Toni, Bruno Lepri, and Andrea Passerini. Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis. *Machine Learning*, 112(4):1389–1409, 2023. ISSN 0885-6125. doi: 10.1007/s10994-022-06293-7.

[31] Sarah Dean, Sarah Rich, and Benjamin Recht. Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 436–445, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372866. URL `https://doi.org/10.1145/3351095.3372866`.

[32] Roel Dobbe and Anouk Wolters. Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context. *Minds and Machines*, 34(2):1–51, 2024.

[33] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103555. URL `https://www.sciencedirect.com/science/article/pii/S0004370221001065`.

[34] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, and Bernhard Schölkopf. On the Adversarial Robustness of Causal Algorithmic Recourse. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5324–5342. PMLR, 17–23 2022.

[35] Michael Downs, Jonathan L. Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. CRUDS: Counterfactual Recourse Using Disentangled Subspaces. *ICML Workshop on Human Interpretability in Machine Learning*, pages 1–23, 2020.

[36] Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schoelkopf, and Setareh Maghsudi. Robustness Implies Fairness in Causal Algorithmic Recourse. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 984–1001, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594057. URL `https://doi.org/10.1145/3593013.3594057`.

[37] Julia El Zini and Mariette Awad. Beyond Model Interpretability: On the Faithfulness and Adversarial Robustness of Contrastive Textual Explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1391–1402. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.100.

[38] Andrew Estornell, Yatong Chen, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Incentivizing Recourse through Auditing in Strategic Classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 400–408. International Joint Conferences on Artificial Intelligence, 8 2023. doi: 10.24963/ijcai.2023/45. URL `https://doi.org/10.24963/ijcai.2023/45`.

[39] Andrea Ferrario and Michele Loi. The Robustness of Counterfactual Explanations Over Time. *IEEE Access*, 10:82736–82750, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3196917.

[40] Susanne Friese, Jacks Soratto, and Denise Pires de Pires. Carrying out a computer-aided thematic content analysis with ATLAS.ti. *IWMI Working Papers*, 18, 04 2018.

[41] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, pages 577–590, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8343-1. doi: 10.1145/3448016.3458455.

[42] Ruijiang Gao and Himabindu Lakkaraju. On the Impact of Algorithmic Recourse on Social Segregation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[43] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. PRINCE: Provider-Side Interpretability with Counterfactual Explanations in Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pages 196–204, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-6822-3. doi: 10.1145/3336191.3371824.

[44] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. ViCE: Visual Counterfactual Explanations for Machine Learning Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pages 531–535, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7118-6. doi: 10.1145/3377325.3377536.

[45] Crystal Grant. Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism, October 2022. URL `https://www.aclu.org/news/privacy-technology/algorithms-in-health-care-may-worsen-medical-racism`. Accessed 22.05.2024.

[46] Maria J. Grant and Andrew Booth. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2):91–108, 2009. doi: https://doi.org/10.1111/j.1471-1842.2009.00848.x.

[47] Stephan Grimmelikhuijsen and Albert Meijer. Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, 5(3):232–242, 03 2022. ISSN 2398-4910. doi: 10.1093/ppmgov/gvac008. URL `https://doi.org/10.1093/ppmgov/gvac008`.

[48] Riccardo Guidotti. Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking. *Data Mining and Knowledge Discovery*, 2022. doi: 10.1007/s10618-022-00831-6.

[49] Riccardo Guidotti and Salvatore Ruggieri. Ensemble of Counterfactual Explainers. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings*, pages 358–368, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-88941-8. doi: 10.1007/978-3-030-88942-5_28.

[50] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Francesca Naretto, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Stable and Actionable Explanations of Black-Box Models through Factual and Counterfactual Rules. *Data Mining and Knowledge Discovery*, 2022. doi: 10.1007/s10618-022-00878-5.

[51] Hangzhi Guo, Feiran Jia, Jinghui Chen, Anna Squicciarini, and Amulya Yadav. RoCourseNet: Robust Training of a Prediction Aware Recourse Model. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 619–628, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615040.

[52] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing Recourse Across Groups. *arXiv*, 2019.

[53] Victor Guyomard, Françoise Fessant, Tassadit Bouadi, and Thomas Guyet. Post-hoc Counterfactual Generation with Supervised Autoencoder. In *Communications in Computer and Information Science*, volume 1524 CCIS, pages 105–114. Springer Science and Business Media Deutschland GmbH, 2021. doi: 10.1007/978-3-030-93736-2_10.

[54] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. Generating Robust Counterfactual Explanations. In *Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2023*, pages 394–409, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-43417-4. doi: 10.1007/978-3-031-43418-1_24.

[55] Suryabhan Singh Hada and Miguel Á. Carreira-Perpiñán. Exploring Counterfactual Explanations for Classification and Regression Trees. In *Communications in Computer and Information Science*, volume 1524 CCIS, pages 489–504. Springer Science and Business Media Deutschland GmbH, 2021. doi: 10.1007/978-3-030-93736-2_37.

[56] Aparajita Haldar, Teddy Cunningham, and Hakan Ferhatosmanoglu. RAGUEL: Recourse-Aware Group Unfairness Elimination. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 666–675, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9236-5. doi: 10.1145/3511808.3557424.

[57] Ian Hardy, Jayanth Yetukuri, and Yang Liu. Adaptive Adversarial Training Does Not Increase Recourse Costs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 432–442, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604704.

[58] Zhian He and Eric Lo. Answering Why-not Questions on Top-k Queries. *2012 IEEE 28th International Conference on Data Engineering*, pages 750–761, 2012. doi: 10.1109/ICDE.2012.8.

[59] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[60] Jacqueline Höllig, Aniek F. Markus, JJef de Slegte, and Prachi Bagave. Semantic Meaningfulness: Evaluating Counterfactual Approaches for Real-World Plausibility and Feasibility. In *Communications in Computer and Information Science*, volume 1902 CCIS, pages 636–659. Springer Science and Business Media Deutschland GmbH, 2023. doi: 10.1007/978-3-031-44067-0_32.

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

[62] Sarathi K, Shania Mitra, Deepak P, and Sutanu Chakraborti. Counterfactuals as Explanations for Monotonic Classifiers. In *CEUR Workshop Proceedings*, volume 3389, pages 177–188. CEUR-WS, 2022.

[63] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. *Transactions of the Japanese Society for Artificial Intelligence*, 36(6), 2021. doi: 10.1527/TJSAI.36-6_C-L44.

[64] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike, Kento Uemura, and Hiroki Arimura. Ordered Counterfactual Explanation by Mixed-Integer Linear Optimization. In *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, volume 13A, pages 11564–11574. Association for the Advancement of Artificial Intelligence, 2021.

[65] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 1846–1870. PMLR, 2022.

[66] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-Agnostic Counterfactual Explanations for Consequential Decisions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 895–905. PMLR, 2020.

[67] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, 33:265–277, 2020.

[68] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Towards Causal Algorithmic Recourse. In *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 139–166, Cham, 2020. Springer International Publishing. ISBN 978-3-031-04082-5. doi: 10.1007/978-3-031-04083-2_8.

[69] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 353–362, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445899.

[70] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5), December 2022. ISSN 0360-0300. doi: 10.1145/3527848.

[71] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4466–4474. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/609. URL `https://doi.org/10.24963/ijcai.2021/609`. Survey Track.

[72] Nwaike Kelechi and Licheng Jiao. Quantifying Actionability: Evaluating Human-Recipient Models. *IEEE Access*, 11:119811–119823, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3324906.

[73] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. Causal Perspective on Meaningful and Robust Algorithmic Recourse. *ICML Workshop on Algorithmic Recourse*, 2021.

[74] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. Improvement-Focused Causal Recourse (ICR). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i10.26398.

[75] Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. Towards Bridging the Gaps between the Right to Explanation and the Right to Be Forgotten. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[76] Anisio Lacerda, Claudio Almeida, Leonardo Ferreira, Adriano Pereira, Gisele L. Pappa, Wagner Meira, Debora Miranda, Marco A. Romano-Silva, and Leandro Malloy Diniz. Algorithmic Recourse in Mental Healthcare. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2023. ISBN 2161-4407. doi: 10.1109/IJCNN54540.2023.10191158.

[77] Derek Leben. Explainable AI as evidence of fair decisions. *Frontiers in Psychology*, 14, 2023. doi: 10.3389/fpsyg.2023.1069426.

[78] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[79] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. FOCUS: Flexible Optimizable Counterfactual Explanations for Tree Ensembles. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, pages 5313–5322, 2022.

[80] Shucen Ma, Jianqi Shi, Yanhong Huang, Shengchao Qin, and Zhe Hou. Minimal-unsatisfiable-core-driven Local Explainability Analysis for Random Forest. *International Journal of Software and Informatics*, 12(4):355–376, 2022. doi: 10.21655/ijsi.1673-7288.00280.

[81] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. In *NeurIPS 2019 Workshop "Do the right thing": machine learning and causal inference for improved decision making*, 2019.

[82] Raphael Mazzine, Sofie Goethals, Dieter Brughmans, and David Martens. Counterfactual Explanations for Employment Services. *International workshop on AI for Human Resources and Public Employment Services*, 2021.

[83] Md Golam Moula Mehedi Hasan and Douglas A. Talbert. Mitigating the Rashomon Effect in Counterfactual Explanation: A Game-theoretic Approach. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS*, volume 35. Florida Online Journals, University of Florida, 2022. doi: 10.32473/flairs.v35i.130711.

[84] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL `https://www.sciencedirect.com/science/article/pii/S0004370218305988`.

[85] Jonathan Moore, Nils Hammerla, and Chris Watkins. Explaining Deep Learning Models with Constrained Adversarial Examples. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I*, pages 43–56, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-29907-1. doi: 10.1007/978-3-030-29908-8_4.

[86] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 607–617, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372850.

[87] Madhumita Murgia. Algorithms are deciding who gets organ transplants. Are their decisions fair?, November 2023. URL `https://www.ft.com/content/5125c83a-b82b-40c5-8b35-99579e087951`. Accessed 22.05.2024.

[88] Philip Naumann and Eirini Ntoutsi. Consequence-Aware Sequential Counterfactual Generation. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II*, pages 682–698, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-86519-1. doi: 10.1007/978-3-030-86520-7_42.

[89] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Providing Actionable Feedback in Hiring Marketplaces using Generative Adversarial Networks. In *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1089–1092. Association for Computing Machinery, 2021. doi: 10.1145/3437963.3441705.

[90] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. CounteRGAN: Generating Counterfactuals for Real-Time Recourse and Interpretability using Residual GANs. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence, UAI 2022*, pages 1488–1497. Association For Uncertainty in Artificial Intelligence (AUAI), 2022.

[91] Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Distributionally Robust Recourse Action. *arXiv*, 2023.

[92] Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Feasible Recourse Plan via Diverse Interpolation. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 4679–4698. PMLR, 2023.

[93] Tuan-Duy H. Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Robust Bayesian Recourse. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180, pages 1498–1508. PMLR, 2022.

[94] Andrew O'Brien and Edward Kim. Toward Multi-Agent Algorithmic Recourse: Challenges From a Game-Theoretic Perspective. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS*, volume 35. Florida Online Journals, University of Florida, 2022. doi: 10.32473/flairs.v35i.130614.

[95] Andrew O'Brien, Edward Kim, and Rosina Weber. Investigating Causally Augmented Sparse Learning as a Tool for Meaningful Classification. In *2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 33–37, September 2023. ISBN 2831-7203. doi: 10.1109/AIKE59827.2023.00013.

[96] Ming Lun Ong, Anthony Li, and Mehul Motani. Explainable and Actionable Machine Learning Models for Electronic Health Record Data. In *IFMBE Proceedings*, volume 79, pages 91–99, Cham, 2021. Springer International Publishing. doi: 10.1007/978-3-030-62045-5_9.

[97] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372, 2021.

[98] Axel Parmentier and Thibaut Vidal. Optimal Counterfactual Explanations in Tree Ensembles. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8422–8431. PMLR, 2021.

[99] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, pages 3126–3132, 2020. doi: 10.1145/3366423.3380087.

[100] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[101] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 4574–4594. PMLR, 2022.

[102] Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. On the Privacy Risks of Algorithmic Recourse. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 9680–9696. PMLR, 2023.

[103] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. ISBN 9780511803161.

[104] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 344–350, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3375627.3375850.

[105] Wenting Qi and Charalampos Chelmis. Improving Algorithmic Decision–Making in the Presence of Untrustworthy Training Data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1102–1108, 2021. doi: 10.1109/BigData52589.2021.9671677.

[106] Marcos M. Raimundo, Luis Gustavo Nonato, and Jorge Poco. Mining Pareto-optimal Counterfactual Antecedents with a Branch-and-Bound Model-Agnostic Algorithm. *Data Mining and Knowledge Discovery*, 2022. doi: 10.1007/s10618-022-00906-4.

[107] Goutham Ramakrishnan, Yun Chan Lee, and Aws Albarghouthi. Synthesizing Action Sequences for Modifying Model Decisions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5462–5469, 2020.

[108] Natraj Raman, Daniele Magazzeni, and Sameena. Shah. Bayesian Hierarchical Models for Counterfactual Estimation. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 1115–1128. PMLR, 2023.

[109] Gomathy Ramaswami, Teo Susnjak, and Anuradha Mathrani. Supporting Students' Academic Performance Using Explainable Machine Learning with Automated Prescriptive Analytics. *Big Data and Cognitive Computing*, 6(4), 2022. doi: 10.3390/bdcc6040105.

[110] Zbigniew W. Ras and Alicja Wieczorkowska. Action-Rules: How to Increase Profit of a Company. In *Principles of Data Mining and Knowledge Discovery*, pages 587–592. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-45372-7.

[111] Peyman Rasouli and Ingrid Chieh Yu. CARE: Coherent Actionable Recourse Based on Sound Counterfactual Explanations. *International Journal of Data Science and Analytics*, 17, 2022. doi: 10.1007/s41060-022-00365-6.

[112] Kaivalya Rawal and Himabindu Lakkaraju. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

[113] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts. *arXiv*, 2021.

[114] Annabelle Redelmeier, Martin Jullum, Kjersti Aas, and Anders Løland. MCCE: Monte Carlo sampling of realistic counterfactual explanations. In *Data Mining and Knowledge Discovery*, pages 421–437. Springer Nature, 2024. doi: 10.1007/s10618-024-01017-y.

[115] Alexis Ross, Himabindu Lakkaraju, and Osbert Bastani. Learning models for actionable recourse. In *Advances in Neural Information Processing Systems*, volume 34, pages 18734–18746, 2021.

[116] Pedram Salimi, Nirmalie Wiratunga, David Corsar, and Anjana Wijekoon. Towards Feasible Counterfactual Explanations: A Taxonomy Guided Template-Based NLG Method. In *Frontiers in Artificial Intelligence and Applications*, volume 372, pages 2057–2064. IOS Press BV, 2023. doi: 10.3233/FAIA230499.

[117] Maximilian Schleich, Zixuan Geng, Yihong Zhang, and Dan Suciu. GeCo: Quality Counterfactual Explanations in Real Time. *Proc. VLDB Endow.*, 14(9):1681–1693, may 2021. ISSN 2150-8097. doi: 10.14778/3461535.3461555. URL `https://doi.org/10.14778/3461535.3461555`.

[118] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1616–1628, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533218.

[119] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL `https://doi.org/10.1145/3287560.3287598`.

[120] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 166–172, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375812.

[121] Shubham Sharma, Alan H. Gee, David Paydarfar, and Joydeep Ghosh. FaiR-N: Fair and Robust Neural Networks for Structured Data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 946–955, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462559.

[122] Sunny Shree, Jaganmohan Chandrasekaran, Yu Lei, Raghu N. Kacker, and D. Richard Kuhn. DeltaExplainer: A Software Debugging Approach to Generating Counterfactual Explanations. In *2022 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pages 103–110, 2022. doi: 10.1109/AITest55621.2022.00023.

[123] Manan Singh, Sai Srinivas Kancheti, Shivam Gupta, Ganesh Ghalme, Shweta Jain, and Narayanan C. Krishnan. Algorithmic Recourse Based on User's Feature-Order Preference. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, CODS-COMAD '23, pages 293–294, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9797-1. doi: 10.1145/3570991.3571039.

[124] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. Directive Explanations for Actionable Explainability in Machine Learning Applications. *ACM Trans. Interact. Intell. Syst.*, 13(4), December 2023. ISSN 2160-6455. doi: 10.1145/3579363.

[125] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*, volume 34, pages 62–75, 2021.

[126] Bevan I. Smith, Charles Chimedza, and Jacoba H. Bührmann. Individualized Help for At-Risk Students Using Model-Agnostic and Counterfactual Explanations. *Education and Information Technologies*, 27(2):1539–1558, March 2022. ISSN 1360-2357. doi: 10.1007/s10639-021-10661-6.

[127] Jan-Tobias Sohns, Christoph Garth, and Heike Leitte. Decision Boundary Visualization for Counterfactual Reasoning. *Computer Graphics Forum*, 42(1):7–20, 2023. doi: 10.1111/cgf.14650.

[128] Nina Spreitzer, Hinda Haned, and Ilse van der Linden. Evaluating the Practicality of Counterfactual Explanations. In *CEUR Workshop Proceedings*, volume 3277, pages 31–50. CEUR-WS, 2022.

[129] Laura State, Salvatore Ruggieri, and Franco Turini. Reason to Explain: Interactive Contrastive Explanations (REASONX). In *Explainable Artificial Intelligence*, volume 1901 CCIS, pages 421–437, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44064-9.

[130] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9:11974–12001, 2021.

[131] Muhammad Suffian and Alessandro Bogliolo. Investigation and Mitigation of Bias in Explainable AI. In *CEUR Workshop Proceedings*, volume 3319, pages 89–94. CEUR-WS, 2022.

[132] Muhammad Suffian, Pierluigi Graziani, Jose M. Alonso, and Alessandro Bogliolo. FCE: Feedback Based Counterfactual Explanations for Explainable AI. *IEEE Access*, 10:72363–72372, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3189432.

[133] Emily Sullivan and Philippe Verreault-Julien. From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 712–722, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534185.

[134] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 465–474, 2017. doi: 10.1145/3097983.3098039.

[135] Maria Tsiakmaki and Omiros Ragos. A Case Study of Interpretable Counterfactual Explanations for the Task of Predicting Student Academic Performance. In *2021 25th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, pages 120–125, July 2021. doi: 10.1109/CSCC53858.2021.00029.

[136] Stratis Tsirtsis and Manuel Gomez-Rodriguez. Decisions, Counterfactual Explanations and Strategic Behavior. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

[137] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable Algorithmic Recourse. In *Advances in Neural Information Processing Systems*, volume 20, pages 16926–19937, 2021.

[138] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287566.

[139] Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133, 2021. doi: 10.1038/s42256-020-00287-7.

[140] Mihaela van der Schaar and Andrew Rashbass. The case for Reality-centric AI, Feb 2023. URL `https://www.vanderschaar-lab.com/the-case-for-reality-centric-ai/`. Accessed 21.05.2024.

[141] Peter M. VanNostrand, Huayi Zhang, Dennis M. Hofmann, and Elke A. Rundensteiner. FACET: Robust Counterfactual Explanation Analytics. *Proc. ACM Manag. Data*, 1(4), December 2023. doi: 10.1145/3626729.

[142] Suresh Venkatasubramanian and Mark Alfano. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 284–293, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372876.

[143] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *arXiv*, 2022.

[144] Sahil Verma, Keegan Hines, and John P. Dickerson. Amortized Generation of Sequential Algorithmic Recourses for Black-Box Models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, pages 8512–8519. Association for the Advancement of Artificial Intelligence, 2022.

[145] Sahil Verma, Ashudeep Singh, Varich Boonsanong, John P. Dickerson, and Chirag Shah. RecRec: Algorithmic Recourse for Recommender Systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 4325–4329, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615181.

[146] Kilian Vieth-Ditlmann. The algorithmic administration: automated decision-making in the public sector, May 2024. URL `https://algorithmwatch.org/en/algorithmic-administration-explained/`. Accessed 22.05.2024.

[147] Marco Virgolin and Saverio Fracaros. On the Robustness of Sparse Counterfactual Explanations to Adverse Perturbations. *Artificial Intelligence*, 316(C), March 2023. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103840.

[148] Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, Gholamreza Haffari, and Dinh Phung. Feature-Based Learning for Diverse and Privacy-Preserving Counterfactual Explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 2211–2222, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599343.

[149] Julius Von Kugelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Scholkopf. On the Fairness of Causal Algorithmic Recourse. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, pages 9584–9594. Association for the Advancement of Artificial Intelligence, 2022.

[150] Julius von Kügelgen, Nikita Agarwal, Jakob Zeitler, Afsaneh Mastouri, and Bernhard Schölkopf. Algorithmic Recourse in Partially and Fully Confounded Settings Through Bounding Counterfactual Effects. *arXiv*, 2021.

[151] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841, 2017.

[152] Paul Y. Wang, Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Demonstration of Generating Explanations for Black-Box Algorithms Using Lewis. *Proc. VLDB Endow.*, 14 (12):2787–2790, July 2021. ISSN 2150-8097. doi: 10.14778/3476311.3476345.

[153] Yongjie Wang, Qinxu Ding, Ke Wang, Yue Liu, Xingyu Wu, Jinglong Wang, Yong Liu, and Chunyan Miao. The Skyline of Counterfactual Explanations for Machine Learning Decision Models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 2030–2039, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8446-9. doi: 10.1145/3459637.3482397.

[154] Yongjie Wang, Hangwei Qian, Yongjie Liu, Wei Guo, and Chunyan Miao. Flexible and Robust Counterfactual Explanations with Minimal Satisfiable Perturbations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 2596–2605, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3614885.

[155] Zhendong Wang, Isak Samsten, Vasiliki Kougia, and Panagiotis Papapetrou. Style-Transfer Counterfactual Explanations: An Application to Mortality Prevention of ICU Patients. *Artif. Intell. Med.*, 135(C), January 2023. ISSN 0933-3657. doi: 10.1016/j.artmed.2022.102457.

[156] Zijie J. Wang, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. GAM Coach: Towards Interactive and User-Centered Algorithmic Recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3580816.

[157] Greta Warren, Mark T. Keane, and Ruth M. J. Byrne. Features of Explainability: How Users Understand Counterfactual and Causal Explanations for Categorical and Continuous Features in XAI. In *CEUR Workshop Proceedings*, volume 3251. CEUR-WS, 2022.

[158] Greta Warren, Barry Smyth, and Mark T. Keane. "Better" Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI). In *Case-Based Reasoning Research and Development: 30th International Conference, ICCBR 2022, Nancy, France, September 12–15, 2022, Proceedings*, pages 63–78, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-14922-1. doi: 10.1007/978-3-031-14923-8_5.

[159] Daniel S. Weld and Gagan Bansal. The Challenge of Crafting Intelligible Intelligence. *Commun. ACM*, 62(6):70–79, may 2019. ISSN 0001-0782. doi: 10.1145/3282486. URL `https://doi.org/10.1145/3282486`.

[160] Anjana Wijekoon, Nirmalie Wiratunga, Ikechukwu Nkisi-Orji, Kyle Martin, Chamath Palihawadana, and David Corsar. Counterfactual Explanations for Student Outcome Prediction with Moodle Footprints. In *CEUR Workshop Proceedings*, volume 2894, pages 1–8. CEUR-WS, 2021.

[161] Nirmalie Wiratunga, Anjana Wijekoon, Ikechukwu Nkisi-Orji, Kyle Martin, Chamath Palihawadana, and David Corsar. DisCERN: Discovering Counterfactual Explanations using Relevance Features from Neighbourhoods. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1466–1473, November 2021. ISBN 2375-0197. doi: 10.1109/ICTAI52525.2021.00233.

[162] Claes Wohlin. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. *EASE '14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 2014. doi: 10.1145/2601248.2601268. URL `https://doi.org/10.1145/2601248.2601268`.

[163] Jingquan Yan and Hao Wang. Self-Interpretable Time Series Prediction with Counterfactual Explanations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[164] Jayanth Yetukuri, Ian Hardy, and Yang Liu. Towards User Guided Actionable Recourse. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 742–751, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604708.

[165] Songming Zhang, Xiaofeng Chen, Shiping Wen, and Zhongshan Li. Density-Based Reliable and Robust Explainer for Counterfactual Explanation. *Expert Syst. Appl.*, 226(C), September 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.120214.

A Extended discussion of the search process

While our discussion of the search process in Section 3.1 in the main body of the document is complete, we also provide an extended version of this discussion to allow for full reproducibility.

We make use of 3 search engines to collect the initial set of studies: ACM Digital Library, IEEE Xplore, and SCOPUS. Given the blurry distinction between AR and CEs, we consider the papers discussing either problem. In a small scoping review, we identify several keywords common to publications on recourse, as well as several equivalent terms to build the query shown below.

```
(‘‘Machine Learning’’ OR ‘‘Artificial Intelligence’’
OR ‘‘Algorithmic Decision*’’ OR ‘‘Consequential Decision*’’
OR Classif* OR Predict* OR ‘‘Explainable AI’’ OR AI OR XAI)
AND (((Counterfactual OR Contrastive OR Actionable) AND Explanation*)
OR ((Algorithmic OR Individual* OR Actionable) AND Recourse)
OR Counterfactual?)
```

We modify this query to account for the semantic differences between the search engines.

For ACM Digital Library:

```
Title:(( "Machine Learning" OR "Artificial Intelligence"
OR "Algorithmic Decision*" OR "Consequential Decision*"
OR classif* OR predict* OR "Explainable AI" OR ai OR xai )
AND ( ( ( counterfactual OR contrastive OR actionable )
AND explanation* ) OR ( ( algorithmic OR individual* OR actionable )
AND recourse ) OR counterfactual? ))
OR Abstract:(( "Machine Learning" OR "Artificial Intelligence"
OR "Algorithmic Decision*" OR "Consequential Decision*"
OR classif* OR predict* OR "Explainable AI" OR ai OR xai )
AND ( ( ( counterfactual OR contrastive OR actionable )
AND explanation* ) OR ( ( algorithmic OR individual* OR actionable )
AND recourse ) OR counterfactual? ))
OR Keyword:(( "Machine Learning" OR "Artificial Intelligence"
OR "Algorithmic Decision*" OR "Consequential Decision*"
OR classif* OR predict* OR "Explainable AI" OR ai OR xai )
AND ( ( ( counterfactual OR contrastive OR actionable )
AND explanation* ) OR ( ( algorithmic OR individual* OR actionable )
AND recourse ) OR counterfactual? ))
```

For IEEE Xplore:

```
((("All Metadata":"Machine Learning"
OR "All Metadata":"Artificial Intelligence"
OR "All Metadata":"Algorithmic Decision*"
OR "All Metadata":"Consequential Decision*"
OR "All Metadata":classif* OR "All Metadata":predict*
OR "All Metadata":"Explainable AI" OR "All Metadata":ai
OR "All Metadata":xai )
AND ((("All Metadata":counterfactual OR "All Metadata":contrastive
OR "All Metadata":actionable ) AND "All Metadata":explanation* )
OR ( ("All Metadata":algorithmic OR "All Metadata":individual*
OR "All Metadata":actionable )
AND "All Metadata":recourse )
OR "All Metadata":counterfactual? )))
```

For SCOPUS:

```
TITLE-ABS-KEY ( ( "Machine Learning" OR "Artificial Intelligence"
OR "Algorithmic Decision*" OR "Consequential Decision*"
OR classif* OR predict* OR "Explainable AI" OR ai OR xai )
AND ( ( ( counterfactual OR contrastive OR actionable ) AND explanation* )
OR ( ( algorithmic OR individual* OR actionable ) AND recourse )
OR counterfactual? ) )
```

The search is carried out on January 12th 2024 in titles, abstracts, and keywords, with 1267 results from ACM Digital Library (The ACM Guide to Computing Literature), 513 results from IEEE Xplore, and 2139 results from SCOPUS. This leads to a total of 3919 results, which are imported to the Zotero reference management software for de-duplication. After removing the duplicates, we are left with 3136 results, 44 of which are the meta-data of conference proceedings that we also remove.

To facilitate the screening process, we employ the open-source ASReview tool, which makes use of an active learning approach to re-order the set of publications, such that the most relevant ones are always "at the top of the stack" [139]. We run ASReview on the default settings, i.e.:

```
Feature extraction technique: TF-IDF
Classifier: Naive Bayes
Query strategy: Maximum
Balance strategy: Dynamic resampling (Double)
```

The researchers behind the tool suggest employing a stopping rule measured in the number of consecutive irrelevant records, which we set to 30, or 1% of the entire dataset. We accept all papers that focus on algorithmic recourse and counterfactual explanations, completing the screening after evaluating 1040 abstracts (33.67% of the dataset), leading to 504 (16.30%) records among which we identify further 4 duplicates to remove. This results in the reported number of 499 relevant records.

We observe that some important publications may be missing from our results. For instance, [151] was published in the Harvard Journal of Law & Technology that is not indexed by computer science search engines. Thus, we decide to augment the set of records by applying snowballing, which has been shown as a good alternative to databases in systematic reviews in software engineering [162].

We decide to make use of citation counts as a proxy for impact. Due to the lack of a suitable tool that would provide unbiased citation counts for *all* papers in our dataset, we collect them from Google Scholar. Unfortunately, citation counts on Google Scholar tend to be inflated, but as we make use of snowballing purely to enrich the dataset, these does not impact the validity of our study. We manually collect Google Scholar citation counts for all 499 results from the first screening on January 27th and 28th, order them descendingly, and collect references for the top 50 (10%) "most impactful" publications. Snowballing results in a total of 1519 new records. Indeed, we observe that [151] (mentioned above) is referenced by 39 of the 50 publications used for snowballing.

While this strategy introduces several pre-prints into our result set [52, 61, 91, 113, 143, 150], we decide not to exclude them. Our review remains primarily concerned with peer-reviewed work. Here, we also note that [114], which we collected as a pre-print has been published between the search and appraisal. As such we decided to evaluate its published version and refer to it in this paper.

After adding the snowballed references into our dataset, we are left with 2018 records for the second screening with ASReview, again on the default settings. This time, we look for publications that specifically refer to the problem of AR, "actionable" CEs, or modifying outcomes of automated decision-making systems. We employ a stricter stopping rule to minimize the risk of false negatives, completing the screening after 60 consecutive irrelevant records. We evaluate 538 results (26.71% of the dataset), with 203 (10.06%) relevant results that are considered for full-text appraisal. This concludes the extended discussion of the search process.

B Evaluation of contributions

Table 1: Evaluation of the collected publications on the types of contributions, 2017-2021.

Year	Reference	Propose methods	Theoretical frameworks	Analyses	Apply	Benchmark	Review
2017	[151]	✓	✓				
2019	[52]	✓					
	[61]	✓					
	[81]	✓					
	[85]	✓					
	[138]	✓					
2020	[35]	✓					
	[86]	✓					
	[136]	✓					
	[20]	✓					
	[26]	✓					
	[44]	✓					
	[67]	✓					
	[66]	✓					
	[99]	✓					
	[104]	✓					
	[107]	✓					
	[120]	✓					
	[112]	✓					
	[13]		✓				
	[142]		✓				
2021	[69]	✓	✓				
	[137]	✓		✓			
	[41]	✓					
	[49]	✓					
	[53]	✓					
	[73]	✓					
	[150]	✓					
	[105]	✓					
	[19]	✓					
	[22]	✓					
	[63]	✓					
	[64]	✓					
	[88]	✓					
	[98]	✓					
	[115]	✓					
	[117]	✓					
	[153]	✓					
	[161]	✓					
	[121]	✓					
	[55]		✓				
	[12]			✓			
	[113]			✓			
	[125]			✓			
	[4]				✓		
	[82]				✓		
	[89]				✓		
	[96]				✓		
	[135]				✓		
	[152]				✓		
	[160]				✓		
	[100]					✓	

Table 2: Evaluation of the collected publications on the types of contributions, 2022.

Year	Reference	Propose methods	Theoretical frameworks	Analyses	Apply	Benchmark	Review
2022	[39]	✓		✓			
	[34]	✓		✓			
	[6]	✓					
	[25]	✓					
	[50]	✓					
	[62]	✓					
	[158]	✓					
	[83]	✓					
	[56]	✓					
	[79]	✓					
	[80]	✓					
	[90]	✓					
	[93]	✓					
	[106]	✓					
	[111]	✓					
	[132]	✓					
	[131]	✓					
	[144]	✓					
	[65]	✓					
	[101]		✓	✓			
	[24]		✓		✓		
	[70]		✓				✓
	[15]		✓				
	[16]		✓				
	[94]		✓				
	[118]		✓				
	[133]		✓				
	[157]		✓				
	[128]		✓				
	[149]			✓			
	[28]				✓		
	[109]				✓		
	[126]				✓		
	[48]					✓	✓
	[143]					✓	✓

Table 3: Evaluation of the collected publications on the types of contributions, 2023-2024.

Year	Reference	Propose methods	Theoretical frameworks	Analyses	Apply	Benchmark	Review
2023	[36]	✓	✓				
	[29]	✓	✓				
	[116]	✓	✓				
	[9]	✓		✓			
	[42]	✓		✓			
	[75]	✓		✓			
	[147]	✓		✓			
	[156]	✓			✓		
	[155]	✓			✓		
	[54]	✓					
	[123]	✓					
	[14]	✓					
	[72]	✓					
	[30]	✓					
	[51]	✓					
	[91]	✓					
	[92]	✓					
	[95]	✓					
	[108]	✓					
	[127]	✓					
	[129]	✓					
	[141]	✓					
	[154]	✓					
	[163]	✓					
	[164]	✓					
	[165]	✓					
	[78]	✓					
	[148]	✓					
	[74]	✓					
	[77]		✓				
	[124]		✓				
	[38]			✓			
	[57]			✓			
	[102]			✓			
	[3]				✓		
	[76]				✓		
	[8]					✓	
	[60]					✓	
2024	[21]	✓					
	[114]	✓					

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claim is that existing research on recourse is disconnected from the practical requirements of systems where it would be applied (see Section 4 and Section 5.1). Our claim is supported by a systematized literature review which is the contribution of this work (Section 3). These are reflected in the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We highlight and discuss the three main limitations of our work in Section 5.2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work, as a literature review, does not rely on theoretical results or proofs. Nonetheless, we are explicit about the "assumptions" in that we discuss our approach to the collection and analysis of results in depth in Section 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: Our work does not rely on any experiments, so this question is not applicable. Nonetheless, we believe that we provide sufficiently in-depth characterization of the review process where other authors should be able to reproduce it (Section 3 and Appendix A).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Our work does not rely on any experiments, so this question is not applicable. Nonetheless, we provide the complete list of publications covered in this review. We will also release the review data upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: Our work does not rely on any experiments, so this question is not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Our work does not rely on any experiments, so this question is not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Our work does not rely on any experiments, so this question is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and we confirm that our work conforms to it in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Although this is not covered in a separate section, positive and negative societal impacts of our work (and algorithmic recourse in general) are a key consideration throughout this paper. See for instance Section 1 or Section 6.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks, so this question is not applicable. We do not introduce any data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Our work does not use existing assets (other than the referenced papers), so this question is not applicable. All papers covered in the review are referenced in sufficient detail, so that the readers can access them.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our work does not release any new assets, so this question is not applicable. We release the paper with the most permissible license available for NeurIPS submissions. Finally, we will release the review data upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing or research with human subjects, so this question is not applicable. The work was in its entirety carried out by the authors.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing or research with human subjects, so this question is not applicable. We did not require an IRB approval or equivalent to carry out this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.