Counterfactual Explanations May Not Be the Best Algorithmic Recourse Approach

Sohini Upadhyay
Harvard University
Cambridge, Massachusetts, USA
supadhyay@g.harvard.edu

Himabindu Lakkaraju
Harvard University
Cambridge, Massachusetts, USA
hlakkaraju@hbs.edu

Krzysztof Z. Gajos
School of Engineering and Applied Sciences
Harvard University
Allston, Massachusetts, USA
kgajos@g.harvard.edu

Abstract

Algorithmic recourse is a rapidly developing subfield in explainable AI (XAI) concerned with providing individuals subject to adverse high-stakes algorithmic outcomes with explanations indicating how to reverse said outcomes. While XAI research in the machine learning community doesn't confine itself to counterfactual explanations, its algorithmic recourse subfield does, adopting the assumption that the optimal way to provide recourse is through counterfactual explanations. Though there has been extensive human-AI interaction research on explanations, translating these findings to the algorithmic recourse setting is non-obvious due to meaningful problem setting differences, leaving the question of whether counterfactuals are the most optimal explanation paradigm for recourse unanswered. While intuitively satisfying, the prescriptive nature of counterfactuals makes them vulnerable to poor outcomes when circumstances unknown to the decision-making and explanation generating algorithms affect re-application strategies. With these concerns in mind, we designed a series of experiments comparing different explanation methods in the recourse setting, explicitly incorporating scenarios where circumstances unknown to the decision-making and explanation algorithms affect re-application strategies. In Experiment 1, we compared counterfactuals with reason codes, a simple feature-based explanation, finding that they both yield comparable re-application success, and that reason codes led to better user outcomes when unknown circumstances had a high impact on re-application strategies. In Experiment 2, we sought to improve on reason code outcomes, comparing them to feature attributions, a more informative feature-based explanation, but found no improvements. Finally, in Experiment 3, we aimed to improve on reason code outcomes with a multiple counterfactual explanation condition, finding that multiple counterfactuals led to higher re-application success but still resulted in comparatively worse user outcomes in the face of high impact unknown circumstances. Taken together, these findings call into question whether the standard counterfactual paradigm is the best approach for the algorithmic recourse problem setting.

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712095

CCS Concepts

• **Human-centered computing** → **User studies**.

Keywords

algorithmic recourse, counterfactual explanations, AI explanations

ACM Reference Format:
Sohini Upadhyay, Himabindu Lakkaraju, and Krzysztof Z. Gajos. 2025. Counterfactual Explanations May Not Be the Best Algorithmic Recourse Approach. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3708359.3712095

1 Introduction

Motivated by AI policy mandates and notions of a "right to explanation," *algorithmic recourse* is a rapidly developing subfield in explainable AI (XAI) concerned with providing individuals subject to adverse high-stakes algorithmic outcomes with explanations indicating how to reverse said outcomes [16, 24, 50]. For example, suppose that an individual is denied a loan due to algorithmic decision-making. Algorithmic recourse aims to provide that individual with an explanation that both illuminates why their application was denied and enables them to reapply for that loan successfully. Additionally, algorithmic recourse is subject to an important constraint: institutions using algorithms for automated decision-making are concerned with applicants "gaming the system" [2] and, consequently, resist making their algorithms transparent. Thus, it is generally assumed that explanations used for algorithmic recourse need to be informative in the context of a specific case without revealing the complete decision-making logic. While the recourse problem has been extensively explored by the machine learning (ML) research community, it has been largely overlooked by human-AI interaction scholars, allowing untested assumptions to take root and under-evaluated solution paradigms with real world policy implications to be promoted.

As background, the bulk of ML recourse research has been devoted to developing efficient and robust methods for computing counterfactual explanations across different types of models [24, 50], in spite of theoretical concerns that counterfactuals may not support individuals seeking to reverse adverse outcomes [2]. Counterfactual explanations identify attribute changes needed to achieve an alternate, desired outcome. In our loan denial example, a counterfactual explanation of that algorithmic decision could be "the loan would be approved if annual income increased by $10k." What distinguishes recourse methods from other counterfactual explanation generation methods is a focus on finding useful, easily actionable

counterfactuals. This can range from prescribing minimal changes that would achieve the desired outcome, such as increasing income by $10k in lieu of $100k, to ensuring that prescribed changes are possible, i.e., decreasing duration of home ownership, a monotonically increasing temporal variable, is impossible. Researchers employ mathematical abstractions to capture salient feature information, like those aforementioned, with the aim of generating useful, actionable counterfactuals. As first highlighted in Barocas et al.'s expository analysis surfacing hidden assumptions about the use of counterfactuals [2], a challenge with the focus on these mathematical abstractions is that there may exist additional contexts not captured in the data available to the algorithm that substantially impact which courses of action are the best suited for the affected individual. To return to our loan denial example, a counterfactual explanation may say "the loan would be approved if income increased by $10k," unaware of childcare obligations preventing an individual from taking on more work shifts. If alternatives exist, a different set of financial recommendations may be more appropriate for their situation. For example, the person may be able to reduce how much of their credit limit they use by paying for some of their purchases with a debit card instead of a credit card. Thus, counterfactual explanations offer a clear suggestion for what a person might do to achieve a desired outcome, but if that one suggested path is not appropriate for that person to follow, counterfactual explanations might not offer enough information for people to know if/what alternative paths to success exist.

There is a clear gap in the literature given this background: are counterfactuals even the correct explanation paradigm to pursue for recourse? As we detail in related work, there is a substantial body of work on AI-assisted decision-making for settings where AI models recommend or justify an optimal course of action, including experiments that consider counterfactuals. The algorithmic recourse setting, while superficially similar, is meaningfully different because the human interacting with the explanation is a decision subject and not a decision-making worker. Specifically, in the recourse setting, counterfactuals serve two purposes: First, they provide an explanation for a negative decision that has already been taken by a different party and, second, they communicate a course of action for achieving different outcome next time. Because the audience of the explanation is the decision subject, there are also constraints on how much information recourse explanations can convey—institutional stakeholders employing decision-making algorithms (such as banks) are unlikely to provide complete information or algorithm access due to proprietary arguments and concerns about individuals "gaming the system" [2].

It is particularly important for HCI researchers to bridge this gap given that our field can and should contribute knowledge in areas where pertinent policy debates are taking place [57]; beyond generating new human-AI interaction insights, further empirical insights could support policy-makers' attempts to empower individuals adversely affected by algorithmic decision-making. Related HCI scholarship has largely focused on the broader field of algorithmic contestation — where adverse algorithmic outcomes may be addressed through a wider range of processes (e.g., appeals, algorithm abolition) than affected individuals accepting institutions' initial decisions and reapplying [22, 27]. At minimum, making research connections to the algorithmic recourse field put forth by the ML community could help policy-makers better parse, scope, and navigate scholarship relevant to empowering affected individuals.

With all this in mind, we conducted an empirical evaluation of whether counterfactuals are the correct explanation paradigm to pursue for algorithmic recourse. This paper is preliminary in its critical technical practice [4] in that we challenge the counterfactual explanation paradigm without challenging any of the other core assumptions of algorithmic recourse, such as not revealing the details of the decision-making algorithms. At a higher level, we are also operating under the traditional recourse assumption that applicants need to make changes to achieve a desired outcome — to assume otherwise (e.g., to demand changes in the logic of the decision-making algorithm or institution) broadens the problem setting to algorithmic contestation, which, while critical and interrelated, is outside the scope of this work.

To interrogate whether the counterfactual paradigm is the most appropriate for recourse, we compared the effects of various explanation types in the recourse problem setting. We designed a college internship reapplication task, conducting three experiments that consecutively build on each other. In each experiment, participants, acting as career counselors, selected courses that applicants (college students) should take before reapplying. Participants had access to the denial letter that included an explanation for why the original internship application was denied. In our task design, we explicitly incorporated circumstances unknown to the decision-making and explanation generation algorithms (the schedule of course offerings) into our reapplication task and we challenged our participants to create course plans that would both lead to a successful reapplication outcome and that could be completed in as few semesters as possible. We manipulated the course schedule design to create conditions where students' urgency and explanations' course recommendations were mutually compatible (the *aligned* condition) or at odds (the *misaligned* condition). Considering the relative strengths and limitations of the counterfactual explanations discussed prior, in our primary experiment we hypothesized that: H1) Employing counterfactual explanations would result in higher reapplication success than feature-based explanations; H2) However, employing counterfactual explanations would result in less optimal reapplication plans (in terms of the number of semesters needed to complete additional courses) compared to feature-based explanations in the misaligned condition, where circumstances unknown to the decision-making and explanation generation algorithms influence the reapplication process.

Support for H1 would bolster the counterfactual paradigm. Support for H2 would do the opposite, serving as an existence proof of when counterfactuals can lead to worse outcomes, namely in the face of circumstances unknown to decision-making and explanation-generation algorithms impacting the reapplication process. This pre-condition for worse outcomes is potentially widespread; despite ML efforts detailed in related work, no algorithmic decision-making or explanation-generating algorithm can plausibly capture all circumstances important to an individual in real-world scenarios.

In Experiment 1, we instantiated feature-based explanations with "reason codes" (simple lists of features that impacted the decision the most), observing no support for H1 and strong support for H2. Specifically, we found that both explanation formats yielded comparable reapplication acceptance, and counterfactuals led to

later (i.e., less optimal) application acceptances when unknown circumstances had a high impact in the misaligned schedule condition. This was our primary experiment designed around our core question of counterfactual use in algorithmic recourse. Subsequent supporting experiments built on the findings of Experiment 1 — that reason codes can outperform counterfactuals — and sought better recourse explanation alternatives than reason codes. In Experiment 2, we sought to improve on reason codes outcomes, comparing them to feature attributions, which unlike reason codes, additionally conveyed the magnitude of the relative feature importance. However, we found no significant improvements. In Experiment 3, we aimed to improve on reason codes outcomes with a multiple counterfactuals explanation condition, finding that showing multiple counterfactuals (in our case 2) led to higher reapplication acceptance but still resulted in later application acceptance compared to reason codes in the misaligned schedule condition. Pursuing these supporting experiments enabled us to enrich the recourse explanation design recommendations that initially emerged from our primary experiment.

Across all experiments, we collected short responses on how participants used the different types of explanations to shape their decisions, revealing that most participants focused on reapplying successfully over reapplying quickly in counterfactual-based explanation conditions, but considered both needs holistically more often in feature-based explanation conditions. We also collected self-reported explanation preferences, finding that reason codes were always the least preferred in each of the experiments, contrasting with participants' performance on objective metrics. In summary, in this work we make the following contributions:

- We conducted the first user study evaluating the untested assumption that counterfactuals are the best explanation format for the algorithmic recourse setting, amassing empirical evidence challenging the counterfactual paradigm, demonstrating that across objective metrics, simple feature-based explanations like reason codes can lead to comparable or better outcomes than counterfactuals.
- We followed up these findings with additional experiments aiming to identify better explanations for algorithmic recourse, finding no overwhelmingly better approach than reason codes, and moderate success with multiple counterfactuals. We consequently recommend the recourse research community to both explore solutions outside counterfactual paradigm, and while within the paradigm, focus on multiple counterfactual solutions.
- We designed a reapplication task where features unknown to the algorithm affect decision-making, a condition critical to ensuring successful recourse outcomes in real-world deployments. Subsequently, this task design serves as a foundational template for effective evaluation of future recourse explanation approaches.

2 Related Work

2.1 Algorithmic Recourse Landscape

Our work was explicitly motivated by the assumptions and questions first surfaced by Barocas et al. [2]. At the onset of their analysis, Barocas et al. consider the differences between counterfactuals and "principle reasons," or reason codes — a simple feature-based explanation approach employed in credit scoring that lists some features contributing to an algorithmic decision in order of importance. This theoretical analysis considering the potential pros and cons of one explanation approach versus another directly influenced the conditions and hypotheses we employed in our primary experiment, stated formally in Section 4. Similarly motivating, Sullivan and Verreault-Julien [43] argue that recourse should be framed as a recommendation problem, and through this lens, it may turn out that supporting individuals in re-application may require a different decision support framework than counterfactuals. Barocas et al. [2], and concurrent work by Venkatasubramanian and Alfano [49], also make the following counterfactual-specific observations: prescribed feature changes often incorrectly assume feature independence, the difficulty of feature changes cannot be determined from the data trivially, and counterfactuals make static, monotonic, and binary classification model assumptions. Some of these concerns are tackled in algorithmic works detailed below.

Since the publication of Wachter et al.'s seminal work introducing the idea of algorithmic recourse [51], over 350 counterfactual generating algorithms, many specifically tailored for recourse, have been proposed, surveyed, and taxonomized [24, 50]. While our work focuses on human-centered outcomes, relevant to these outcomes are algorithmic research directions that aim to make explanations useful to individuals. Many approach this objective indirectly, aiming to produce counterfactual explanations that recommend changes that are consistent with values and trends present in the data, or "realistic," with the implication being that realistic recommendations are more useful to individuals. To this end, some methods aim to generate counterfactuals that are on the underlying data-manifold [15, 23], while others employ causal modeling of the data space directly [25, 26]. Others incorporate salient feature information unrepresented in raw data, like monotonicity and immutability, into counterfactual generation optimization constraints [46, 54]. All of these approaches are vulnerable to inaccurate or incomplete knowledge about the data or circumstances relevant to any given individual.

Some recourse methods defer the issue of counterfactual usefulness to designing a human-centered cost function to incorporate into the counterfactual generation optimization objective, making it a problem for metric learning research, or other related CS subfields. To this end, Rawal and Lakkaraju [39] proposed using the Bradley-Terry method to learn a recourse cost function from crowd-sourced or user inputted pairwise feature comparisons, where comparisons inquire which feature is more difficult to change. This type of approach may be limited due to the number of comparisons necessary to learn an effective cost function. In GAM Coach, Wang et al. [54] introduced an integer programming-based approach to incorporate user preferences like feature value range and perceived difficulty to change a feature, but their approach is only applicable when generating counterfactuals for linear (or linearly approximated) decision-making algorithms, and makes the common but often unrealistic assumption of feature independence.

Other methods tackle counterfactual usefulness by presenting multiple counterfactuals at once, the implication being that individuals can choose to follow whichever one is best suited for their circumstances. These approaches often focus on ways to maximize

the "diversity" of counterfactual set, or how different counterfactuals are from one another, in effort to present the widest set of options [37]. Relatedly, some works propose search [41], or iterative preference refinement algorithms [53, 54] that operate on a set of multiple counterfactuals. One of the primary critiques of multiple counterfactual approaches is whether they can be adopted in any realistic application given institutional information constraints. Given enough examples, multiple counterfactuals can be leveraged to game the decision-making algorithm, finding superficial improvements external to the decision-making objective that enable re-application success, or in the extreme, recreate the decision-making rule, posing proprietary concerns [2, 54].

In the realm of user studies focused on algorithmic recourse, GAM Coach [54] was also proposed as a means to achieving positive human-centered outcomes. As discussed above, GAM Coach introduced a novel approach to incorporating user preferences into counterfactual generation iteratively over a set of multiple counterfactuals. Wang et al. also presented an extensive interface and tool for GAM Coach, enabling comparisons across the multiple counterfactuals it produces. Their user studies centered around subjective measures, finding that participants found their system useful and usable, and preferred personalized plans over default counterfactuals. While both this work and our own are primarily concerned with human-centered algorithmic recourse outcomes, we seek to answer fundamentally different and complementary questions. We are interested in interrogating and explicitly understanding the prerequisite question of whether counterfactuals are even the correct explanation paradigm to pursue for recourse, particularly while explicitly modeling scenarios where circumstances unknown to the decision-making and explanation generation algorithms affect re-application.

2.2 Comparing Counterfactuals with other Explanation Types

Barocas et al. [2] conclude their work with a call to engage directly with decision subjects to better understand what is useful for recourse. Keane et al. [29] surveyed over 100 counterfactual explanation generation methods, finding that only 21% of them have been user tested, often with limitations. Of studies empirically evaluating counterfactuals, even fewer made comparisons to other explanation types.

Some works outside the domain of algorithmic recourse provide evidence for the goal-directed benefits of counterfactuals, supporting our intuition behind H1 where we hypothesized that due to their explicit, prescriptive nature, counterfactuals will result in more re-application acceptances (stated formally in Section 4). Warren et al. [55] compared counterfactual explanations with causal ones, finding that counterfactual explanations improved users' predictions of AI decisions (system understanding) compared to no explanation baselines, but not when compared to causal explanations. In a complementary work leveraging the same task design, Celar and Byrne [11] performed additional analyses, finding that counterfactuals improved the accuracy of participants' own decision-making compared to causal explanations. Celar and Byrne related their work to psychological theories positing that counterfactuals are goal-directed towards future decisions and prompt individuals into considering both the facts and the alternative to reality at once, a richer mental representation state with more information than that prompted by causal explanations. The downside of the richer representation is a higher cognitive load, especially given evidence that individuals do not like cognitively effortful explanations [6]. Celar and Byrne's findings demonstrate that in the XAI domain, the cognitive costs of counterfactuals are outweighed by the goal-directed benefits of available explicit information.

Other findings indicated that feature-based explanations promote model understanding, supporting our intuition behind H2 (stated formally in Section 4), because model understanding could improve individuals' ability to make optimal choices when circumstances unknown to decision-making and explanation algorithms affect the re-application process. Namely, Wang and Yin [52] evaluated the effect of counterfactual and feature-based explanations on model understanding, uncertainty awareness, and trust, in familiar (recidivism prediction) and unfamiliar (forest cover prediction) domains. Effects were observed in the familiar domain, with feature-based explanations having a positive effect on the three aforementioned goals.

The findings of other works comparing counterfactuals with different explanation types [3, 14, 20, 32, 34, 47, 55, 58] were less closely tied to our hypotheses. These works produced mixed findings with respect to counterfactuals outperforming or underperforming other explanation types, on varied decision making tasks and metrics, reinforcing the need for usage-informed evaluation practices that has been well established in the broader XAI field [5, 33]. This further motivates the need for specific user studies of the recourse setting, where the decision making task, re-application under an explanation information constraint, has not been addressed.

2.3 Designing Effective AI-Assisted Decision Support

While the above works center on recourse and the comparative strengths and weaknesses of counterfactuals relative to other explanation types, other scholarship has interrogated broader and related research questions surrounding the effectiveness of AI explanations in supporting decision-making.

In a qualitative study, Yacoby et al. [56] investigated the effectiveness of counterfactual explanations in supporting pre-trial risk assessment, finding that judges, the decision-makers, initially misinterpreted and ultimately ignored counterfactual explanations, making decisions based on their domain expertise [56]. Other works systematically manipulated feature-based explanations [21, 30] and the stated confidence of recommendations [1] to demonstrate how AI explanation-assisted decision support can be misleading.

While these demonstrations of ineffective AI explanation-assisted decision support do not map cleanly to understanding the recourse problem setting, they are instances of more wide ranging human-AI interaction concerns regarding prescriptive AI-assisted decision support. HCI scholarship on effective AI-assisted decision support is extensive, with research identifying over- and under-reliance on AI recommendations and attempting to calibrate this reliance via various interventions. This includes attempts to align human-AI mental models [28, 38], promote cognitive engagement with the information provided [6, 19], or adapt the human-AI interaction

to the individual differences in reliance on AI [44]. Lee and Chew [32], introduced in the previous section, engaged with notions of over and under-reliance explicitly, finding that clinical experts and laypeople over-relied on "wrong" AI assessments of a physical condition less when presented with counterfactual explanations compared to feature based explanations [32]. Some scholars argue against providing explicit decision recommendations because they take agency away from human-decision makers and fail to promote the mental model alignment or cognitive engagement critical for effective decision-making. Instead, some of these scholars argue for adaptive support that is optimized for human learning and for the sub-processes involved in decision-making [7, 8, 19]. Others propose evaluative support that helps decision makers assess arguments for and against specific decisions [36]. In addition to the recourse and counterfactual explanation literature, we situate our study in this body of effective human-AI decision support work by drawing connections to these perspectives when interpreting our results.

3 Methods

We conducted one primary experiment and two supporting experiments that consecutively built on each other, first to interrogate whether the counterfactual paradigm best suits the recourse setting and then to identify the most effective alternatives. In this section we outline our task design and other methodological decisions common to all three experiments.

3.1 Task Design

We asked participants to put themselves in the shoes of a college career counselor and help students who were initially denied internships improve their resumes based on hiring manager feedback and reapply. For each task instance, participants were presented with a simplified resume (Figure 1) which showed an applicant's experience in four skill areas. The experience was visualized as a number of stars (0–5). Under the guise of hiring manager's feedback, participants were also shown an explanation for why a particular applicant was denied an internship (Figure 2).

Next, participants were shown a tool (Figure 3), which they had to use to suggest to an applicant what they should do to improve their skill set in order to be accepted for the internship the next time they applied. The tool presented a schedule of courses in the 4 skill areas. The schedule covered the next 3 semesters. Taking a course in a particular skill area would improve an applicant's corresponding skill level by one star. For most skills, courses were available in some semesters but not others. Additionally, participants could recommend a maximum of 4 courses (communicated in the interface as the number of available credits, Figure 3 top left). Underneath the schedule, participants were shown a reminder, reproducing the information they were told prior about the applicant and the explanation for the initial denial (Figure 3 bottom).

Participants were instructed to select a set of courses that would satisfy 2 criteria: First, if the student took the courses, their reapplication for the internship should be accepted. Second, participants were asked to find solutions that would allow a successful reapplication in as few semesters as possible. To explain the importance of the second criterion, participants were informed that the students they would be assisting were first generation students who were struggling to make ends meet and to afford rent near campus. If evicted, they would have to dropout and move back in with their families, putting their dreams of being the first college graduates in their families at risk. Participants were informed that their students had applied to these paid internships to further their academic goals and to meet their financial needs. To incentivize careful consideration of both criteria, participants were told that they would receive a bonus that depended on how quickly the applicant would be ready to reapply ($0.10 if all courses could be completed in 1 semester, $0.05 for 2 semesters, and $0.01 for 3 semesters) if the recommendation they offered led to a successful reapplication (i.e., no bonus would be offered for advice that did not lead to the applicant being hired for the internship).

We manipulated the course schedule design such that in some instances, directly following the advice contained in an explanation (i.e., choosing exactly the courses mentioned in the counterfactual or choosing courses aligned with the top 2 reason codes) would allow a student to reapply successfully in 2 semesters. We refer to these as the *aligned* condition because the algorithm's priorities aligned with the students' need to reapply as quickly as possible. In other instances — in the *misaligned* condition — choosing the courses implied by the explanation would require 3 semesters of additional coursework before the student could reapply while a 2-semester solution was possible. In Experiment 3 (where we explore using multiple counterfactuals) we will also introduce a semi-aligned condition.

Given this overarching task structure, each of our experiments employed a within-subjects design where conditions were born of 2 orthogonal factors: explanation type and schedule alignment. There were 2 explanation types and 2 or 3 alignment conditions per experiment, detailed in each of the experiment sections that follow.

In each experiment, every task was randomly assigned a unique student name, internship name, and set of skill names. Each task was also randomly assigned a unique permutation rearranging the skill order on the resume and schedule. These choices made the resumes and associated tasks appear outwardly different, while internally consistent and comparable. The order of conditions was randomized at the level of individual task instances as well.

3.1.1 Implementation Details. We generated a synthetic dataset to underlie our tasks. This synthetic dataset was created by exhaustively generating every possible 4 feature, 6 ordinal level data point, i.e., $\{[0,0,0,0],[1,0,0,0],\ldots,[5,5,5,5]\}$ because resumes had 4 skills, each with a 0–5 star rating. Each resume mapped to a data point in this dataset. We created our set of candidate names from the most popular baby names in preceding years, split evenly across males and females. We created our set of internship and corresponding skill names by referencing the names of majors and required subjects in university course catalogs.

Underlying application acceptance was a logistic regression classifier binning the resumes into accepted or denied groups. As our resumes were synthetic data points, the weights of the logistic regression classifier ware selected by pseudo-random uniform sampling without replacement. The pseudo-random seed was selected by grid-searching for attributes that would make the reapplication

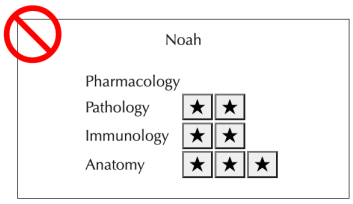

Figure 1: Example denied resume. The stars next to each skill name indicate experience in that skill area.

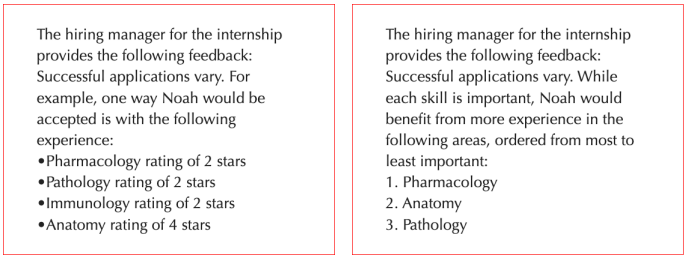

Figure 2: Counterfactual on left, reason codes on right.

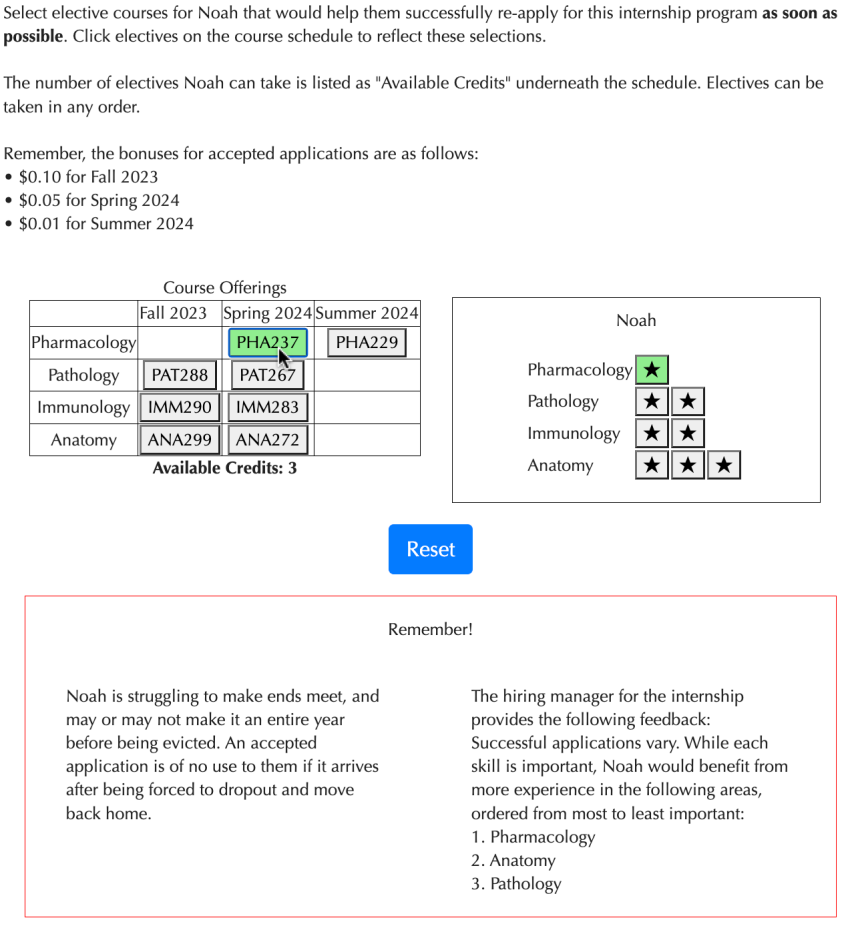

Figure 3: Schedule tool with reason codes reminder.

tasks uniform in difficulty across explanation conditions. Specifically, we sought to maximize the number of denied resumes with:

(1) No maximal 5 star ratings, to avoid making resume improvements too easy.
(2) A minimum of 3 star increments needed to change application decision, fixing task difficulty uniformly.

Ultimately, the resulting logistic regression weights were [0.8, 0.3, 0.2, 0.5]. Of the denied resumes, we chose the four that had the fewest number of ways to change the application decision (7 ways). We wanted denied data points with the fewest ways to change the application decision in order to avoid scenarios where acceptances could be achieved by making changes randomly. As discussed prior, we applied feature order permutations to generate additional tasks. Notice that the regression model only had access to the features captured by the resume data; the temporal urgency motivating the students was unknown. The same will hold for the explanation conditions described in the individual experiment sections that follow.

3.2 Procedures

3.2.1 Recruitment. Participants were recruited via Prolific. We targeted a US $12/hour rate before bonuses resulting in base payments of $4 for Experiments 1 and 2, and $4.50 for Experiment 3, based on median completion times. As described prior, we also paid participants a bonus for each accepted application contingent on the courses selected ($0.10 if all courses could be completed in 1 semester, $0.05 for 2 semesters, and $0.01 for 3 semesters).

3.2.2 Pre-Tasks. At the onset of the experiment, each participant was shown information about the study, followed by an informed consent form. Next we asked participants to answer optional demographics questions. Then participants completed an abbreviated (4-item) Need for Cognition questionnaire (as previously used in [18] and derived from [10]).

Next, participants were given detailed instructions about the task. At the end of these instructions, participants were asked a multiple choice question about the task objective (to apply both quickly and successfully). If answered incorrectly, participants were provided with answer feedback and asked to select the correct choice in order to proceed.

After the instructions, participants completed two practice application tasks. The practice tasks included one instance of each explanation type, and the most misaligned schedule condition to prime users with the more difficult tasks. The practice task order was randomized. Participants were prompted to try the practice tasks again, up to three times, if the application was denied or if they selected courses that could only be completed in three semesters. After each practice task, users were able to see if their final course selection resulted in successful reapplication. If the first practice task did not result in successful reapplication, the 3 try limit was removed - users had to complete the second task successfully before being allowed to proceed, and were given unlimited tries to do so.

3.2.3 Tasks. Then users were shown 3 application tasks per distinct explanation/alignment condition, in a randomized order, plus two attention checks disguised as tasks, appearing one third and two thirds of the way into the real tasks. In these attention check tasks, the explanation text was replaced with instructions to skip course selection for those tasks. This was meant to detect participants who selected courses arbitrarily without referring to the content in the explanation text.

3.2.4 Post-Tasks. After the last task, participants were asked a short form question about how they made their course selection for the final task (persisted on the screen). After completing this question, participants were informed of the number of reapplications that were successful and the cumulative bonus they earned. Next, participants were presented with a denied resume, both explanation formats, but no schedule, and asked to select which format of hiring manager feedback (explanation) they preferred, followed by a request to provide a brief explanation for their preference. Finally, participants were shown a form consisting of optional questions inquiring whether they encountered technical difficulties or cheated in some way, and then provided with a code to enter into Prolific to ensure payment.

3.3 Design & Analysis

3.3.1 Quantitative Methods. As previously mentioned, each of our experiments used a within subject design with 2 factors: explanation type and alignment. On an individual participant level, we measured: average application acceptance for each explanation condition and the average semesters taken to achieve acceptance. Given our task design, the minimum possible number of semesters was 2 and the maximum possible was 3. As the participants were aiming to select courses that would allow a student to reapply quickly and successfully, an average semesters of 2 reflects the best outcome and an average semesters of 3 reflects the worst outcome. For subjective preferences between the two explanation conditions, we report the percent breakdown of the self reported preferences. For supplementary analyses disaggregating these metrics based on Need for Cognition (NFC), refer to the appendix.

None of these metrics were normally distributed. Thus we used Wilcoxon Signed Rank tests to determine whether metric differences were statistically significant. We computed effect sizes as $r = \frac{Z}{\sqrt{n}}$ (where Z is the test statistic produced by the Wilcoxon signed rank test and n is the number of participants in the sample). This is a common approach for Wilcoxon non-parametric tests [17, 45]. The magnitudes of effect sizes computed this way are typically interpreted using Cohen's guidelines for r: .5 is interpreted as a large effect, .3 a medium effect, and .1 is a small effect [13]. As lack of evidence for an effect is not evidence for lack of an effect, we also compute 95% bootstrapped confidence intervals on 1000 bootstrap samples for r. Then in cases where we do not detect an effect, we are 95% confident that an effect would fall within that interval if it does exist.

We removed participants from our analyses if they failed an attention check. In measuring average application acceptance, the number of samples for our statistical tests was equivalent to the number of participants. In measuring average semesters for acceptance, some participants had null values for some explanation/alignment conditions because they had 0 applications accepted for those conditions. We used robust implementations of the Wilcoxon tests that dropped these samples and their pairs from the analyses. We report both the number of participants in our analyses (N) and the number

of non null pairs (N_w) powering Wilcoxon tests in our results. The participants for all three experiments are summarized in Table 1.

3.3.2 Qualitative Methods. Recall that post-tasks, participants were posed short answer questions inquiring about their course selection approach and explanation preference. The responses to these questions were open coded by the first author. The goal of this supplementary analysis was to further understand the reapplication strategies and explanation preferences driving our main quantitative results. While we conducted this analysis for all three experiments, we detail our findings most extensively for our first experiment and report on similarities and differences from those findings in subsequent experiments.

For the short response question asking how participants decided which courses to select on the last task, the short response text was coded alongside the participant's actual course selection, supplementing interpretation of their responses with observations of their actual choices. Codes identified course selection strategies, for example *followed counterfactual recommendation exactly* or *prioritized most important skill*. In our results, we detail these strategies and report on their relative prevalence across explanation conditions.

For the short response question about which explanation participants preferred and why, codes summarize explanation or task attributes participants' named and how those attributes contributed to their preference, for example *preferred counterfactual, more specific* or *preferred feature attribution, visual representation better*. In our results we convey the range of reasons why participants preferred one explanation over another, reporting on attributes participants described and how those attributes contributed to their preferences.

4 Experiment 1: Counterfactuals vs. Reason Codes

In this experiment, we evaluated whether employing counterfactuals leads to better recourse outcomes than feature-based explanations. Feature-based explanations—explored in depth by human-AI researchers in other areas of XAI but not much in the context of algorithmic recourse—are a popular, alternative class of AI explanations to counterfactuals. These explanations aim to indicate how each feature in the input data contributes to an algorithmic decision [12]. As initially debated by Barocas et al. [2], the rationale conveyed by feature-based explanations may enable individuals to better navigate situations where circumstances unknown to the decision algorithm affect their reapplication strategy. However, the lack of explicit guidance they offer could also lead individuals to construct unsuccessful reapplication plans. With this in mind, we hypothesized:

- **H1:** Employing counterfactual explanations would result in higher reapplication acceptance than feature-based explanations.
- **H2:** Employing counterfactual explanations would result in less optimal reapplication plans compared to feature-based explanations when circumstances unknown to the decision-making and explanation generation algorithms influence the reapplication process.

4.1 Conditions

Experiment 1 included 2 factors: explanation type (counterfactuals vs. reason codes) and schedule alignment (aligned or misaligned) for a total of 4 distinct explanation/alignment conditions.

4.1.1 Explanations. In all of our experiments we operated under the following minimal information constraint: explanations for recourse cannot reveal the complete decision-making rule, which would provide the decision-maker with the ability to assess the reapplication success of every possible course plan. Accordingly, in this experiment we instantiated feature-based explanations with "reason codes", as coined by Barocas et al. [2], a simple list-based explanation showing the features that contributed to the outcome the most, ordered by the magnitude of their impact. We chose reason codes as our starting point because we wanted to compare counterfactuals to a very rudimentary feature-based explanation, containing a low granularity of information in contrast with the explicit prescriptive content of counterfactuals. Beyond meeting our minimal information constraint, versions of reason codes are already widely employed by institutions in credit scoring [2], where they sometimes get referred to as "principal reasons".

Per our task design, explanations were presented under the guise of hiring manager feedback, either as a counterfactual (Figure 2 left) or reason codes (Figure 2 right). A counterfactual explanation showed one instance of how the application could be improved for the applicant to be hired. The reason codes, instead, provided a ranked list of three skills that the applicant should improve ordered by the relative importance of those skills. We chose a list length of three for representation brevity. We state at the explanation onset that each skill is important, so the fourth unlisted skill is understood to be least important implicitly.

4.1.2 Schedule Alignment. We manipulated the course schedule design to create two conditions where students' urgency and explanations' course recommendations were either aligned or misaligned. This enabled us to evaluate H2, that employing counterfactual explanations may result in worse user outcomes in the presence of circumstances unknown to the decision-making and explanation generation algorithms. Specifically, in the aligned condition, we fixed the course schedules such that the counterfactual recommendations could be completed within 2 semesters and the most important skill in the reason codes explanations had courses offered both in first and second semesters. In the misaligned condition, we designed course schedules such that the counterfactual recommendations would take 3 semesters to complete and the most important skill in the reason codes explanations was only offered in the second and third semesters, never the first. Given our task design, explained in the preceding methods section, across both aligned and misaligned schedules, there were 7 ways (i.e., combinations of courses) that would lead to successful reapplication within the 3 semesters. The aligned schedules allowed for all 7 routes to be reached in 2 semesters whereas misaligned only allowed for 1 route to be reached in 2 semesters.

4.1.3 Implementation Details. As we were interested in how users engage with the information contained in counterfactuals and reason codes, rather than specific algorithmic methods used to generate them, we employed the following strategies to construct accurate

		Experiment 1	Experiment 2	Experiment 3
N_{tot}		118	125	130
N		100	100	102
N_w	**Aligned**	94	94	93
	Semialigned	NA	NA	96
	Misaligned	82	93	79
Age		18-63, M=33, SD=9.9	19-66, M=35, SD=11.1	19-71, M=37, SD=13.4
Gender		female: 44 male: 53 non-binary: 2 not responded: 1	female: 52 male: 42 non-binary: 6 not responded: 0	female: 51 male: 48 non-binary: 2 not responded: 1

Table 1: Summary of participants. N_{tot} refers to the number of participants who completed the study, N refers to the number of participants who passed both attention checks and were included in the analysis, and N_w refers to the number of non null pairs in the Wilcoxon tests comparing average semesters taken for accepted applications.

explanations. For the counterfactual explanation condition, we grid-searched for the nearest accepted neighbor to the denied resume in the exhaustive synthetic resume dataset. This method mirrors recourse method growing spheres [31] and results in nearby, valid, on-manifold counterfactuals, a goal in many algorithmic recourse methods [24, 50]. For the reason codes explanation, we listed the top three most important features in order based on the logistic regression classifier weight magnitudes. The feature order permutations were applied so that explanations and tasks were internally consistent.

4.2 Results

4.2.1 Main Results. Our main results for Experiment 1 are summarized in Figure 4. Contrary to H1, we observe comparable application acceptance across both explanation conditions. The average application acceptance for counterfactuals (M=78.5%) was not significantly greater ($Z = -0.57, p = 0.57, r = -0.06, r_{CI} = [-0.25, 0.00]$) than the average application acceptance for reason codes (M=77.2%). Disaggregating further to aligned and misaligned conditions also yielded no significant differences in average application acceptance between counterfactuals and reason codes (details not reported).

We see strong support for H2, observing that counterfactuals yield comparatively worse user outcomes than reason codes in the misaligned schedule condition. For misaligned schedules, the average semesters taken when participants were presented with counterfactuals (M=2.90) was significantly greater ($Z = -5.51, p < 0.0001, r = -0.61, r_{CI} = [-0.71, -0.47]$), and thus less optimal, than the average semesters taken for the reason codes condition (M=2.47). For the aligned schedule condition, the average semesters taken for counterfactuals (M=2.00) was slightly but significantly lower ($Z = -3.69, p = 0.0002, r = -0.38, r_{CI} = [-0.46, -0.29]$), and thus more optimal, than the average semesters taken for reason codes (M=2.08).

4.2.2 Course Selection Approaches. Recall that after the last task we asked participants to fill in a short response indicating how they made their course selection for that task. Given our randomization of task order, we collected roughly equal amounts of responses for each explanation/alignment condition.

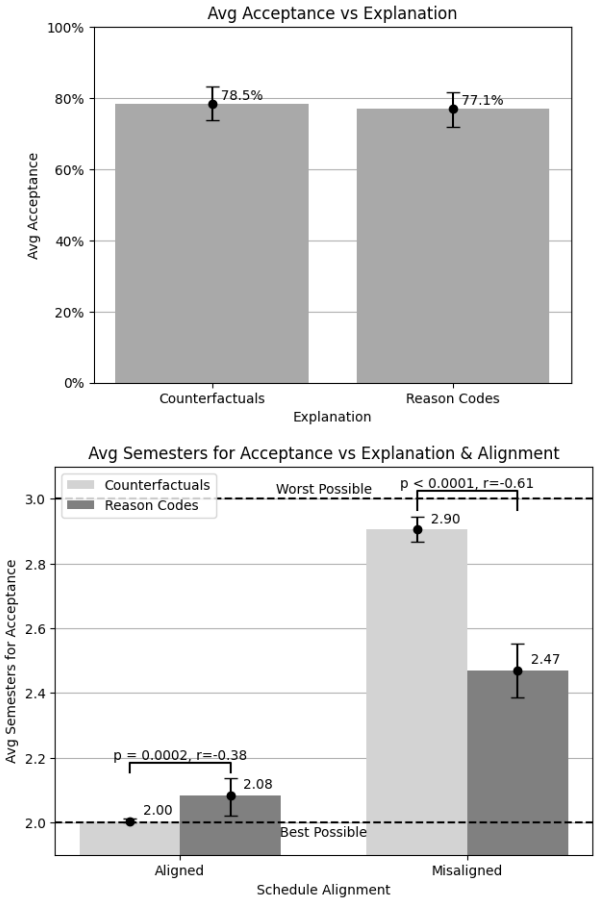

Figure 4: Experiment 1 results. Average application acceptance (top), average semesters taken to achieve acceptance (bottom). Error bars reflect 95% bootstrapped confidence intervals on 1000 bootstrap samples.

By analyzing their responses and their corresponding course selection, we observe that the vast majority of participants closely

followed the counterfactuals' course recommendations. The minority of participants who did not closely follow counterfactual recommendations prioritized time. Some partially followed the counterfactual explanations, and compensated by taking additional classes based on the applicant's original skill profile making statements like "I didn't want her to have to wait all the way until the end of Summer so I thought that having a higher rating in Geometry would accommodate for the lower rating in Stats." Others compensated by interpreting the magnitude of counterfactual skill improvements as skill importance, and selecting courses based on this perceived skill importance, for example saying "I think Criminal Law needs at least something, but I didn't want to wait until the third semester because she would lose her apartment. I tied [sic] to give Criminal Law as much as I could and allocated the remaining points to International Law and Civil Procedure since she was already strong and they requested 4 starts in Civil Procedure."

The majority of responses about reason codes indicated selecting the most important skill, using a plurality of their course credit limit in the aligned condition. In these aligned tasks, most participants treated the second and third important skills equivalently, prioritizing the second important feature over the third only in a minority of cases. In the misaligned condition, some participants continued to max out the most important skill, for example stating "I wanted to make sure she had all the classes she needed with the highest star rating to make sure she got the internship. I wasn't focused on how fast she could finish, but just if she could pass the internship requirements or not." This approach was largely outweighed by participants preferring to prioritize time, picking the top three skills and maximizing the second most important to do so. In these scenarios, participants would share explanations like "I decided to choose the electives that were happening as soon as they were available based on James most important ones. Time here was more important than the elective that's why I didn't choose summer courses for human computer interaction [most important] and opted to go with spring systems [second most important] instead." In other cases, participants would mention prioritizing time, and follow the ranking less closely after selecting the top skill. Some would then pick one course from each skill, making statements like "I had to balance out the requirements for the internship while accounting for getting her the internship as quickly as possible. so I gave her, her international law skill [most important] while buffing her already established skills in hopes of her getting accepted." Others would pick additional courses based on the applicant's original skill profile, sharing comments like "Algebra was most important so it needed another credit. Since that already would be taken Spring 24 I decided to add on the next important that had zero stars which was calculus." Overall, the temporal dimension of the task was mentioned more in comments pertaining to reason codes.

4.2.3 Self-Reported Preferences. In Experiment 1, 74% of participants reported preferring counterfactuals, while 26% reported preferring reason codes. This difference was statistically significant ($Z=-4.80$, $p<0.0001$, $r=-3.39$).

Analyzing the short responses explaining preference selection yielded insights into this phenomenon. Participants who preferred counterfactuals praised their specificity and appreciated seeing explicit numbers. They perceived counterfactuals as containing more information and conveying clearer expectations, thereby making course selection easier. To this end, participants commented "This format seems to me to provide a bit more information on what's required. It seems to be more specific," and "Knowing exactly and specifically, where the student could improve makes it so much easier to assign the classes knowing what they needed exactly."

Participants who preferred reason codes held different perspectives. Some found counterfactuals "to be more misleading," also commenting "[Counterfactuals] only shows an example of a successful application, leaving room for belief that there are other paths that could lead to acceptance as well. [Reason Codes] is more general, but still specifies which skills are the most important." Compared to those who preferred counterfactuals, participants who preferred reason codes indicated more concern for timing in their responses, stating "While it is not as precise as a star rating system [counterfactual condition], which gives a bit more exact of a weight than a ranking to each metric, it is also harder to fulfill course offerings that are not provided in a timely manner based on a weight system—if is [sic] easier to chose the next ranked class rather than choosing between two equally weighted metrics, or between two metrics that are weighted differently, but also offered in the course timeline differently." Relatedly, participants felt that reason codes gave them more autonomy in course selection, commenting "[Reason codes] feels more like advice, and it reads like I have more of a choice to how to pick my answer."

5 Experiment 2: Feature Attributions vs Reason Codes

In Experiment 1, we found that reason codes resulted in course selections that better aligned with applicants' schedules (in the misaligned conditions) while resulting in similar overall acceptance rates compared to the counterfactual explanations. We designed Experiment 2 to see if we could improve on reason codes with more informative feature-based explanations—feature attributions. Feature attribution explanations are a popular class of feature-based explanations that encompass methods like LIME, SHAP, and saliency maps [35, 40, 42]. Like the reason codes condition, feature attribution explanations indicate the order of feature importance for a decision. Additionally, beyond just rank, they typically convey how much more important each feature is relative to others. Given this additional information, feature attributions may help users better determine alternate paths to reapplication success when circumstances unknown to the decision-making and explanation generating algorithms affect reapplication. In order to meet our information constraint, recall that feature attribution methods for recourse cannot reveal the decision-making rule. This could be feasible by employing decision-boundary approximations (as in LIME, SHAP), local instead of global explanations, or withholding information.

5.1 Conditions

Experiment 2 included 2 factors: explanation type (feature attributions vs. reason codes) and schedule alignment (aligned or misaligned) for a total of 4 distinct explanation/alignment conditions. Reason codes explanation and schedule alignment conditions were identical to those in Experiment 1.

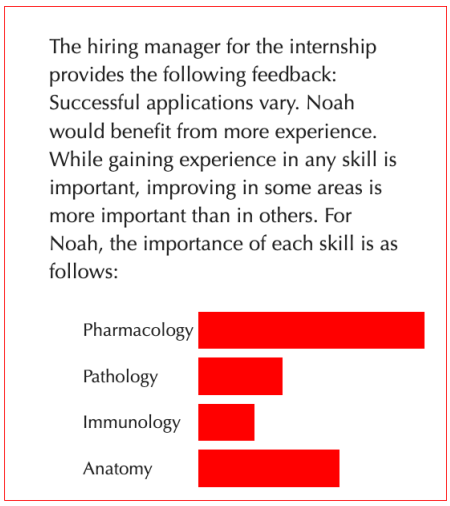

Figure 5: Feature attributions

5.1.1 Feature Attribution Instantiation. Mirroring representations of popular feature attribution methods like LIME [40] and SHAP [35], our feature attribution explanations represented each feature's contribution via bar graph (Figure 5). The length of each bar corresponds to the logistic regression weight for that feature, or skill area. Thus more important skills have longer bars and the relative importance of skills can be deduced by comparing bar lengths. While our instantiation of feature attributions did not involve an approximation of the decision boundary, we did withhold information. We did not display the numerical logistic regression weight associated with each bar, nor conveyed that logistic regression is underlying the decision-making, preventing participants from recreating the decision-making rule. Thus, this design met our minimal information constraint.

5.2 Results

5.2.1 Main Results. Our main results for Experiment 2 are summarized in Figure 6. Feature attributions did not improve on reason code outcomes. The average application acceptance for feature attributions (M=81.0%) was not significantly different ($Z = -1.30, p = 0.19, r = -0.13, r = [-0.31, 0.00]$) from the average application acceptance for reason codes (M=79.2%). Disaggregating further to aligned and misaligned conditions also yielded no significant differences between average application acceptance rates (details not reported).

For misaligned schedules, the average semesters taken for feature attributions (M=2.62) was significantly greater ($Z = -2.56, p = 0.01, r = -0.27, r_{CI} = [-0.44, -0.07]$), and thus less optimal, than the average semesters taken for the reason codes condition (M=2.52). For the aligned schedule condition, the average semesters taken for feature attributions (M=2.09) was not significantly different ($Z = -0.73, p = 0.47, r = -0.08, r_{CI} = [-0.27, 0.00]$) from the average semesters taken for reason codes (M=2.06).

5.2.2 Course Selection Approaches. In Experiment 2, participant explanations of course selection for both explanation reason codes and

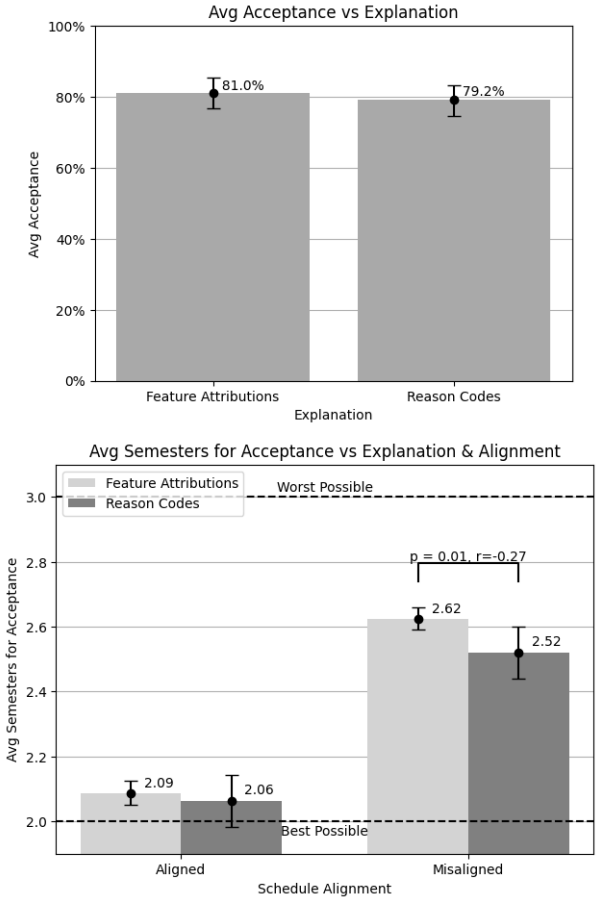

Figure 6: Experiment 2 results. Average application acceptance (top), average semesters taken to achieve acceptance (bottom). Error bars reflect 95% bootstrapped confidence intervals on 1000 bootstrap samples.

feature attributions, were consistent with reason codes comments in Experiment 1, primarily describing attempts to select the top skill and select courses within two semesters. While the strategies described were similar across explanation types, the distribution of strategies appeared different. In Experiment 2, participants reported prioritizing the top features more in feature attributions tasks compared to reason codes tasks. This seemed to be motivated by the relative skill importance conveyed in the feature attributions conditions, as participants made comments like "Astronomy [top skill] is almost double the importance of the others so I took 2."

5.2.3 Self-Reported Preferences. In Experiment 2, 76% of participants reported preferring feature attributions and 24% reported preferring reason codes. The average feature attributions preference rank was significantly higher than average reason codes preference rank (Z=-5.2, p<0.0001, r=-3.68).

Participants that preferred feature attributions explained that visual representations were easier to interpret, sharing "I like the

look of graphs. It's very quick and easy to pick up on the information needed which makes the task quicker and more efficient." Participants also often commented that the feature attributions helped them make more informed decisions through conveying the relative importance of skills, for example remarking "The bars are useful because they they say more about just how much more or less important the priorities are. Knowing that calculus was twice as important as stats, I probably would've asked Henry to finish the calc series instead of introducing stats."

Participants who preferred the reason code explanations liked their concise list representation, making statements like "I prefer to have ordered lists. They help me stay organized and focused." Some of these participants found the feature attribution visualizations confusing or misleading, commenting "The relative rankings on the right [feature attributions] are more confusing than the left [reason codes] and also harder to remember" and "[reason codes] is a more useful format because it shows a ranking without it having such drastically different lengths in value that it may skew decision making. For example in the [feature attributions] format, robotics is so much farther ahead that it might make you think you NEED to take summer 2024 classes just to appease it if they otherwise did not have experience in robotics, but that risk is not worth overall eviction."

Some participants found both explanation formats comparable, commenting "To be honest, both worked almost equally. One is easier to read instantly (a) [reason codes], while the other has a visual [feature attributions]."

6 Experiment 3: Reason Codes vs Multiple Counterfactuals

In Experiment 1, we found that reason codes yielded comparable or better recourse outcomes compared to counterfactuals. In Experiment 2, we tried to improve on reason codes outcomes with feature attributions to no avail. We designed Experiment 3 with the same objective, to see if we could improve on reason codes with a more informative counterfactual option. In light of research advocating for the use of multiple counterfactuals being presented for algorithmic recourse [37, 54], we choose to compare reason codes to a multiple counterfactuals explanation. Multiple counterfactuals convey fundamentally more information than a singe counterfactual, which could help users better understand the landscape of reapplication acceptances and better navigate scenarios when circumstances unknown to the decision-making and explanation generating algorithms affect reapplication. As detailed in related work, sufficiently large sets of counterfactuals can be employed to recreate the decision-making rule. To meet our information constraint, we instantiate multiple counterfactuals minimally, with two counterfactuals.

6.1 Conditions

Experiment 3 included 2 factors: explanation type (multiple counterfactuals vs. reason codes) and schedule alignment (aligned, semi-aligned, misaligned) for a total of 6 distinct explanation/alignment conditions. Reason codes were identical to those in Experiments 1 and 2.

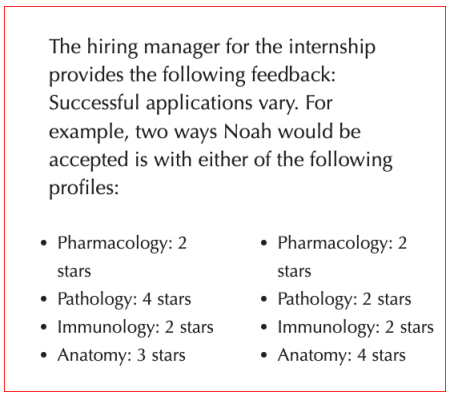

Figure 7: Multiple counterfactuals

6.1.1 Multiple Counterfactuals Instantiation. We showed two counterfactuals side by side for the multiple counterfactual condition (Figure 7). For each task, one counterfactual was the same used in Experiment 1, and the second was selected a priori from the remaining 6 successful solution options. The second counterfactual was selected to accommodate the new schedule alignment conditions detailed below. The display order of the two counterfactuals was randomized.

6.1.2 Schedule Alignment. We adapted our schedule alignment conditions as follows. The aligned and misaligned conditions closely paralleled those employed in Experiments 1 and 2. For the aligned condition, we designed our course schedules such that either counterfactual recommendations could be completed within two semesters and the top two reason code skills were offered in the first and second semesters. For the misaligned condition, we designed our course schedules such that either counterfactual recommendations could only be completed in 3 semesters and the most important reason code skill was only offered in the second and third semesters, never the first. Additionally, we added a third semialigned condition, where one of the counterfactual recommendations could be completed in 2 semesters and the other could only be completed in 3. Also in the semialigned condition, the most important reason code skill was offered in the first 2 semesters but the second most important was only offered once in the last semester. These conditions enabled us to vary the impact of schedule alignment more systematically for this new experiment.

6.2 Results

6.2.1 Main Results. Our main results for Experiment 3 are summarized in Figure 8. Multiple counterfactuals partially improved on reason code outcomes. The average application acceptance for multiple counterfactuals (M=88.1%) was significantly greater ($Z = -5.17, p < 0.0001, r = -0.51, r_{CI} = [-0.65, -0.36]$) than the average application acceptance for reason codes (M=72.3%). This trend persisted even when disaggregating further to aligned, semialigned, and misaligned conditions, detailed in the Appendix.

For misaligned schedules, the average semesters taken for multiple counterfactuals (M=2.95) was significantly greater ($Z = -6.67, p < 0.0001, r = -0.75, r_{CI} = [-0.82, -0.68]$), and thus less optimal, than

the average semesters taken for the reason codes condition (M=2.36). For the semialigned schedule condition, the average semesters taken for multiple counterfactuals (M=2.13) was significantly less ($Z = -4.37, p < 0.0001, r = -0.44, r_{CI} = [-0.57, -0.30]$), and thus more optimal, than the average semesters taken for reason codes (M=2.34). For the aligned schedule condition, the average semesters taken for multiple counterfactuals (M=2.00) was not significantly different ($Z = -1.0, p = 0.32, r = -0.10, r_{CI} = [-0.21, 0.00]$) from than the average semesters taken for reason codes (M=2.01).

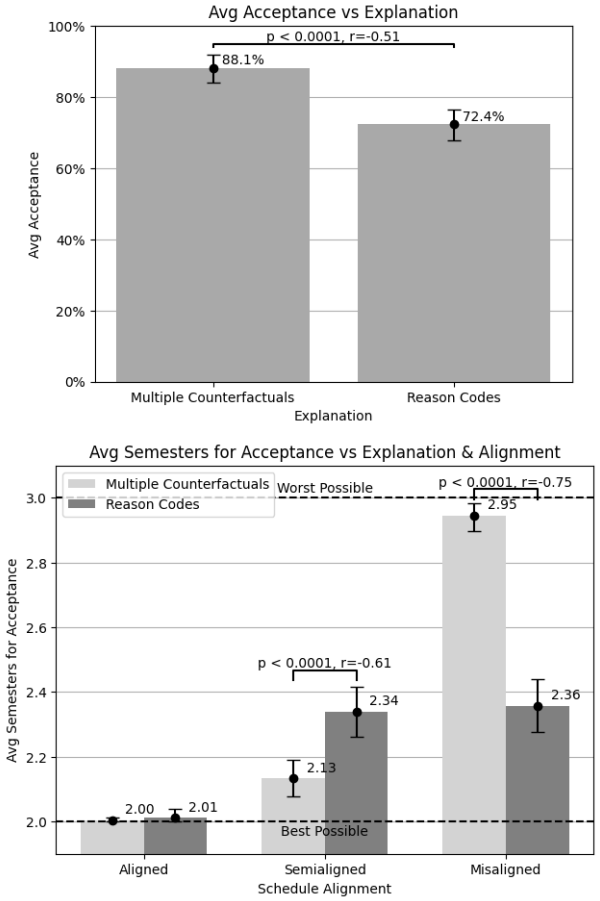

Figure 8: Experiment 3 results. Average application acceptance (top), average semesters taken to achieve acceptance (bottom). Error bars reflect 95% bootstrapped confidence intervals on 1000 bootstrap samples.

6.2.2 Course Selection Approaches. The strategies described for multiple counterfactuals were similar to those described for counterfactuals in Experiment 1. In the semi-aligned condition, specific to Experiment 3, participants preferred to follow the counterfactual recommendations that could be completed within 2 semesters, making statements like "I looked at the two paths recommended by the hiring manager and especially prioritized getting him graduated before the end of spring 2024. He may be evicted, so getting him graduated earlier was better." In the misaligned condition, some participants indicated feeling no choice but to follow one of the multiple counterfactual recommendations, making comments like "I chose based on the fact that no matter what Harper would be risking coming close to being evicted due to calculus not being available in the fall semester." Across the alignment conditions, there was some evidence of participants trying to interpolate information about application acceptance by cross-referencing both counterfactuals. For example, one participant explained "I wanted at least one point in microeconomics since it may be important but not too much. I then put the leftover points in econometrics and statistics because they were rated overall more important by the managers."

Understandably, in Experiment 3, the course selection strategies described for reason codes were also consistent with those reported in Experiments 1 and 2. In the semi-aligned condition, participants overwhelmingly prioritized course selections that could be complete within two semesters, deliberately choosing to skip over the second most important skill offered later in the schedule, with participants sharing explanations like "I didn't want her to take a course too late so I decided to skip that one [second most important skill] and instead delegate her credits to the next 2 most important courses."

6.2.3 Self-Reported Preferences. In Experiment 3, 75% of participants reported preferring multiple counterfactuals and 25% reported preferring reason codes. The average multiple counterfactuals preference rank was significantly higher than average reason codes preference rank (Z=-4.95, p<0.0001, r=-3.5).

The short responses explaining these preferences were consistent with those in Experiment 1. Like those who preferred counterfactuals in Experiment 1, participants who preferred multiple counterfactuals in Experiment 3 praised their specificity, perceived them as containing more information or guidelines for application acceptance, and found they made course selection easier. Likewise, those who preferred reason codes shared the rationale of their counterparts in Experiment 1, finding reason codes more flexible or amenable to timely acceptance, and multiple counterfactuals confusing or misleading.

7 Discussion

We interpret our results with the following focal points in mind. We designed our experiments such that the metric of average semesters for acceptance more accurately reflected an applicant's holistic desired outcome compared to the less granular application acceptance metric. Given that algorithmic recourse methods are envisioned for use in high stakes applications, and given the near impossibility of algorithms knowing all circumstances relevant to individuals' decision making processes in said settings, the misaligned schedule condition likely reflects a more realistic scenario than the aligned schedule condition. Accordingly, in interpreting our results, we center average semesters for acceptance results in the misaligned schedule conditions. As we detail below, reason codes, not any of the counterfactual based explanation conditions, consistently performed the best at this metric, and were even comparable to counterfactuals on the less granular application acceptance metric, forming a systematic pattern of counterfactual underperformance

that calls for a serious re-imagination of explanation paradigms for algorithmic recourse.

In our primary experiment, Experiment 1, we found no evidence for H1 — compared to reason codes (a simple feature-based explanation), counterfactuals did not offer advantages in terms of the rate of reapplication acceptance. We speculate that H1 was unsupported because participants were capable of effectively interpreting the reason codes; participants' short responses about their use of reason codes indicated that most participants tried to select coursework in the most important skill area and turned to the next important skill areas when selecting the most important skill conflicted with applicants' urgency. We did, however, observe support for H2, finding that counterfactuals were significantly more prone to worse outcomes, as measured by average semesters for acceptance, when circumstances unknown to the decision-making and explanation algorithms had a high impact on the reapplication process, simulated by our misaligned schedule condition. Shedding light on this outcome, we observed that students' urgency in reapplication was emphasized more in short responses pertaining to reason codes than counterfactuals and considering this dimension was critical to achieving optimal outcomes in the misaligned schedule condition.

Both the quantitative result supporting H2, and participants' short responses clarifying this outcome, align with Celar and Byrne's work demonstrating that psychological theories of counterfactuals' goal-directed benefits hold in the broader XAI domain [11], fixating focus on what the explanation was designed for (reapplication success), and with Barocas et al.'s theorizing that counterfactuals and feature based explanations have different advantages [2]. Given Barocas et al.'s theorizing [2] and Wang et al.'s findings indicating that feature-based explanations promote model understanding [52], we can also read this result as a confirmation of perspectives against prescriptive AI-assistive decision support if we link model understanding to more comprehensive human reasoning. The explanation that was less prescriptive and informative with respect to reapplication success but promoted more comprehensive human reasoning (reason codes) better supported applicants than the explanation that was more prescriptive and informative with respect to reapplication success but promoted less comprehensive human reasoning (counterfactuals). Interpreted in terms of over-reliance, our findings suggest that individuals over-rely on counterfactual explanation recommendations in the misaligned condition. This is in stark contrast to Lee and Chew's [32] finding that counterfactual explanations reduce over-reliance relative to feature-based explanations in clinical decision-making, likely due to the meaningful problem setting differences that distinguish algorithmic recourse.

Outside of negative results, in Experiment 1, the first redeeming finding for counterfactuals was that there was some evidence (medium effect size) that when circumstances unknown the decision-making and explanation algorithms have a low impact on the reapplication process, as simulated by our aligned schedule condition, counterfactuals lead to better outcomes than reason codes, as measured by semesters for acceptance. Unfortunately, this can rarely be guaranteed in real-world settings, and given the negative objective outcomes (large effect size) associated with misalignment, the drawbacks of counterfactuals appear more impactful. The second redeeming finding for counterfactuals was that they were significantly preferred over reason codes by participants. While this doesn't mitigate their negative objective outcomes, the contrast between the objective and subjective findings is consistent with results discussed in related work comparing counterfactuals to other explanation formats. We speculate that this could be because reason codes require more cognitive steps to act on and individuals often dislike cognitively effortful interventions in AI-assisted settings [6].

While people supported by reason codes performed as well or better than when they were supported by counterfactuals in Experiment 1, there was still room for improvement—average application acceptance for reason codes in the misaligned condition was M=77% and the average semesters taken to achieve acceptance was M=2.47 (with the best possible being 2, and the worst possible 3). This motivated us to search for ways to improve on reason codes. To this end, Experiment 2 employing feature attributions was unsuccessful: Compared to reason codes, feature attributions resulted in comparable application acceptance (no significant difference) and significantly worse or comparable outcomes as measured by semesters for acceptance in the misaligned and aligned conditions respectively. As captured in participants' short responses, these results are likely because course selection approaches across both explanation conditions were comparable, with the relative skill importance conveyed by feature attributions occasionally influencing participants to prioritize the top features more, which would result in poor outcomes in the misaligned schedule condition. This result also lends weight to arguments for aligning AI decision support with human cognition [7, 19], as the more informative but misleading feature attributions under-performed relative to reason codes, which were less informative but promoted more comprehensive human reasoning. Similar to Experiment 1, subjective measures contradicted objective measures, with participants significantly preferring feature attributions over reason codes, with short responses indicating that this trend arose from visual information representation preferences and appreciation for the relative importance information conveyed in feature attributions, despite acknowledgements that they likely made similar decisions when shown either explanation.

In Experiment 3 we made a final attempt to improve on reason codes by comparing them to multiple counterfactuals. This time we observed a significant improvement in application acceptance, with multiple counterfactuals outperforming reason codes, M=88.1% to M=72.3%, and the significance of this trend persisted even when disaggregating by schedule alignment. On the flip side, consistent with Experiment 1 findings, the multiple counterfactual condition led to significantly worse outcomes, as measured by average semesters for acceptance, when circumstances unknown to the decision-making and explanation algorithms had a high impact on the reapplication process. On a more positive but less impactful note, multiple counterfactual condition led to significantly better outcomes, as measured by average semesters for acceptance, semi-aligned condition. These trends could be attributed to short responses indicating that participants overwhelmingly followed the course recommendations of one of the multiple counterfactuals, even in the misaligned condition, where they felt like they had no choice not to, despite applicants' competing need for urgency. Like in Experiment 1, and likely for similar reasons, subjective measures conflicted with objective outcomes, with participants significantly preferring multiple counterfactuals to reason codes.

7.1 Generalizability & Limitations

Despite taking care to evaluate on an actual decision-making task (reapplication) in lieu of proxy tasks or subjective measures (e.g., knowledge comprehension, self-reported preferences) that may lead to misleading results [5], our work is still prone to the limitations of crowd-sourced empirical HCI work situated in the West. We employ a synthetic dataset, simple logistic regression, and our users are not the individuals actually facing application denials. It's not obvious how our findings would be affected by varying task characteristics like data complexity and explanation accuracy. Accordingly, significant positive results may not generalize to specific domains, geographies, or complex, deeply personal, real world decision-making.

However, our calls to re-examine counterfactual use rest on significant negative results. While significant positive results may not generalize, significant negative results serve as a compelling existence proof of how counterfactual explanations may fail in recourse settings. This is because real world scenarios would amplify the complexities that make our relatively simple and streamlined task challenging, namely the number and impact of features unknown to the algorithm affecting decision-making.

8 Conclusion & Future Work

Our findings lead us to make three main recommendations for algorithmic recourse explanation development. First, we urge that new recourse explanation approaches are evaluated on reapplication tasks with conditions where there are features unknown to the algorithm that affect decision making, in order to effectively understand the utility of the explanation. Explanations need to be useful across these circumstances to ensure successful recourse outcomes in real-world deployments and future work can leverage our task design as a starting point. Second, we urge against exclusively confining recourse explanation development to counterfactuals and instead encourage innovating new explanation modalities. Given concurrence between our findings and perspectives in AI-assisted decision support arguing against providing more but perhaps misleading information in lieu of promoting more comprehensive human reasoning, designers should be mindful of the pitfalls of explanation prescriptiveness when exploring new explanation modalities. Given the relative success of reason codes, interactive feature-based explanations should also not be discounted. Third and finally, we recommend that when exploring new recourse explanation methods within the counterfactual paradigm, to focus on multiple counterfactual solutions. Based on our partially positive multiple counterfactual results in Experiment 3, more advanced, interactive multiple counterfactual methods like GAM Coach [54] are promising candidates for such evaluation, but may need to grapple with institutional information constraints more explicitly.

Orthogonally, our findings also suggest a limit to the utility of explanations in assisting individuals in reversing adverse outcomes, indicating that it may be productive for future research (and for policy-making) to consider more expansive conceptions of the algorithmic recourse problem, including challenging its other underlying assumptions and shifting attention towards algorithmic contestation. While reason codes and multiple counterfactuals matched or outperformed counterfactuals on the evaluated metrics, their performance was still suboptimal on the misaligned condition, the setting designed to most accurately simulate the real-world and its unknowns. In this regard, our findings contribute to the body of human-AI interaction research demonstrating limitations to AI-explanation assisted decision-support [21, 30, 56]. These limitations also lend weight to the suggestion posed by Karusala et al.'s [27] work on algorithmic contestation, namely that the very framing of algorithmic recourse as an AI-explanation assisted reapplication task may be a poor conception of the actual issues individuals impacted by adverse algorithmic decisions face, and addressing those issues may require more complex socio-technical solutions than explanations [27]. Our work challenged the assumption that counterfactuals are the best explanation format for recourse, but not the assumption that institutional information constraints exist, precipitating the need for explanations over simple algorithm access that would allow individuals to evaluate the success of any possible path. Deeper engagement with these information constraints is an important research direction (and a topic that may benefit from a policy rather than technical intervention), as trying to empower individuals affected by adverse algorithmic outcomes while balancing institutional information constraints may be a solution paradigm that consistently favors institutions given broader societal power structures.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2107391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank the Intelligent Interactive Systems Group at Harvard for their insightful discussions throughout the course of this project and Zana Buçinca for her feedback on the manuscript.

References

[1] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (April 2023), 17 pages. https://doi.org/10.1145/3579460

[2] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 80–89.

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[4] Kirsten Boehner, Shay David, Joseph Jofish 'kaye, and Phoebe Sengers. 2004. Critical Technical Practice as a Methodology for Values in Design. https://api.semanticscholar.org/CorpusID:1738633

[5] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. ACM, New York, NY, USA.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. https://doi.org/10.1145/3449287

[7] Zana Buçinca, Alexandra Chouldechova, Jennifer Wortman Vaughan, and Krzysztof Z. Gajos. 2022. Beyond End Predictions: Stop Putting Machine Learning First and Design Human-Centered AI for Decision Support. *NeurIPS Human-Centered AI Workshop (HCAI)* (2022).

[8] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Susan A. Murphy, and Krzysztof Z. Gajos. 2024. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning. (2024).

[9] John T. Cacioppo and Richard E. Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42, 1 (1982), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

[10] J T Cacioppo, R E Petty, and C F Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

[11] Lenart Celar and Ruth MJ Byrne. 2023. How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition* (2023), 1–16.

[12] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 370 (oct 2023), 32 pages. https://doi.org/10.1145/3610219

[13] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd edition ed.). Lawrence Erlbaum Associates.

[14] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[15] Michael Downs, Jonathan L Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. 2020. Cruds: Counterfactual recourse using disentangled subspaces. *ICML WHI* 2020 (2020), 1–23.

[16] European Parliament and Council of the European Union. [n. d.]. *Regulation (EU) 2016/679 of the European Parliament and of the Council.* https://data.europa.eu/eli/reg/2016/679/oj

[17] Catherine O Fritz, Peter E Morris, and Jennifer J Richler. 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General* 141, 1 (2012), 2.

[18] Krzysztof Z. Gajos and Krysta Chauncey. 2017. The Influence of Personality Traits and Cognitive Load on the Use of Adaptive User Interfaces. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. ACM, New York, NY, USA, 301–306. https://doi.org/10.1145/3025171.3025192

[19] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[20] Lujain Ibrahim, Mohammad M Ghassemi, and Tuka Alhanai. 2023. Do Explanations Improve the Quality of AI-assisted Human Decisions? An Algorithm-in-the-Loop Analysis of Factual & Counterfactual Explanations. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 326–334.

[21] Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S. Valley, Ella A. Kazerooni, Nikola Banovic, Jenna Wiens, and Michael W. Sjoding. 2023. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. *JAMA* 330, 23 (12 2023), 2275–2284. https://doi.org/10.1001/jama.2023.22295

[22] Nari Johnson, Sanika Moharana, Christina Harrington, Nazanin Andalibi, Hoda Heidari, and Motahhare Eslami. 2024. The Fall of an Algorithm: Characterizing the Dynamics Toward Abandonment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM. https://doi.org/10.1145/3630106.3658910

[23] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615* (2019).

[24] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.* 55, 5, Article 95 (dec 2022), 29 pages. https://doi.org/10.1145/3527848

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

[26] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems* 33 (2020), 265–277.

[27] Naveena Karusala, Sohini Upadhyay, Rajesh Veeraraghavan, and Krzysztof Z. Gajos. 2024. Understanding Contestability on the Margins: Implications for the Design of Algorithmic Decision-making in Public Services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 478, 16 pages. https://doi.org/10.1145/3613904.3641898

[28] Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Kate Glazko, Matthew Lee, Scott Carter, Nikos Arechiga, Haiyi Zhu, and Kenneth Holstein. 2023. Training Towards Critical Use: Learning to Situate AI Predictions Relative to Human Knowledge. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 63–78. https://doi.org/10.1145/3582269.3615595

[29] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *arXiv preprint arXiv:2103.01035* (2021).

[30] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[31] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2017. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443* (2017).

[32] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. 7, CSCW2, Article 369 (Oct. 2023), 22 pages. https://doi.org/10.1145/3610218

[33] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.

[34] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[35] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[36] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[37] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[38] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 36. 5323–5331.

[39] Kaivalya Rawal and Himabindu Lakkaraju. 2020. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems* 33 (2020), 12187–12198.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[41] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.

[42] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 611–636.

[43] Emily Sullivan and Philippe Verreault-Julien. 2022. From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 712–722.

[44] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2025. Personalising AI assistance based on overreliance rate in AI-assisted decision making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. ACM Press, New York, NY, USA. https://doi.org/10.1145/3708359.3712128

[45] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences* 1, 21 (2014), 19–25.

[46] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*. 10–19.

[47] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[48] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (2018), 1080–1088.

[49] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 284–293.

[50] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. 2020. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).
[51] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
[52] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*. 318–328.
[53] Yongjie Wang, Qinxu Ding, Ke Wang, Yue Liu, Xingyu Wu, Jinglong Wang, Yong Liu, and Chunyan Miao. 2021. The Skyline of Counterfactual Explanations for Machine Learning Decision Models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 2030–2039. https://doi.org/10.1145/3459637.3482397
[54] Zijie J Wang, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. 2023. GAM Coach: Towards Interactive and User-centered Algorithmic Recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[55] Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. 2023. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 171–187. https://doi.org/10.1145/3581641.3584090
[56] Yaniv Yacoby, Ben Green, Christopher L Griffin Jr, and Finale Doshi-Velez. 2022. "If it didn't happen, why would I change my decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 219–230.
[57] Qian Yang, Richmond Y. Wong, Steven Jackson, Sabine Junginger, Margaret D. Hagan, Thomas Gilbert, and John Zimmerman. 2024. The Future of HCI-Policy Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 820, 15 pages. https://doi.org/10.1145/3613904.3642771
[58] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

Appendix

A.1 Audit for Intervention Generated Inequalities

AI explanation interventions, like any technology design decision, can introduce intervention generated inequalities [48] if they prove more useful to some groups over others. This is particularly concerning when the group that benefits is already privileged in a particular setting. While audits for intervention generated inequalities often disaggregate results based on demographic traits, following the precedents set by [6], we disaggregate our results based on Need for Cognition (NFC), a stable personality trait indicating how much individuals engage in and enjoy effortful cognitive activities [9]. We think this is useful for our experiments because interpreting explanations requires cognitive effort, and the ways in which individuals may benefit from different explanations may vary across NFC [19].

We collected NFC, measured on a 1–5 scale, at the onset of each of our experiments. We divided participants into the high NFC group (NFC>3.5) and the low NFC group (NFC$\leq$ 3.5), which resulted in roughly even groups for each experiment. We reproduced our analyses for each experiment, this time disaggregating results by level of NFC. With one exception discussed below, the aggregate and disaggregated NFC group trends were identical for each experiment, with comparable p-values and effect sizes (details not reported).

The one exception was that in Experiment 2 we observed that for misaligned conditions, the average number of semesters taken for application acceptance, when presented with feature attributions, was significantly greater, and thus less optimal, than the average number of semesters taken when presented with reason codes. After disaggregating by the level of NFC, we found that this held true for the high NFC group, (Z=2.77, p=0.003, r=0.38) but was not significantly greater for the low NFC group (Z=0.49, p=0.31, r=0.08). For misaligned conditions in the high NFC group, the average semesters taken for feature attributions and reason codes were M=2.60 and M=2.45 respectively. For misaligned conditions in the low NFC group, the average semesters taken for feature attributions and reason codes were M=2.66 and M=2.62 respectively. The lower average semesters taken (more optimal) in the high NFC group, suggests that the high NFC group benefited from reason codes compared to feature attributions, while the low NFC group the did equally poorly in both conditions.

In line with our interpretation of Experiment 2 results in Section 7, we suspect that this is because the high NFC group paid more attention to the relative importance information in feature attributions and was thus more prone to prioritizing the top features at the cost of poor outcomes in the misaligned schedule condition.

A.2 Experiment 3 : Disaggregated Average Application Acceptance

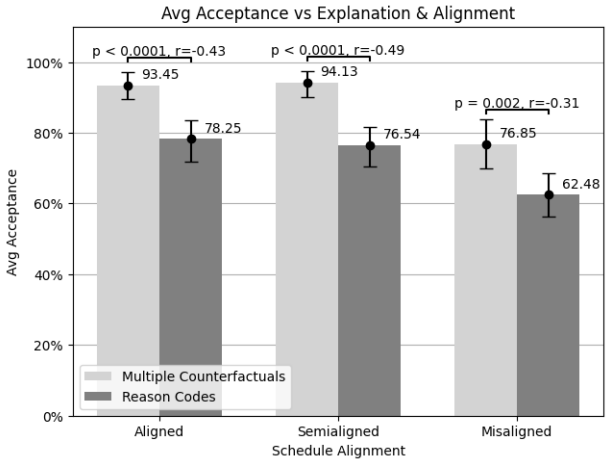

Figure 9: Experiment 3, disaggregated average application acceptance. Error bars reflect 95% bootstrapepd confidence intervals on 1000 bootstrap samples.

The average application acceptance for multiple counterfactuals was significantly greater than the average application acceptance for reason codes, even when disaggregating to aligned, semialigned, and misaligned conditions, as summarized in Figure 9.