Toward System-Level Algorithmic Recourse: Capacity, Competition, and Evaluation

by

Zahra Khotanlou

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2026

© Zahra Khotanlou 2026

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Decision makers are increasingly relying on machine learning in sensitive, high-stakes settings. In such contexts, *algorithmic recourse* aims to provide individuals with actionable and minimally costly steps to reverse unfavorable AI-driven decisions.

While existing research predominantly focuses on single-individual (i.e., *seeker*) and single-model (i.e., *provider*) scenarios, real-world applications involve multiple interacting stakeholders operating under shared resource constraints. Optimizing outcomes under an *individual welfare* perspective overlooks this inherently multi-agent nature, where individuals interact and compete for limited resources. Individually computed recourse recommendations no longer remain valid in such settings: when multiple seekers simultaneously compete for limited provider capacity, recommendations that appear optimal in isolation can become infeasible at the system level, and stakeholder interactions directly affect recourse validity.

To address this limitation, we introduce a novel framework for multi-agent algorithmic recourse that explicitly models system-level interactions among multiple seekers and providers. We formulate this many-to-many setting as a *capacitated weighted bipartite matching problem*, where matches are determined by recourse costs and constrained by provider capacities. Edge weights, derived from recourse costs, enable the system to optimize *social welfare* while quantifying the *welfare gap* between individually optimal outcomes and those achievable under resource constraints.

Building on this formulation, we propose a three-layer optimization framework. First, we compute welfare-maximizing assignments under fixed capacity constraints via capacitated matching. Second, we optimize the distribution of provider capacities to minimize the welfare gap, aligning system resources with seeker preferences. Third, we incorporate cost-aware adjustments by penalizing deviations from existing capacity configurations, enabling practical deployment under real-world operational constraints.

We further extend the framework to support inequality-averse objectives through a concave social welfare formulation that prioritizes the most disadvantaged seekers, allowing a principled trade-off between efficiency and equity. In addition, we present a unified perspective and reproducible benchmarking framework for evaluating recourse methods, enabling consistent comparison across system-level and individual-level criteria.

Experimental results on synthetic and real-world datasets demonstrate that our framework achieves near-optimal social welfare with minimal modifications to system configurations. The results also highlight an inherent efficiency–fairness trade-off, showing how recourse systems can be designed to balance aggregate welfare with distributive considerations in resource-constrained environments.

Overall, this work extends algorithmic recourse from isolated individual recommendations to system-level design, providing a tractable and principled approach for improving collective outcomes while preserving individual actionability.

Acknowledgements

This thesis marks the end of a chapter in my life and the beginning of a new journey. Along the way, I have been supported by people whose presence gave me strength, even in moments when the path felt uncertain.

I would like to express my appreciation to my supervisor, Prof. Amir-Hossein Karimi, for his guidance throughout this journey. His perspective and direction have greatly helped to the development of my research and my experience during my graduate studies.

To my parents and my brother, thank you for being my constant source of love and support. Although I spent many days far from you during this journey, I never truly felt alone. Your belief in me, your encouragement, your prayers, and your presence in my life have been my greatest strength.

Most importantly, I am grateful to God for all the blessings in my life, and for guiding every second of my life and helping me grow into a better person.

To my friends and peers, thank you for the conversations, kindness, and moments of light that made this path brighter.

Dedication

To my family, for their endless love and sacrifices.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements v

Dedication vi

List of Figures ix

List of Tables xi

1 Introduction 1

1.1 Motivation 1

1.2 Algorithmic Recourse 2

1.3 Objectives and Organization 6

2 From Individual to Multi-Agent Recourse 8

2.1 Introduction 8

2.2 Related Work 11

2.2.1 Multi-Agent Recourse 11

2.2.2 Multi-Agent Systems and Resource Allocation 12

2.2.3 Social Welfare and Inequality Aversion 14

3 Capacitated Bipartite Matching for Social Welfare Optimization **17**
3.1 Many-to-Many Recourse Optimization . . . 17
3.2 Minimizing the Welfare Gap . . . 20
3.2.1 Capacity Redistribution with Penalizing Modifications . . . 22
3.3 Inequality-Averse Recourse Optimization . . . 24
3.4 Experiments . . . 25
3.4.1 Hyperparameter Sensitivity Analysis . . . 29
3.5 Discussion . . . 32
3.6 Conclusion . . . 33

4 Unification and Evaluation for Recourse **35**
4.1 A Unified Formalization . . . 36
4.2 A Reproducible Benchmarking Framework . . . 40

5 Summary and Future Work **43**
5.1 Summary . . . 43
5.2 Future Directions . . . 45

References **49**

APPENDICES **62**
A.1 Proof of Theorem 3.4.1 . . . 62
A.2 Provider Classification Performance . . . 64
A.3 Additional Experiments . . . 66

List of Figures

2.1 Various algorithmic recourse setups: (a) the classical one-to-one setting, where an individual s seeks recourse recommendations from a provider p with minimal cost required to reverse the output; (b) the many-to-one setting, where multiple individuals are seeking recourse from a single recourse provider. (c) Our proposed many-to-many framework generalizes prior settings by simultaneously optimizing for multiple recourse seekers and providers. . . . 9

3.1 The effect of capacity distribution on social welfare outcomes. The pink dashed plot (with shaded region) shows the welfare achieved under an arbitrary capacity allocation (Equation (3.1)), which may result in a non-zero welfare gap even when sufficient total capacity is available. The dark red plot represents the optimized distribution obtained by solving Equation (3.2), which maximizes social welfare under a total capacity constraint. The blue horizontal line shows the global individual-optimal welfare, attainable only when total capacity is unconstrained. The gap between the pink and red curves highlights the inefficiency introduced by uncoordinated capacity allocations. 21

3.2 Comprehensive illustration of our proposed framework on the Two Moon dataset and using the MACE method. (a) model decision boundaries, (b) weight matrix showing seeker-provider preferences and individually optimal matches, (c) welfare curve showing how total capacity impacts social welfare and its gap from the individual-welfare upper bound, (d) and the step-by-step optimization of provider capacities, (e) transitioning from initial assignments, to welfare-maximizing allocations, and finally (f) balancing welfare gains with capacity adjustment costs 27

3.3 Trade-off analysis on the COMPAS dataset with high sensitivity ($\gamma = 50$). a Total Social Welfare shows a minor decline as the system becomes more fair (lower α). (b) The Equity Metric $\mathcal{E}$ (Minimum Utility) significantly improves, validating that the concave formulation effectively protects the worst-off seekers. 30

3.4 Hyperparameter sensitivity analysis on the large-scale Credit setting ($n = 200$, $m = 15$). (a) As γ increases, $\text{SW}_{\text{Eq.1}}/\text{IW}$ exhibits a U-shaped trend, initially decreasing due to capacity-induced displacement and later recovering as the matching becomes dominated by maximum-weight edges. The stable elbow at $\gamma \in [50, 150]$ motivates our choice of $\gamma = 100$. (b) $\text{SW}_{\text{Eq.3}}/\text{IW}$ interpolates smoothly between Equation (3.2) at $\beta = 0$ and Equation (3.1) at $\beta \approx 0.40$. 32

A.1 Illustration of our proposed framework: Weight matrix and individually optimal matches (a), Welfare curve showing the gap between social and individual optima. (c–e) Optimization process: transitioning from initial assignments (c) to welfare-maximizing allocations (d), and balancing welfare gains against capacity adjustment costs (e). 66

List of Tables

3.1 Comparative Welfare attainable ratio (SW/IW) on three different datasets of the three optimization formulations. 29

4.1 An overview of recourse algorithms for consequential decision-making settings is presented. Ordered chronologically, we summarize the *goal*, *formulation*, *solution*, and *properties* of each algorithm. Symbols are used to indicate supported settings in the experimental section of the paper (●), settings that are natural extensions of the presented algorithm (◍), settings that are partially supported (◐), and settings that are not supported (○). The models cover a broad range of tree-based (TB), kernel-based (KB), differentiable (DF), or other (OT) types. Actionability contraints (unconditional or conditional), plausibility constraints (domain-, density-, and prototypical-consistency), and additional constraints (diversity, sprasity) are also explored. While the primary datatypes used in consequential settings are tabular ▦ (involving a mix of numeric, binary, categorical, and ordinal variables), we also include additional works that generate recourse for non-tabular (images 📷 and document 🗎) datasets. Furthermore, papers that present analysis of such properties as optimality (opt.), coverage (cov.), and run-time complexity (rtm.) are specified in the table. Finally, we make note of those papers that provide *open-source* implementations of their algorithm. 39

A.1 Classification performance of the 15 providers on the held-out test set (large-scale DiCE setting). Predictions use each provider's tier-specific threshold. 65

A.2 Comparative results of the three optimization formulations. Each cell shows social welfare (SW), capacity vector $\mathbf{k}$, and the percentage of the individual-welfare upper bound (IW) attained in each optimization problem. 67

Chapter 1

Introduction

1.1 Motivation

Machine learning systems are increasingly deployed to inform consequential decisions that fundamentally shape individuals' lives, determining credit eligibility [61], employment opportunities [113], healthcare access [108], educational admissions [37], and criminal justice outcomes [95]. As these automated decision-making systems transition from experimental tools to operational infrastructure providing resource allocation and opportunity distribution, they raise pressing questions about transparency, accountability, and individual agency. When an algorithm denies a loan application, rejects a job candidate, or recommends pretrial detention, affected individuals face not only an adverse outcome but also a critical informational asymmetry: they often cannot understand why the decision was made, nor what actions they might take to achieve a favorable outcome in the future [37, 108].

This challenge has motivated two interrelated research agendas within explainable artificial intelligence (XAI). The first seeks to answer "why?", providing post-hoc explanations that illuminate the reasoning behind algorithmic decisions through techniques such as feature attribution, saliency maps, and local approximations [37, 108]. The second, more action-oriented agenda addresses "what if?" and "how?", offering *algorithmic recourse* that guides individuals toward actionable interventions capable of reversing unfavorable predictions [113, 108, 37, 59, 95]. Algorithmic recourse, rooted in foundational work on counterfactual explanations [113], represents a paradigm shift from *passive understanding* to *active empowerment*, transforming explanations from diagnostic tools into prescriptive recommendations [61, 95].

Since these early formulations, the algorithmic recourse literature has grown substantially, expanding into causality, robustness, fairness, privacy, and personalization [59, 108]. Yet across this growth, the dominant framing has treated recourse as a private interaction between a single individual and a single decision-making system. While a growing body of work has begun to move beyond this one-to-one view, offering population-level recourse summaries [95], studying collective action in online settings [23], and formulating group counterfactual explanations [29], these efforts share a common limitation which they do not model the system-level resource allocation problem that arises when multiple seekers simultaneously compete for multiple providers with finite capacity. When recourse operates at scale, individuals compete for limited provider capacity, their collective actions shift data distributions, and the decisions made for one person affect the recourse available to another [32, 64]. Addressing recourse as the population-level, resource-constrained problem it truly is requires a different formulation. Before developing that formulation, we first establish the technical foundations of algorithmic recourse as it has been studied in the individual setting.

1.2 Algorithmic Recourse

Algorithmic recourse has been defined in the literature as "an actionable set of changes a person can undertake in order to improve their outcome" [60], "the ability of a person to obtain a desired outcome from a fixed model" [107], and "the systematic process of reversing unfavorable decisions by algorithms across a range of counterfactual scenarios" [59]. Across these definitions, the common thread is *actionability*: unlike post-hoc explanations that describe why a decision was reached, recourse prescribes what an individual can concretely do to change it [113, 108, 37]. The field sits at the intersection of explainable AI, causal inference, and optimization, and has grown substantially since its early formulations into a research area that addresses causality, robustness, fairness, privacy, and personalization [59, 108].

The intellectual roots of algorithmic recourse lie in the intersection of counterfactual reasoning and data protection law. Wachter et al. [113] argued that the discourse around a "right to explanation" under the EU General Data Protection Regulation (GDPR) [36] had focused narrowly on opening the black box of algorithmic systems. They proposed a more tractable alternative: rather than explaining how a model works, data controllers should offer *unconditional counterfactual explanations*, statements describing the smallest change to an individual's features that would have produced a different, more favorable outcome, without requiring any disclosure of the model's internal logic. From the perspective of affected individuals, they identified three aims such explanations should serve: to inform

why a decision was reached, to provide grounds to contest it, and to understand what would need to change to receive a desired result in the future. Formally, given a classifier $h : \mathcal{X} \to \{-1, +1\}$ and an individual x with $h(x) = -1$, a counterfactual explanation is found by solving:

$$x'^{\star} \in \arg\min_{x' \in \mathcal{X}} \ \text{dist}(x, x') \quad \text{s.t.} \quad h(x') = +1, \tag{1.1}$$

where $\text{dist}(\cdot, \cdot)$ is a proximity measure such as weighted squared Euclidean distance [113]. The central design principle is *proximity*: the recommended change should be as small as possible to minimize the individual's effort.

While this formulation was foundational, it had a critical limitation: it identifies *where* an individual needs to arrive but says nothing about whether getting there is possible. A counterfactual that recommends reducing one's age, altering a protected attribute, or moving to a practically unreachable region of feature space may be mathematically valid but entirely useless as a guide for action [61, 95]. Ustun et al. [107] formalized this gap with the concept of *actionable recourse*: the ability of a person to change the decision of a model by altering *actionable* input variables. They drew an explicit distinction between features a person can meaningfully change (e.g., income, years of employment) and those that cannot or should not be changed (e.g., age, gender, birthplace), and introduced integer programming tools to generate *flipsets* — minimal-cost sets of feasible actions that guarantee a positive classification outcome. Their empirical analysis revealed that recourse availability and cost are substantially affected by modeling decisions such as regularization and feature selection, making actionability an auditable property of deployed models.

Building on these foundations, the standard recourse problem is formalized as a constrained optimization over a feasible action set [59, 107]. Let $a \in \mathcal{A}$ denote an action vector representing changes to the individual's features, and $A(x) \subseteq \mathcal{A}$ the set of feasible actions given domain constraints. The recourse problem seeks the minimum-cost action that flips the prediction:

$$a^{\star}(x) \in \arg\min_{a \in A(x)} \ \text{cost}(a;\, x) \quad \text{s.t.} \quad h(x + a) = +1, \tag{1.2}$$

where $\text{cost}(a;\, x)$ quantifies the difficulty of performing action a. The optimal value $c^{\star}(x) = \text{cost}(a^{\star}(x);\, x)$ is the *recourse cost* for individual x. The cost function is typically instantiated as a weighted ℓ_p norm, $\text{cost}(a; x) = \sum_k w_k |a_k|^p$, where $p = 1$ encourages sparsity — recommending changes to as few features as possible — and $p = 2$ produces smoother recommendations [113, 59]. Ustun et al. [107] propose percentile-shift measures that normalize changes by the population distribution, so that moving from the 50th to the 55th percentile of a feature is treated as less costly than moving from the 90th to the 95th, a personalized notion of effort that has influenced subsequent work on personalized recourse costs [104, 47].

The feasible action set $A(x)$ is the key mechanism for encoding real-world constraints. Karimi et al. [59] identify three categories of features that determine its structure: *actionable* features the individual can directly change (e.g., bank balance, years of employment); *mutable but non-actionable* features that cannot be changed directly but may change as a downstream consequence of other interventions (e.g., a credit score that improves following debt repayment); and *immutable* features such as birthplace, age, or protected attributes that cannot be changed under any circumstances and are excluded from $A(x)$ entirely. Additional constraints can encode directionality (e.g., income can only increase), bounds, and conditional feasibility [107, 59].

A key conceptual distinction in the recourse literature, formalized by Karimi et al. [59], is between *actionability* and *plausibility*. Actionability constrains the action set to changes the individual can actually perform. Plausibility constrains the post-action state $x + a$ to lie on or near the empirical data manifold: the resulting profile should resemble realistic individuals in the population, not merely satisfy the model's decision boundary [59, 108]. The two concepts are distinct and can conflict: a feasible action may produce an implausible post-action state, and a plausible state may not be reachable by any feasible action. Most recourse methods must balance both simultaneously, and several approaches have proposed using generative models to operationalize plausibility [83, 59].

A deeper tension concerns the relationship between the feature changes a model recommends and their causal consequences in the real world. Karimi et al. [59] draw a fundamental distinction between *contrastive explanations*, which answer "what would have been different?" and report the minimal feature change needed to flip the prediction, and *consequential recommendations*, which answer "what should the individual actually do?" and model actions as interventions in a causal system. Both can be framed as counterfactual reasoning in the philosophical sense [75], but they differ critically in their assumptions about feature dependence. Early methods following Wachter et al. [113] treated features as independently mutable, ignoring structural dependencies between them. Karimi et al. [61] showed that this can produce recommendations that are ineffective or counterproductive: recommending an increase in income without accounting for its causal effect on debt-to-income ratio and credit score may flip the model's prediction in theory but not in practice. Their *causal algorithmic recourse* reformulates the problem using structural causal models (SCMs), representing actions as formal interventions $do(a)$ in a causal graph and optimizing over downstream effects [61, 60]. A key negative result is that without full specification of causal relations, recourse validity cannot be guaranteed [59]. Subsequent work has extended causal recourse to partial causal knowledge [60], unobserved confounding [72], and improvement-focused objectives that distinguish boundary-crossing from genuine real-world improvement [71], with a growing body of work proposing frameworks that explicitly account for downstream

causal consequences [70, 27, 81].

The recourse literature has converged on a set of properties that high-quality recommendations should satisfy [59, 108, 37], and evaluating methods along a common rubric is a prerequisite for meaningful comparison [108, 82]. *Validity* is the most basic: the post-action state must receive a positive prediction from the model [113, 107]. *Actionability* requires the recommended changes to lie within $A(x)$, respecting immutability, directionality, and domain-specific constraints [107, 59]. *Plausibility* requires the post-action state to resemble realistic individuals, preventing exploitation of boundary artifacts [83, 59]. *Sparsity* favors changing as few features as possible, since multi-feature interventions are cognitively demanding [113, 107]. *Diversity* generates multiple distinct valid recommendations so individuals can choose the path best suited to their circumstances [83, 59]; Mothilal et al. [83] address this with DiCE (Diverse Counterfactual Explanations), which uses determinantal point processes to generate counterfactuals that are maximally spread in feature space while remaining valid. *Robustness* requires the recommendation to remain valid under perturbations to the individual's features, to execution noise, or to future model changes [87, 106, 111]; Pawelczyk et al. [87] show that many methods produce recommendations indistinguishable from adversarial examples and therefore brittle to small input changes, while Pawelczyk et al. [89] introduce a probabilistic robustness notion that explicitly navigates the trade-off between recourse cost and the probability of invalidation under distributional shift. These properties frequently trade off against each other: robustness typically increases cost [89], diversity can reduce plausibility [83], and causal grounding can increase cost [61]. Moreira et al. [82] empirically document disagreement across methods on nearly all properties, showing that which method performs best depends critically on which property is prioritized.

A consistent finding across the literature is that optimizing individual recourse cost does not guarantee equitable outcomes at the population level. Ustun et al. [107] showed that common modeling decisions can systematically affect recourse cost and availability differently across subpopulations, even when aggregate accuracy is good. Karimi et al. [59] formalize this concern: actionability constraints such as non-decreasing age impose differential costs across groups in ways that reflect and reinforce structural inequities. Perello et al. [92] show that standard recourse formulations can generate discrimination even when the underlying classifier is group-fair, due to the interaction between cost structure and population heterogeneity, while Kügelgen et al. [72] demonstrate that cost-based recourse can embed structural inequities even when classifiers are causally fair. Several approaches have responded by incorporating group fairness constraints directly into the recourse formulation [123, 31, 76].

This connects algorithmic recourse to the broader literature on welfare-based fairness. Heidari et al. [46] propose evaluating algorithms through a family of cardinal social welfare

functions motivated by the Rawlsian veil of ignorance, showing that lower-bounding these functions provides the first computationally tractable mechanism for controlling individual-level inequality in algorithmic outcomes, parameterized by a concavity coefficient $\alpha \in (0, 1)$ that interpolates between the utilitarian sum ($\alpha \to 1$) and the Rawlsian max-min ($\alpha \to 0$). Bertsimas et al. [9] characterize the efficiency cost of this shift as the *price of fairness*, showing it is bounded by a closed-form expression in the number of agents, establishing that fairness and efficiency need not be in severe conflict. Bell et al. [7] argue that minimizing individual recourse cost is insufficient when individuals face systematically different opportunity structures, and Gao et al. [34] show that large-scale recourse implementation induces measurable distribution shifts and social segregation, demonstrating that population-level consequences of individual recourse are real, not merely theoretical. Together, these findings establish that recourse is not only an individual problem but a distributional one, and that designing systems that are both individually valid and collectively equitable requires moving beyond the one-to-one framing that has dominated the literature.

1.3 Objectives and Organization

The technical foundations established in the previous section reveal a structural gap in the recourse literature. Most of the traditional recourse studies assume a single individual and a single decision-making system interacting in isolation from everyone else who faces the same situation. This assumption is convenient analytically but rarely holds in practice. In high-stakes domains such as lending, hiring, and admissions, multiple individuals simultaneously seek recourse from multiple providers, each operating under its own decision model and finite capacity to grant favorable outcomes. When provider capacity is limited, recourse recommendations compete for scarce resources: the system can no longer deliver the individually optimal outcome to every seeker at once, and the resulting welfare loss cannot be recovered by any improvement to how individual recourse is computed. It can only be addressed by treating the allocation of recourse as a system-level problem.

In this study we have developed a principled framework for this system-level problem: given a population of recourse seekers and a set of capacity-constrained providers, how should recourse recommendations be coordinated to remain valid and maximize social welfare? The framework is agnostic to the choice of individual recourse method, it operates as a coordination layer on top of any existing recourse infrastructure, taking as input only the matrix of pairwise recourse costs already defined in (1.2).

The remainder of the thesis is organized as follows. Chapter 2 introduces the multi-agent recourse setting formally. It opens by surveying related work on multi-agent recourse,

resource allocation, and social welfare theory, then defines the many-to-many matching problem, the welfare gap, and the bipartite graph structure that connects individual recourse costs to the system-level allocation problem. Chapter 3 presents the full technical contribution: a three-layer optimization framework for social welfare maximization, an efficient algorithm for optimal capacity redistribution, an inequality-averse extension for equitable allocation, and comprehensive empirical validation on synthetic and real-world datasets. Chapter 5 summarizes the contributions and outlines directions for future research, including game-theoretic extensions, dynamic and temporal settings, and fairness-aware multi-agent recourse design.

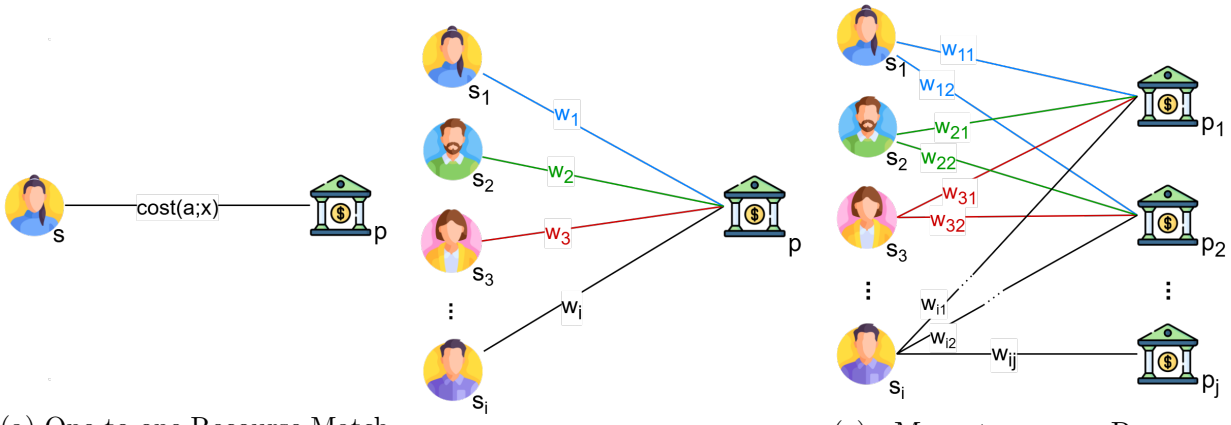
Chapter 2

From Individual to Multi-Agent Recourse

Individual recourse, as formalized in Section 1.2, treats each affected person in isolation: a single seeker, a single provider, and a cost to be minimized. This framing is analytically clean but practically incomplete. In realistic deployment, multiple individuals simultaneously seek recourse from multiple decision-makers, each with its own model and finite capacity to accommodate favorable outcomes. When capacity is limited, individually optimal recourse recommendations cannot be realized simultaneously, and the question of *who gets access to what* becomes unavoidable. This chapter develops the formal machinery for reasoning about recourse under these conditions, situating it within the broader literature on multi-agent resource allocation and social welfare theory.

2.1 Introduction

Existing studies on algorithmic recourse predominantly address how an individual would need to change their attributes to achieve the desired outcome [59]. Such settings generally assume a single individual impacted by a single decision-making model, as shown in Figure 2.1(a). In real-world scenarios, however, AI decision-making systems often interact with multiple individuals whose actions can influence outcomes and, consequently, recourse recommendations for others. Furthermore, individuals seeking recourse may engage with multiple recourse providers (Figure 2.1(c)) to choose the most suitable recommendation among those available. The driving insight of this study is that *ignoring the effects of*



(a) One-to-one Recourse Matching

(b) Many-to-one Setting

(c) Many-to-many Recourse Matching

Figure 2.1: Various algorithmic recourse setups: (a) the classical one-to-one setting, where an individual s seeks recourse recommendations from a provider p with minimal cost required to reverse the output; (b) the many-to-one setting, where multiple individuals are seeking recourse from a single recourse provider. (c) Our proposed many-to-many framework generalizes prior settings by simultaneously optimizing for multiple recourse seekers and providers.

other stakeholders may invalidate the recourse recommendation. To make this concrete, consider the following running example that will recur throughout this study. A group of loan applicants, each rejected by one or more banks, simultaneously seeks recourse recommendations. Each bank has a fixed loan approval capacity: it can only approve a limited number of applications per period. If each applicant independently follows the individually optimal recourse recommendation, they may all be directed toward the same bank offering the lowest recourse cost, exhausting that bank's capacity and leaving many seekers without a feasible match. Meanwhile, other banks with available capacity remain underutilized. The aggregate outcome is inefficient, and the welfare loss is entirely attributable not to any failure of individual recourse methods, but to the absence of coordination across a population of seekers under capacity constraints.

While algorithmic recourse is typically framed as an individual intervention, providing a single affected person with minimal-cost actions to reverse a fixed model decision, recourse objectives also implicitly induce a system-level allocation problem. In realistic decision-making systems, providers have limited capacity to provide favorable outcomes or actionable support, and multiple individuals may simultaneously seek recourse from multiple decision-makers. When such capacity constraints are present, recourse recommendations compete

for scarce resources, transforming recourse from a purely individual optimization problem into a question of how limited intervention capacity should be allocated across a population. This observation aligns algorithmic recourse with broader questions of welfare maximization under resource constraints, a connection that has received comparatively limited attention in the recourse literature.

From this perspective, *social welfare* provides a natural objective for evaluating recourse systems at scale. Rather than considering recourse as an isolated right exercised independently by each individual, a welfare-based formulation evaluates how effectively a system converts limited intervention capacity into meaningful, low-cost opportunities for improvement across the population. Welfare-based views have played a central role in the study of fairness in automated decision-making, where individual utilities, distributive justice, and aggregate outcomes are explicitly traded off [11, 49, 46, 22, 10]. In our setting, we characterize the best achievable population-level outcome under explicit capacity constraints. The resulting *welfare gap* quantifies the efficiency loss that arises when individually optimal recourse recommendations are infeasible at the system level.

Recent work has begun to demonstrate that recourse objectives themselves can induce inequitable or inefficient outcomes when optimized locally. In particular, Perello et al.[92] show that standard algorithmic recourse formulations may generate discrimination even in the absence of biased decision models, due to the interaction between recourse costs and population-level constraints. Kügelgen et al.[72] further show that cost-based recourse can embed structural inequities even when classifiers are causally fair, and Bell et al.[7] argue that minimizing individual recourse cost is insufficient when individuals face systematically different opportunity structures. These findings collectively motivate our shift from individual cost minimization to population-level welfare optimization.

There remains a gap in the literature regarding algorithmic recourse in situations involving multiple recourse providers, each potentially impacting outcomes with their own decision models and capacity constraints. Existing approaches typically overlook how providing recourse recommendations simultaneously to multiple recourse seekers can benefit society and overall recourse actionability through interactions among individuals. We model provider capacities and recourse costs jointly, exposing a structural source of inefficiency that cannot be addressed by improving individual recourse methods alone.

2.2 Related Work

2.2.1 Multi-Agent Recourse

The vast majority of algorithmic recourse research operates in the one-to-one setting with a single seeker interacting with a single provider, and the recourse problem is solved independently for each individual without reference to other affected parties. A small but growing body of work has begun to challenge this framing, motivated by the observation that in real-world settings, recourse seekers do not act in isolation; they share providers, compete for limited capacity, and collectively shape the environment in which future recourse is sought.

O'Brien et al. [85] are among the first to formally examine the consequences of treating recourse as a collective rather than purely individual problem. Drawing on game-theoretic principles, they introduce two alternatives to standard individual cost minimization, *Social-Welfare-Efficient Recourse*, which maximizes aggregate welfare across the seeker population, and *Pareto-Efficient Recourse*, which requires that no seeker can be made better off without making another worse off. Their central illustration is a prisoner's dilemma over recourse. When seekers act independently, each choosing the recommendation that minimizes their own cost, the resulting Nash equilibrium is collectively suboptimal and Pareto-dominated by a coordinated solution. This provides both a game-theoretic diagnosis of the inefficiency of individual recourse and a normative argument for coordination, even when individual recourse methods are individually optimal, they may be collectively harmful.

Fonseca et al. [32] approach the multi-agent question from a dynamic perspective, proposing an agent-based simulation framework to study the effects of a continuously changing environment on the reliability of recourse recommendations over time. Their framework identifies two distinct mechanisms by which recourse reliability degrades: (1) competition among existing seekers who act simultaneously on similar recommendations, shifting the decision boundary as a result of their collective actions, and (2) competition from new seekers who continuously enter the environment and displace earlier recourse-holders. A finding is that only a small set of specific parameterizations yield recourse recommendations that remain reliable over time, and that existing recourse methods fail to adequately reward the effort of seekers who follow recommendations in good faith but are later displaced. Fonseca et al. [32] conclude that substantial additional work is needed to understand recourse reliability under dynamic competition and to develop methods that account for temporal shifts.

Building on this framework, Bell et al. [8] introduce *The Game of Recourse*, an agent-based simulation designed to synthesize realistic recourse data while explicitly accounting for

both reliability and fairness over time. Their work highlights that recourse recommendations may be particularly unreliable and unfair for members of disadvantaged groups, whose initial circumstances already require more effort to achieve favorable outcomes, compounding disparity over time as competition intensifies. Together, these works establish that fairness in recourse is inseparable from the multi-agent dynamics of competition and resource scarcity.

A complementary population-level concern is raised by Gao et al. [34], who study the delayed societal impact of algorithmic recourse at scale. Analyzing the recourse output by state-of-the-art algorithms, they provide one of the first theoretical and empirical analyses of how large-scale recourse implementation leads to data distribution shifts and potential social segregation, a phenomenon where seekers who successfully obtain recourse cluster into distinct regions of feature space, reinforcing existing structural divides. Their work underscores that individual recourse methods, however well-designed, can have unintended population-level consequences that only become visible when recourse is examined at scale.

Perello et al. [92] make a closely related point from a different angle, demonstrating that standard algorithmic recourse formulations can generate discrimination even in the absence of biased classifiers. The mechanism is structural. Recourse cost functions, combined with heterogeneous initial circumstances across demographic groups, produce systematically different burdens for different populations even when the underlying decision model is group-fair. This finding illustrates the inadequacy of optimizing individual recourse costs in isolation and provides a formal motivation for population-level welfare criteria.

Taken together, these works leave a clear and important gap. Existing multi-agent recourse research considers either (1) a single provider with multiple seekers, or (2) temporal dynamics with a fixed provider, but never multiple providers simultaneously interacting with multiple seekers under explicit capacity constraints. This is precisely the many-to-many setting formalized in this study.

2.2.2 Multi-Agent Systems and Resource Allocation

The many-to-many recourse setting connects naturally to a broader literature on how scarce resources should be allocated among competing agents to maximize collective welfare. We draw on three related bodies of work, namely, capacitated bipartite matching, regret analysis in online learning, and two-sided matching theory.

Bipartite Matching Under Capacity Constraints. The maximum-weight bipartite matching problem is one of the most fundamental problems in combinatorial optimization,

solvable in polynomial time via the Hungarian algorithm [67]. With capacity constraints, where each resource can serve at most k_j agents, the problem remains polynomial but requires more sophisticated LP or MILP formulations [67]. This family of problems has been studied extensively in domains where the allocation of agents to limited resources has direct welfare consequences.

Dickerson et al. [26] study online bipartite matching with reusable resources under adversarial arrival distributions, developing LP-relaxation-based algorithms that achieve near-optimal competitive ratios. Their central result is that system-level coordination, in which a central planner makes globally informed assignments, consistently and substantially outperforms greedy per-agent assignment when resources are constrained and agent preferences overlap. Li et al. [77] address weighted bipartite matching in the context of kidney exchange, introducing a hybrid online model that incorporates all compatible donor and recipient pairs rather than restricting to pre-specified compatible subsets. Their approach substantially increases the number of successful transplants, demonstrating that welfare gains from coordinated matching can accrue most strongly to the most disadvantaged subpopulations.

These works motivate the use of capacitated bipartite matching as the algorithmic backbone of our framework. However, the recourse setting introduces a crucial distinction. The edge weights in our bipartite graph are not exogenous compatibility scores or logistical preferences. Instead, each weight is derived from the solution to an individual recourse optimization problem (Chapter 3), so improving the matching directly reduces the recourse cost by individual seekers. This creates a principled coupling between the combinatorial matching layer and the individual recourse objectives, it is designed to coordinate a connection that has no analog in prior bipartite matching work.

Regret and Welfare Gaps in Online Learning. The welfare gap we define has structural similarities to regret in online learning. In the classical multi-armed bandit (MAB) setting, regret measures the cumulative difference between the rewards an algorithm obtains and those of the best fixed policy in hindsight [15]. Our welfare gap plays an analogous role at the system level. It quantifies the welfare loss that arises when seekers act independently, each choosing the provider with the lowest individual recourse cost, relative to the welfare achievable under centrally coordinated assignment. Both quantities measure the cost of decentralized, nearsighted decision-making relative to a centralized optimum.

Hossain et al. [48] extend the MAB framework to the multi-agent setting, introducing a model in which multiple agents simultaneously pull arms and receive heterogeneous stochastic rewards. Their key contribution is defining regret in terms of the *Nash social*

welfare [21] the product of individual utilities and designing algorithms that achieve sublinear Nash regret. This is significant because Nash social welfare is itself an inequality-averse objective, it is zero whenever any agent receives zero utility, making it sensitive to the worst-off agent in ways that utilitarian regret is not. Our concave welfare function with parameter α generalizes Nash social welfare ($\alpha = 0$) to a continuous family of inequality-averse objectives.

Patil et al. [86] study the Fair-SMAB problem, in which a single agent must satisfy minimum pull-frequency constraints for each arm alongside maximizing cumulative reward. The minimum pull frequency constraint in Patil et al. [86] is conceptually analogous to our provider capacity lower bounds; both enforce a minimum usage level on each resource, ensuring that no provider is systematically underutilized even if it is not the individually preferred option for most seekers. Our welfare gap can be understood as a static, full-information version of these regret notions. We measure the loss from decentralized assignment in a single-shot setting rather than over a learning horizon, and the source of the gap is capacity misallocation rather than epistemic uncertainty.

Two-Sided Matching. The deferred acceptance algorithm of Gale et al. [33] produces a stable two-sided matching in polynomial time, no seeker–provider pair would mutually prefer each other over their assigned match. The algorithm has been successfully deployed in medical residency matching, school admissions, and organ exchange, where stability is a key desideratum because it prevents manipulation by participants who could otherwise benefit from deviating. Our framework is related in spirit, but we optimize for welfare rather than stability, and our matching is driven by recourse costs rather than ordinal preference rankings. These objectives can diverge substantially when recourse costs are heterogeneous across seeker–provider pairs, as is typical in practice. Exploring stability as a complementary constraint alongside welfare maximization is a natural extension of this work, discussed further in Chapter 5.

2.2.3 Social Welfare and Inequality Aversion

The question of how to aggregate individual utilities into a meaningful system-level objective is classical in welfare economics and social choice theory, and has become increasingly central in algorithmic fairness. We draw on this literature to motivate both the base utilitarian matching objective and the inequality-averse extension in Chapter 3.

Utilitarianism and Rawlsian Justice. The two canonical poles of normative welfare theory are utilitarianism [93] and Rawlsianism [96]. The utilitarian criterion, maximizing the sum of individual utilities, is the standard benchmark in classical welfare economics and corresponds directly to our base matching objective, which is maximizing the total weight of the matching. The Rawlsian criterion originates in Raw's [96] *veil of ignorance* thought experiment. A social arrangement is one that rational agents would choose without knowing the position they will occupy in the resulting distribution. Rawls argues that under the veil of ignorance, rational agents would opt for the *difference principle*, inequalities are just only insofar as they benefit the least-advantaged members of society, which, in the welfare maximization context, corresponds to the max-min criterion; maximize the welfare of the worst-off individual. The veil of ignorance has gained renewed traction in AI and algorithmic decision-making research as a principled device for eliciting fairness-based reasoning. Gale et al. [33] show empirically across five studies that participants behind the veil more frequently choose, and upon reflection endorse, principles for AI governance that prioritize the worst-off, driven by fairness-based reasoning rather than risk aversion alone. This empirical evidence supports using Rawlsian criteria as normative anchors in algorithmic allocation design.

Heidari et al. [46] operationalize the Rawlsian perspective for algorithmic decision-making. They propose a family of welfare-based fairness measures that correspond directly to cardinal social welfare functions from economics, justified by the veil of ignorance. The measures that ransfers benefit from a high-benefit to a low-benefit individual strictly increases welfare, so the measures are intrinsically sensitive to inequality. A central study is that lower-bounding these measures provides a computationally tractable mechanism for bounding individual-level inequality in algorithmic outcomes. The measures are parameterized by a concavity coefficient α, which interpolates between the utilitarian sum ($\alpha \to 1$) and the Rawlsian max-min. This parameterization is precisely what we adopt for the inequality-averse layer of our framework in Chapter 3: by varying α, the planner can navigate the full spectrum between maximizing aggregate welfare and protecting the worst-off seeker.

The Price of Fairness and the Efficiency–Equity Trade-Off. Moving from a utilitarian to a fairness-aware objective generally reduces aggregate efficiency. Bertsimas et al. [9] formalize this trade-off as the *price of fairness*. For a broad family of resource allocation problems, they define the price of fairness as the relative efficiency loss incurred by switching to either proportional fairness or max-min fairness. The implication is practically significant. Fairness need not be prohibitively costly in terms of aggregate welfare. Bertsimas et al. [10] extend this analysis by deriving the full Pareto frontier between aggregate efficiency and a broad set of fairness criteria, characterizing how the frontier shifts as the fairness

requirement tightens. They show that navigating this frontier corresponds precisely to varying the concavity of the welfare function, the same mechanism we exploit through the parameter α in our framework. Rea et al. [97] take a complementary empirical perspective, distinguishing between *equity* (input-weighted allocation) and *equality* (uniform outcomes) and constructing a multi-objective optimizer to derive the Pareto frontier between the two. Their analysis shows that substantial gains in equity can often be achieved at small losses in aggregate efficiency, motivating the use of smooth inequality-averse objectives rather than hard demographic constraints. Hu et al. [49] provide a cautionary finding: enforcing strict demographic parity constraints in classification tasks can reduce welfare for *both* advantaged and disadvantaged groups simultaneously, because hard constraints force the system to reallocate resources away from high-utility assignments without compensating gains for low-utility individuals. This leveling down impact is precisely the failure mode that our parameterized concave welfare function avoids: by softly penalizing inequality rather than enforcing hard parity, the system can improve the welfare of the worst-off seekers without necessarily reducing the welfare of others.

Welfare Maximization in Algorithmic Allocation. Welfare-centric perspectives have become increasingly prominent in the algorithmic fairness literature, motivated by the observation that purely statistical notions of group fairness can fail to capture the distributive justice concerns that motivate fairness in the first place [11]. Corbett-Davies et al. [22] analyze the efficiency–fairness trade-off in algorithmic decision-making, showing that fairness constraints on classifiers necessarily reduce utility, but the size of the reduction depends on the choice of constraint, and welfare-based constraints often achieve better equity–efficiency trade-offs than demographic parity constraints. Related work in operational research studies the allocation of limited resources fairly across heterogeneous needs, developing both theoretical characterizations of welfare– equity trade-offs and practical optimization methods [97].

Taken together, this body of work establishes that welfare-based formulations are both normatively grounded and computationally tractable, and that the efficiency–equity trade-off can be navigated systematically through the concavity of the welfare function. To the best of our knowledge, this study is the first to apply this welfare-maximization perspective to algorithmic recourse, formalizing the allocation of limited provider capacity as a welfare optimization problem and exposing the efficiency–equity Pareto frontier through a parameterized concave objective.

Chapter 3

Capacitated Bipartite Matching for Social Welfare Optimization

3.1 Many-to-Many Recourse Optimization

In this section, we formalize the matching problem that serves as the foundation of our optimization framework. The problem is modeled as a bipartite graph as shown in Figure 2.1(c), with loan seekers on one set and loan providers (e.g., banks) on the other, as follows:

- **Seekers** $\mathcal{S}$: $\{s_i \mid s_i \in \mathcal{S},\ \forall\ i \in [n]\}$, each characterized by a feature vector x_i.
- **Providers** $\mathcal{P}$: $\{p_j \mid p_j \in \mathcal{P},\ \forall\ j \in [m]\}$, each equipped with a classifier (w.l.o.g. binary model) h_j to accept or reject the seekers and a matching capacity k_j. All the seekers are initially rejected by all providers, i.e., $h_j(x_i) = -1\ \forall\ i, j$, meaning that each seeker will have a recommendation from all providers.

Furthermore, it is assumed that a *central planner* will coordinate matches between seekers and providers (i.e., eq. (3.1)) and potentially redistribute existing capacity among providers (i.e., eq. (3.2) and eq. (3.3)).

Recourse cost computation The standard algorithmic recourse setting assumes that a seeker (e.g., s_i) seeks to obtain recourse recommendations from a provider (e.g., p_j). This minimal change, as defined by Ustun et. al [107], is defined as the minimal effort or cost

required to change an individual's input features such that a predictive model will reverse its output from an undesirable outcome to a desirable one. Formally, given provider p_j's decision model h_j and an input feature vector x_i corresponding to the characteristics of seeker s_i, such that $h_j(x_i) = -1$ (assumed binary w.l.o.g.), the recourse cost is defined as the solution to the following optimization problem:

$$c_{ij} = \min_{a \in A(x_i)} \text{cost}(a; x_i) \quad \text{s.t.} \quad h_j(x_i + a) = +1 \quad \forall\ i, j$$

where a is an action vector representing feasible changes to the features of x_i, and $A(x_i)$ is the set of allowed actions based on domain constraints (e.g., mutability and bounds on feature changes). The function $\text{cost}(a; x_i)$ quantifies the difficulty of applying action a to instance x_i. If a feasible action exists that satisfies the constraint, the minimal value of $\text{cost}(a; x_i)$ is the *recourse cost*. Therefore, c_{ij} represents the minimal change required for seeker s_i to achieve approval from provider p_j. Our proposed framework is agnostic to the choices of recourse method and providers' model, operating only on the minimum cost of change required for each pair of seeker s_i and provider p_j.

Bipartite Graph Construction. Once all minimal recourse costs, c_{ij}, between seekers and providers are precomputed, we construct a *weighted bipartite graph* $\mathcal{G} = (\mathcal{V}, \mathcal{W})$, where nodes $\mathcal{V} = \mathcal{S} \cup \mathcal{P}$ and

$$\mathcal{W} := \left\{ w_{ij} \;\middle|\; w_{ij} = e^{-\gamma \cdot c_{ij}},\ \forall\ i, j \right\}.$$

where $\gamma > 0$ is a scaling parameter controlling the sensitivity of the transformation. This exponential transformation converts costs into edge weights, enabling algorithms such as the maximum-weight bipartite matching [67] to prioritize low-cost (i.e., efficient) recourse assignments while maximizing overall match coverage. Furthermore, this transformation is monotone in c_{ij} and consistent with an exponential utility representation exhibiting Constant Absolute Risk Aversion (CARA), a standard model in economics for capturing diminishing marginal disutility of cost or effort [2, 93]: lower recourse costs yield higher utility, while differences among low-cost interventions are weighted more heavily than differences among high-cost ones. Beyond its utility interpretation, γ represents a meaningful modeling decision about how strongly cost differences between providers are reflected in the matching weights. A formal analysis of γ's effect on social welfare is provided in Section 3.4.1.

Optimization Model Next, with consideration for real-world constraints on provider capacity, we formulate a *capacitated weighted bipartite matching problem*. If we denote the maximum weight for seeker i accordingly as $w_i^* = \max_j(w_{ij})$ (corresponding to the least

costly recommendation received), we can then measure the ideal scenario in which each seeker attains its optimal outcome, independently of other seekers, as:

$$\text{Individual Welfare} := \sum_{i=1}^{n} w_i^*$$

However, this ideal scenario assumes that providers have unbounded capacity (w.l.o.g., at least the number of seekers for each provider), meaning that they can freely provide the resources, which is not realistic. In practice, each provider has a limited capacity k_j, meaning they can serve only a finite number of seekers. Taking a systems-level view and aiming to minimize the overall cost of recourse across all seekers and providers,[1] the optimal matching under capacity constraints is determined as:

$$\text{Social Welfare} := \max_{z_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}\, z_{ij}.$$

where z_{ij} are binary decision variables that indicate whether seeker i is assigned to provider j. This equation has the structure of a classical capacitated bipartite matching problem, solvable in polynomial time via assignment algorithms [67]. In our implementation, to obtain the optimal matching, we encode the above formulation as a mixed-integer linear program (MILP) and solve it with the Gurobi Optimizer [40][2] as follows:

$$\begin{aligned} \text{Social Welfare} := \max_{z_{ij}} \quad & \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}\, z_{ij} \\ \text{s.t.} \quad & \underbrace{\sum_{j=1}^{m} z_{ij} \le 1 \quad \forall i,}_{\text{Matching Constraint}} \quad \underbrace{\sum_{i=1}^{n} z_{ij} \le k_j \quad \forall j,}_{\text{Capacity Constraint}} \quad \underbrace{z_{ij} \in \{0,1\} \quad \forall i,j}_{\text{Edge Constraint}} \end{aligned} \tag{3.1}$$

Under unbounded provider capacity, the optimal matching over z_{ij} is achieved when each seeker is matched with the provider that minimizes its individual recourse cost. However, capacity constraints may leave some seekers to accept suboptimal assignments ($w_{ij} < w_i^*$), resulting in higher recourse costs than their ideal match. This discrepancy is quantified by

[1] This assumes that providers do not have ulterior preferences affecting the matching process.

[2] Gurobi's branch-and-bound engine—augmented with presolve, cutting-plane generation, and heuristic warm-starts–guarantees global optimality.

the gap between the ideal individual welfare and the realized social welfare:

$$\text{Welfare Gap} := \left(\sum_{i=1}^{n} w_i^*\right) - \left(\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}\, z_{ij}\right).$$

This gap highlights a critical design challenge: given a fixed total amount of provider capacity, how should these limited resources be distributed across providers to minimize the welfare gap? A naive uniform distribution of provider capacities may lead to significant welfare losses. In contrast, allocating more capacity to providers most preferred by seekers, i.e., those associated with lower recourse costs, can substantially reduce the welfare gap, even when the total capacity remains fixed. In the following section, we introduce a systematic approach that not only identifies distinct allocation scenarios based on varying resource availability but also provides an optimization approach to allocate capacity effectively and minimize the welfare gap.

3.2 Minimizing the Welfare Gap

Under a fixed total provider capacity $K = \sum_{j=1}^{m} k_j$, the welfare gap can vary depending on how capacity is distributed among providers. In fact, for any given K, there is an optimal allocation of provider capacities k_j that minimizes this welfare gap. This observation leads to a new optimization problem involving two sets of decision variables namely, the integer variables $k_j \;\forall\; j$, representing provider capacities in the optimal solution, and the matching variables z_{ij}, as previously defined, indicating the best matching under the system settings.

$$\max_{z_{ij},\, k_j} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}\, z_{ij} \quad \text{s.t.} \quad \sum_{j=1}^{m} k_j = K \quad \forall\, j \quad \text{Total Capacity Constraint} \tag{3.2}$$

The matching, capacity, and edge constraints remain the same as before in Equation (3.1), with the additional constraint on the total capacity. Figure 3.1 represents how this welfare gap varies according to the total available resources and their distribution across providers. In particular, the figure highlights two notable cutoffs and three distinct areas. In the first area, where the total available resources are fewer than the number of seekers, i.e., $K = \sum_{j=1}^{m} k_j < |S|$, a welfare gap greater than zero is inevitable. Under these circumstances, some seekers will remain unmatched, and the depicted minimum gap represents the best achievable outcome through optimal capacity distribution. The first critical point is reached once available resources equal the number of seekers, where the welfare gap *can* reach zero

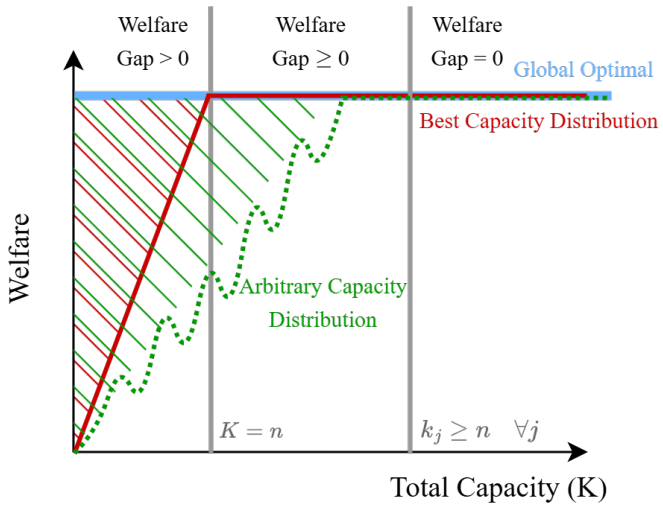


Figure 3.1: The effect of capacity distribution on social welfare outcomes. The pink dashed plot (with shaded region) shows the welfare achieved under an arbitrary capacity allocation (Equation (3.1)), which may result in a non-zero welfare gap even when sufficient total capacity is available. The dark red plot represents the optimized distribution obtained by solving Equation (3.2), which maximizes social welfare under a total capacity constraint. The blue horizontal line shows the global individual-optimal welfare, attainable only when total capacity is unconstrained. The gap between the pink and red curves highlights the inefficiency introduced by uncoordinated capacity allocations.

if the distribution of these resources could be optimized and aligned with the best choice of each seeker. At the second critical point, where each provider individually possesses resources equal to the number of seekers, the welfare gap is guaranteed to be zero as seekers can freely match with their preferred providers. In the area between the two critical points, while resources are plentiful enough for a zero welfare gap distribution over capacities to exist, any suboptimal distribution of capacities among providers will result in a welfare gap greater than zero. Therefore, identifying the optimal capacity distribution becomes an essential challenge, specifically, determining how capacities can best be allocated to minimize the welfare gap given fixed total resources.

To determine the optimal distribution of provider capacities, Equation (3.2) could again be solved using a Mixed-Integer Linear Programming (MILP) method with solvers such as Gurobi [40]. However, the structure of the matching weights reveals two key patterns,

suggesting an approach simpler than solving an MILP directly. In Algorithm 1, we propose a systematic method that assigns capacities based on each seeker's top-ranked matching weight, previously defined in Section 3.1 as w_i^*. For each seeker i, the value w_i^* represents the maximum possible contribution of that seeker to the total welfare, as no assignment can surpass this highest-weight edge. Moreover, since at most K seekers can receive recourse, excluding any of the top K highest-weight edges directly reduces the achievable welfare. Thus, the welfare of any feasible solution is bounded above by $\sum_{i \in \mathcal{S}_K} w_i^*$, where $\mathcal{S}_K$ denotes the set of seekers corresponding to the K highest-ranked edges, and $j_i^* = \arg\max(w_i^*)$ denotes the index of the provider most preferred by seeker s_i. Then the capacity vector k^* is defined by counting how often each provider appears among the top-K individually preferred matches (e.g. w_i^*).

$$k_j^* = |\{i \in S_K \mid j_i^* = j\}|, \quad j = 1, ..., m$$

These insights lead to Algorithm 1, which offers a more direct and efficient alternative to MILP. After computing all w_i^* values in $O(nm)$ time (n seekers and m providers), we sort the resulting values once in $O(n \log(n))$ time. Capacities are then assigned to providers precisely according to the top K dominant edges. This approach ensures provider capacities align directly with seeker preferences, thereby naturally minimizing the welfare gap by emphasizing matches that are both individually optimal and beneficial for the overall system. Although, Equation (3.2) could be solved using a solver such as Gurobi, Algorithm 1 exploits the special structure of the problem to provide a direct, closed-form solution and it recovers the same optimal social welfare as the solution from the solver.

3.2.1 Capacity Redistribution with Penalizing Modifications

While the previous section addressed how provider capacities can be optimally distributed to minimize the welfare gap for a given total system capacity K, additional considerations may be warranted. In practice, recourse methods often operate within established configurations determined by existing organizational structures, resource availability, and operational constraints. Transitioning from the current provider capacity configuration to an optimal setup typically involves real-world adjustment costs. Simply recommending an entirely new capacity distribution may be impractical or expensive to implement. Therefore, even though optimally reallocating capacities reduces the welfare gap, the costs of these adjustments must be carefully balanced against minimizing the welfare gap. To address this challenge, we extend our optimization framework by explicitly penalizing deviations from the initial capacities.

Algorithm 1 Optimal Capacity Distribution

Input: seekers $\mathcal{S}$, providers $\mathcal{P}$, weights w_{ij}, total capacity K
Output: provider capacities $k = (k_1, \ldots, k_{|\mathcal{P}|})$
Initialize empty list $\mathcal{L}$
for each seeker $i \in \mathcal{S}$ **do**
    $w_i^* = max_j(w_{ij})$
    $j_i^* \leftarrow \arg\max_{j \in \mathcal{P}} \; w_i^*$ {best provider for seeker i}
    Append triple $(i,\ j_i^*,\ w_i^*)$ to $\mathcal{L}$
end for
Sort $\mathcal{L}$ in descending order of weight
Select the first K triples of $\mathcal{L}$ {top-K matches}
Initialize $k_j \leftarrow 0$ for all $j \in \mathcal{P}$
for each selected triple $(i,\ j,\ w)$ **do**
    $k_j \leftarrow k_j + 1$
end for
return capacity vector k

Let $\hat{k_j}$ denote the initial capacity of provider j, and $\tilde{k_j}$ represent the target capacity after configuration. The change in capacity $\Delta k_j = \tilde{k_j} - \hat{k_j}$ can penalize large changes with $\beta_j |\Delta k_j|$, where $\beta_j \geq 0$ controls the penalty sensitivity for each of the providers accordingly. Integrating this penalty into our optimization leads to a multi-objective problem, balancing social welfare maximization with minimization of capacity adjustment penalties. The modified objective function is:

$$\text{Welfare} = \max_{z_{ij},\ k_j} \left(\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} z_{ij} - \sum_{j=1}^{m} \beta_j |\Delta k_j| \right) \tag{3.3}$$

subject to the same matching, capacity, total capacity, and edge constraints previously defined in Equation (3.2). Unlike Equation (3.1), the penalized formulation no longer reduces to a standard bipartite matching or flow problem, due to the interaction between capacity constraints. As a result, this enhanced formulation constitutes a MILP problem, solvable by recent versions of solvers like Gurobi [40]. Ultimately, this optimization simultaneously identifies the optimal matching z_{ij} and optimal provider capacities k_j, clearly indicating how capacities should be adjusted from their initial configuration. The resulting solution provides a practical recourse recommendation system effectively balancing improvements in social welfare with realistic operational constraints.

3.3 Inequality-Averse Recourse Optimization

We have introduced the three-layer optimization framework in Section 3.1 under a utilitarian objective, which is maximizing the aggregate sum of weights across the population. While the approach in Section 3.2 minimizes the total welfare gap, it is theoretically blind how that welfare is distributed among individuals. In resource-constrained settings, a purely utilitarian objective naturally favors *efficient* matches, seekers with low recourse costs (high edge weights w_{ij}), while it may inadvertently disadvantage those with higher costs. This creates a risk that efforts to maximize overall performance systematically disadvantage the most vulnerable seekers. If a recourse system achieves high total welfare by serving only the easiest cases while leaving others with no actionable recommendation, it fails to provide the meaningful opportunities required by the right to recourse. To address this, we extend our framework to explicitly model *distribution-sensitive welfare*, trading off allocative efficiency for distributional equity [10, 49].

To account for inequality aversion in the matching process, we use a concave social welfare function. Unlike the utilitarian approach in Equation (3.1), which assigns constant marginal value to utility gains, a concave objective places diminishing marginal value on increasing the utility o well-off seekers. This functional form prioritizes improvements for those with the lowest weights, a standard approach in welfare economics for modeling inequality aversion [4, 2].

Let $u_i = \sum_{j=1}^{m} w_{ij} z_{ij}$ denote the utility attained by seeker i under a given matching. We replace the linear objective with the following concave welfare function, consistent with Constant Absolute Risk Aversion (CARA) assumptions [93]:

$$\max_{z_{ij}} \sum_{i=1}^{n} (u_i)^{\alpha}, \quad \text{where } \alpha \in (0, 1)$$

subject to the matching and capacity constraints defined in Equation (3.1). The parameter α serves as a coefficient of *inequality aversion*. When $\alpha = 1$, the objective recovers the standard utilitarian function. As $\alpha \to 0$, the optimization approaches the Rawlsian Max-Min principle [96], prioritizing the maximization of the worst-off seeker's utility above all else. This formulation provides a principled mechanism to balance aggregate efficiency with equity without requiring rigid constraints or manual quotas.

To evaluate the fairness of allocations produced by our framework, we define a specific Equity metric, denoted as $\mathcal{E}$. While the social welfare shows the system's total efficiency, $\mathcal{E}$ quantifies the Rawlsian difference principle by measuring the *floor* of the recourse

distribution. We define $\mathcal{E}$ as the minimum utility attained by any successfully matched seeker:

$$\mathcal{E} = \min_{i \in \mathcal{M}} u_i = \min_{i \in \mathcal{M}} \left(\sum_{j=1}^{m} w_{ij} z_{ij} \right)$$

where $\mathcal{M} = \{i \mid \sum_j z_{ij} = 1\}$ represents the set of matched seekers. An increase in $\mathcal{E}$ indicates that the system has successfully increased the outcome for the least advantaged individual. In Section 3.4.1, we use this metric to demonstrate that our concave formulation can significantly increase $\mathcal{E}$ with only marginal costs to social welfare.

3.4 Experiments

This section demonstrates how the proposed many-to-many recourse framework behaves in practice. All experiments are fully reproducible from the scripts accompanying this paper.

Datasets and Models We evaluate our framework on three datasets across two experimental scales. For small-scale mechanistic illustration we use the synthetic **Two-Moon** benchmark, whose non-linear geometry yields heterogeneous classifiers with disagreements on decision boundaries. For large-scale validation we use the real-world **Credit** [50] and **COMPAS** [73] datasets, widely used in recourse research.

In the *small-scale* setting, four classification models serve as the $m = 4$ providers: logistic regression, a multilayer perceptron (MLP), a decision tree, and a random forest. We subsample $n = 8$ seekers for Two-Moon, instances that are rejected by all providers, forming the pool $\{s_i\}_{i=1}^{8}$ used for visualization in Figure 3.2.

In the *large-scale* setting, we construct $m = 15$ providers spanning different performance across the same four model families. Provider heterogeneity is introduced by varying model-specific hyperparameters (e.g., regularization strength, tree depth, number of estimators), yielding providers with genuinely different decision boundaries and predictive behaviors (Table A.1 shows the preformance of these providers' model in Section A.2). Additionally, as the number of providers grows, finding seekers rejected by all providers becomes difficult, since any single permissive provider can disqualify a candidate from the seeker pool. To ensure a sufficient pool of seekers rejected by all providers, we apply provider-specific classification thresholds: some providers use more conservative (higher) decision thresholds, increasing the instances they reject and thereby enlarging the pool of seekers who are denied by all $m = 15$ providers.

Counterfactual Generation Our framework is agnostic to the choice of recourse method, operating only on the minimum cost c_{ij} for each seeker–provider pair, precomputed using a recourse method of choice. We exploit this agnosticism by using two recourse methods across our settings. For the small-scale Two-Moon experiments we use Model-Agnostic Counterfactual Explanations (MACE) [58], and for the large-scale Credit and COMPAS experiments we use Diverse Counterfactual Explanations (DiCE) [83]. We use the distance norm ℓ_1 for DiCE experiments and ℓ_∞ for MACE as the recourse cost measure. For every pair (s_i, p_j) the recourse method outputs a counterfactual instance together with its cost c_{ij}. This yields $8 \times 4 = 32$ weights for Two-Moon and $200 \times 15 = 3{,}000$ weights for each of Credit and COMPAS.

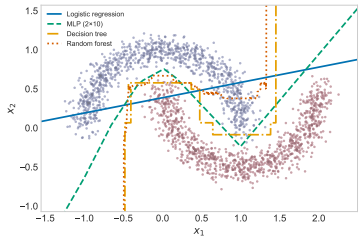
Optimization Setup The optimization proceeds in three layers across all experiments using an off-the-shelf MILP solver [40]. *(i)* Given a total capacity K equal to the number of seekers and an initial capacity vector $\hat{k}$,[3] Equation (3.1) finds the matching with the maximum social welfare under specified constraints. *(ii)* Algorithm 1 solves Equation (3.2), tracing the full welfare curve as K ranges from 0 to $n \cdot m$ and validating the theoretical cutoffs of Section 3.2. *(iii)* solves Equation (3.3), jointly optimizing matching and capacities while penalizing deviations from the initial vector. Hyperparameter selection is discussed in Section 3.4.1; the values used here are $\gamma = 10$, $\beta = 0.03$ for Two-Moon (MACE) and $\gamma = 100$, $\beta = 0.15$ for Credit and COMPAS (DiCE).

Metrics To evaluate the proposed framework, we introduce the *Welfare Attainment Ratio*, defined as:

$$\frac{\mathrm{SW}}{\mathrm{IW}} = \frac{\max_{z_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}\, z_{ij}}{\sum_{i=1}^{n} \max_j w_{ij}} \tag{3.4}$$

where the Social Welfare is feasible under constraints defined in Equation (3.1). This ratio measures the fraction of the theoretically optimal individual welfare that the system achieves under capacity constraints. A value of 1 indicates that the system delivers the same aggregate welfare as if every seeker were matched to their highest-weight provider with no capacity limitations. A value below 1 quantifies the welfare loss introduced by capacity constraints and their distribution across providers.

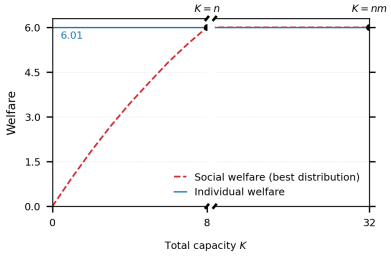
[3] For the small-scale setting the initial capacity vector is drawn uniformly at random; for the large-scale setting it is drawn from a Poisson distribution, reflecting a more realistic scenario of unequal baseline capacities whereby most providers have small capacity and some providers have large capacity.



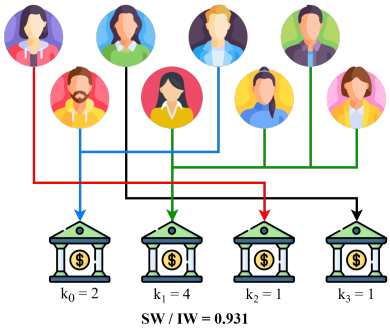
(a) Two-Moon dataset overlaid with the decision boundaries of the four provider models, underscoring that seekers face genuinely different models.

$$\mathbf{W} = \begin{bmatrix} 0.548 & 0.425 & 0.611 & 0.594 \\ 0.667 & 0.413 & 0.695 & 0.703 \\ 0.539 & 0.488 & 0.521 & 0.949 \\ 0.488 & 0.687 & 0.666 & 0.484 \\ 0.773 & 0.354 & 0.793 & 0.896 \\ 0.576 & 0.582 & 0.557 & 0.834 \\ 0.558 & 0.765 & 0.545 & 0.626 \\ 0.417 & 0.557 & 0.558 & 0.415 \end{bmatrix}$$

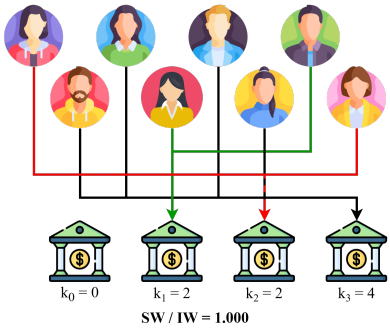
(b) Illustration of Algorithm 1 on the weight matrix. For each seeker (row), the largest edge weight w_i^* is highlighted; its column index j_i^* designates the preferred provider.



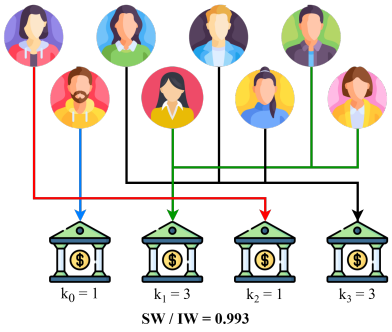
(c) Social welfare attained by Algorithm 1 (red dashed) as the total capacity K grows from 0 to $n * m$, against the individual-welfare upper-bound.



(d) Optimal seeker-to-provider assignments obtained by solving Equation (3.1) under the initial capacity vector.



(e) Provider capacities are distributed according to the capacity vector founded by Equation (3.2), eliminating the welfare gap.



(f) Matching produced by Equation (3.3), balancing social-welfare against the cost of deviating from the initial capacities.

Figure 3.2: Comprehensive illustration of our proposed framework on the Two Moon dataset and using the MACE method. (a) model decision boundaries, (b) weight matrix showing seeker-provider preferences and individually optimal matches, (c) welfare curve showing how total capacity impacts social welfare and its gap from the individual-welfare upper bound, (d) and the step-by-step optimization of provider capacities, (e) transitioning from initial assignments, to welfare-maximizing allocations, and finally (f) balancing welfare gains with capacity adjustment costs .

Results Figure 3.2(c) illustrates the social welfare achieved with the capacity distribution returned by Algorithm 1 for each total capacity K as it increases. The dashed red curve rises monotonically and meets the individual-welfare upper bound (solid blue) exactly at the point $K = n$. Beyond that, further capacity is non-essential, confirming the welfare gap would be zero in the case of finding the optimized distribution of capacities.

Figure 3.2(d) for Two-Moon dataset fixes $K = n$ (with $n = 8$) and starts from an arbitrary reference capacity vector $\hat{\mathbf{k}} = (2, 4, 1, 1)$; the first stage of our framework then computes the best matching for this initial configuration as shown in the figure. Although every seeker is matched, the allocation falls short of the global individual welfare. Figure 3.2(b) exposes the source of this gap. The next step of the algorithm finds each seeker's best edge w_i^*; distributing capacity according to Algorithm 1 and finds $\mathbf{k}^* = (0, 2, 4, 4)$. With this distribution, the system reaches the welfare upper bound, which closes the gap entirely. Figure 3.2(f) shows the outcome of the cost-aware optimization that penalizes deviations from $\hat{\mathbf{k}}$. The solver selects a softer move, $\mathbf{k} = (1, 3, 1, 3)$ yet the matching in this setting retains 0.9938 of the global optimum. Hence, almost the full efficiency gain is achievable with limited changes, demonstrating the practical value of our proposed framework. (Section A.3 shows result of the same setting experiment with ℓ_1 distance norm.)

To extend our evaluation, we summarize results for all datasets in Table 3.1, which highlights the welfare attainable ratio for each optimization setting. For the Credit and COMPAS datasets, the best matching from Equation (3.1) attained a social welfare that constitutes 0.841 and 0.832 of the individual welfare, respectively. The cost-aware optimization in Equation (3.3) selects a softer redistribution that preserved 0.948 and 0.931 of the optimal welfare for Credit and COMPAS datasets while making a more balanced capacity modification, compared to the sharper concentration found in Equation (3.2). To the best of our knowledge, this is the first work that systematically aims to close the social welfare gap by moving beyond individual recourse recommendations and instead considering coordinated, system-level optimization across multiple seekers and providers. These results align with the findings from the synthetic dataset, reinforcing the effectiveness of targeted redistribution and demonstrating that even for complex, real-world data, high social welfare can be achieved with moderate system changes.

Table 3.1: Comparative Welfare attainable ratio (SW/IW) on three different datasets of the three optimization formulations.

	Two-Moon	**Credit**	**COMPAS**
Equation (3.1)	0.931	0.841	0.832
Equation (3.2)	1.000	1.000	1.000
Equation (3.3)	0.993	0.948	0.931

3.4.1 Hyperparameter Sensitivity Analysis

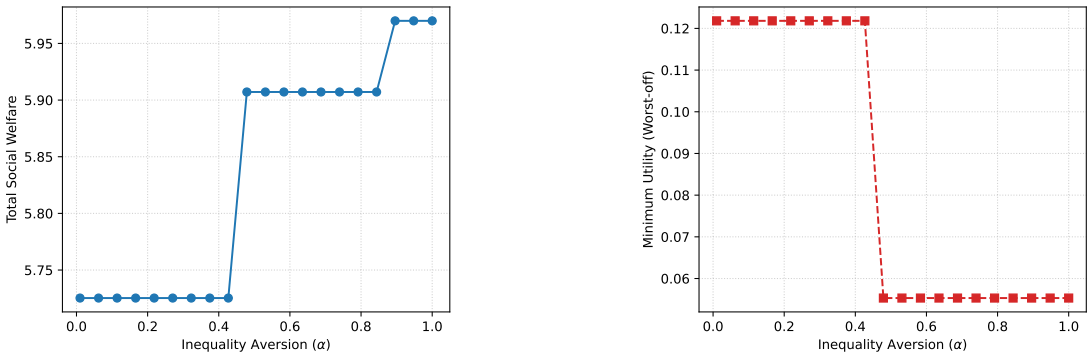
Sensitivity to α We evaluate the impact of the concave social welfare formulation (Section 3.3), focusing specifically on the COMPAS dataset which offers a more complex feature space with significant cost disparities between providers, in contrast to the Two-moon and Credit datasets. This structural heterogeneity is essential for this experiment, as it creates the necessary conditions to observe a meaningful trade-off between aggregate efficiency and individual equity.

To test the framework under these conditions, we use a high sensitivity parameter ($\gamma = 50$), which further amplifies the utility gap between preferred and secondary providers. We vary the inequality aversion parameter α from 0.01 (highly equity-focused) to 1.0 (utilitarian) and measure system performance using two competing metrics, Social Welfare (aggregate efficiency) and the Equity Metric $\mathcal{E}$ (Section 3.3).

The results, illustrated in Figure 3.3, demonstrate a trade-off between these objectives. Under the standard utilitarian baseline ($\alpha = 1.0$), the system achieves its maximum aggregate efficiency (≈ 5.95) but leaves the most disadvantaged seekers with a low utility baseline ($\mathcal{E} \approx 0.06$). As we introduce inequality aversion by decreasing α toward 0.01, the system reallocates capacity to protect these vulnerable individuals.

This redistribution results in a monotonic but minor reduction in social welfare, which drops to approximately 5.75, a relative efficiency loss of only 3.4%. In contrast to this marginal cost, the gain in equity is substantial, the minimum matched utility $\mathcal{E}$ doubles from ≈ 0.06 to ≈ 0.12. This finding validates that the concave welfare formulation successfully elevates the recourse floor for the worst-off seekers without causing a significant burden on overall system efficiency.

Sensitivity to β. The penalty coefficient β in Equation (3.3) controls the trade-off between welfare maximization and deviation from the initial capacity configuration. Fig-



(a) Efficiency Cost

(b) Equity Gain

Figure 3.3: Trade-off analysis on the COMPAS dataset with high sensitivity ($\gamma = 50$). a Total Social Welfare shows a minor decline as the system becomes more fair (lower α). (b) The Equity Metric $\mathcal{E}$ (Minimum Utility) significantly improves, validating that the concave formulation effectively protects the worst-off seekers.

ure 3.4 shows how SW/IW evolves as β increases from 0 to 1 on the large-scale experiment setting. At $\beta = 0$, the penalty term vanishes and Equation (3.3) reduces exactly to Equation (3.2)), recovering the individual welfare completely. As β increases, the growing penalty discourages capacity modifications, and SW/IW decays monotonically until converging to Equation (3.1)), where no capacity redistribution takes place. This confirms that β provides a smooth interpolation between the two extremes of the framework, and that practitioners can select any value in $[0, 1]$ to trade off welfare gain against the cost of capacity modification.

Sensitivity to γ. The scaling parameter γ controls how strongly cost differences between providers are reflected in the matching weights $w_{ij} = e^{-\gamma c_{ij}}$. Figure 3.4 shows how the attainable welfare ratio SW/IW varies as γ ranges from 0 to 15,000 in the large-scale experiment settings, revealing a characteristic U-shaped curve with three distinct regimes.

At $\gamma = 0$, all weights equal 1 and the matching is purely capacity-driven, treating every seeker-provider pair as equally desirable regardless of cost; SW/IW reflects capacity utilization alone. Since $K = n$, every seeker is matched and $SW/IW = 1$.

As γ increases from 0, weights become heterogeneous, and the matching starts favoring high-weight assignments. However, capacity constraints prevent some seekers from accessing their preferred provider. This gap grows exponentially with γ, pulling SW/IW downward.

The minimum value observed around $\gamma \approx 1{,}000$ in Figure 3.4, represents the worst-case welfare loss under the initial capacity vector $\hat{k}$: the point where cost discrimination is strong enough to create significant displacement losses but the dominant term has not yet taken over. The depth of this minimum characterizes how poorly $\hat{k}$ is aligned with the cost structure of the system.

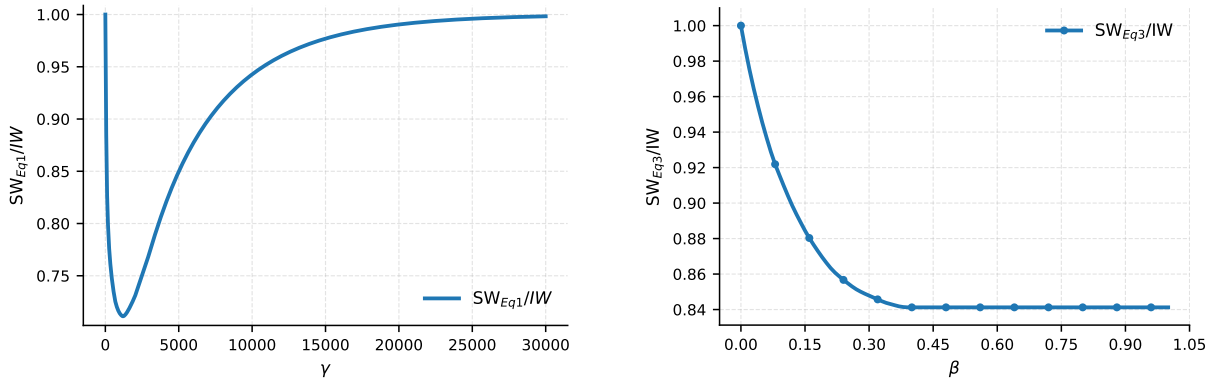
At very large γ, both SW and IW become dominated by the globally highest-weight edges in the system. Specifically, $IW(\gamma) \sim |S_{\min}| \cdot e^{-\gamma c_{\min}}$ where $c_{\min} = \min_{i,j} c_{ij}$ and $S_{\min} = \{i : c_i^* = c_{\min}\}$. When the capacity vector $\hat{k}$ permits the matching to include these globally maximum-weight edges, $SW(\gamma)$ is governed by the same dominant exponential and the ratio recovers. This asymptotic behavior is formalized in Proposition 3.4.1.

Proposition 3.4.1. *Let $c_{\min} = \min_{i,j} c_{ij}$ and $S_{\min} = \{i : c_i^* = c_{\min}\}$, and define $J_i^* = \{j : c_{ij} = c_i^*\}$. Then:*

$$\lim_{\gamma \to \infty} \frac{SW(\gamma)}{IW(\gamma)} = \frac{\max_{z \in \mathcal{Z}} \sum_{i \in S_{\min}} \sum_{j \in J_i^*} z_{ij}}{|S_{\min}|} \tag{3.5}$$

The limit equals 1 when $\hat{k}$ permits all seekers in $S_{\min}$ to access their best choice provider, and is less than 1 when capacity constraints prevent this. The limit depends on whether the initial capacity vector $\hat{k}$ can accommodate such minimum-cost assignments for seekers in S_{min}, not on the absolute magnitudes of the costs. The proof is provided in Section A.1.

The U-shaped curve reveals that SW/IW is not monotone in γ, and that neither extreme ($\gamma = 0$ or $\gamma \to \infty$) is appropriate for practical use. The elbow region identifies the most informative operating range: γ should be large enough that cost differences between providers are meaningfully reflected in the matching weights, but small enough that near-optimal alternatives remain distinguishable for seekers who cannot access their top-choice provider due to capacity constraints. When γ is very small, the matching is entirely insensitive to costs; and when it is very large, the matching treats all non-top-choice alternatives as identical, discarding useful information precisely when capacity constraints force displacement. In practice, we recommend identifying the elbow of the SW/IW curve, the range where the curve transitions from steep decay to recovery, as the operating region for γ. In our large-scale experiments this corresponds to $\gamma \in [50, 150]$, and we therefore use $\gamma = 100$ as our operating value. Importantly, the relative ordering of the three optimization equations and all qualitative findings remain stable across this range, confirming that our conclusions are robust to the precise choice of γ.



(a) Sensitivity to γ

(b) Sensitivity to β

Figure 3.4: Hyperparameter sensitivity analysis on the large-scale Credit setting ($n = 200$, $m = 15$). (a) As γ increases, $\mathrm{SW}_{\mathrm{Eq.1}}/\mathrm{IW}$ exhibits a U-shaped trend, initially decreasing due to capacity-induced displacement and later recovering as the matching becomes dominated by maximum-weight edges. The stable elbow at $\gamma \in [50, 150]$ motivates our choice of $\gamma = 100$. (b) $\mathrm{SW}_{\mathrm{Eq.3}}/\mathrm{IW}$ interpolates smoothly between Equation (3.2) at $\beta = 0$ and Equation (3.1) at $\beta \approx 0.40$.

3.5 Discussion

Our experiments show that welfare losses arise immediately in multi-agent settings when seekers apply independently under fixed, random capacities, even in low-dimensional feature spaces (e.g., the Two-Moon dataset). However, these losses can be substantially reduced with minimal system-level intervention. When total system capacity matches the number of seekers, a single pass of Algorithm 1 restores the full social welfare. This suggests that most of the efficiency gap stems not from resource scarcity but from poor allocation. Moreover, introducing a modest penalty for deviating from the initial capacity distribution (Equation (3.3)) had little impact on overall performance. This indicates that near-optimal outcomes are achievable through small, targeted adjustments, supporting practical deployment.

The results also highlight the importance of model diversity. Providers whose classifiers align well with the underlying feature distribution of a particular sub-population accrue many high-weight edges. Directing additional capacity to such specialists both increases overall welfare and improves individual actionability. In contrast, giving every provider the same capacity can waste resources on providers that do not add much value.

While our results demonstrate the effectiveness of the proposed framework across datasets and settings, it is important to note that several design choices can influence the specific quantitative outcomes in figures and tables. These include the choice of the edge-weight transformation function, the scaling parameter γ, and the penalty coefficients β_j used in cost-aware optimization. However, the central insight remains robust; there are fundamental differences between isolated one-to-one settings and resource-constrained many-to-many environments that need to be accounted for in system-level optimization. Looking forward, this framework can be enriched by considering game-theoretic extensions in which providers act strategically. Such formulations could capture competitive or decentralized recourse environments, opening new directions for modeling strategic behavior and fairness in multi-agent recourse systems.

Finally, our inequality-averse formulation reframes algorithmic recourse allocation as a system-design problem rather than a purely efficiency-driven optimization task. The choice of social welfare objective implicitly encodes assumptions about whose outcomes should be prioritized when capacity is scarce. In practice, this choice may be constrained by organizational risk tolerance or domain-specific fairness norms, suggesting that recourse systems should expose their welfare objectives as design parameters rather than fixed technical choices.

3.6 Conclusion

We have introduced a many-to-many view of algorithmic recourse in which multiple seekers obtain recommendations from multiple decision-making models whose resources are limited, and demonstrated that individually optimal recourse recommendations may no longer remain valid when real-world capacity constraints and stakeholder interactions come into play. Further, we quantified the welfare gap between the socially optimal solution, computed by a central planner, and the individually optimal outcome, where each seeker selects the provider offering the lowest recourse cost, without coordination. Experiments demonstrated that explicit capacity management improves social welfare to the near-optimal upper bound while modifying minor capacity units. We have demonstrated how recourse allocation systems can incorporate inequality-averse objectives that explicitly prioritize individuals with the poorest outcomes. This extension highlights that recourse is not solely a technical optimization problem, but also a value-laden system design choice in which efficiency and fairness must be carefully balanced when resources are limited.

Beyond the static setting studied here, several adjustments deserve closer attention. On the seeker side, adding or removing applicants or allowing limited edits to their feature

vectors directly reshapes the social welfare landscape. On the provider side, changing the number of providers or altering classifier parameters transforms the attainable optimum and the route toward it. Future studies can link these adjustment options to policy goals like ensuring diversity, fairness, or profit. Also, real recourse ecosystems are not one-shot games; seekers reapply, models retrain, and resources fluctuate. Embedding the matching layer inside an online setting would let the system adapt capacities and decision boundaries in real time, continuously shrinking the welfare gap as new data arrive. Overall, the proposed framework bridges the space between individual prescriptions and system-wide outcomes, offers a tractable path toward higher social welfare, and opens several promising directions for adaptive and fairness-aware extensions.

Chapter 4

Unification and Evaluation for Recourse

The algorithmic recourse literature has grown rapidly across multiple directions, including causality, robustness, fairness, privacy, and multi-agent settings. However, the field has evolved in a largely fragmented manner, with individual papers adopting different formulations, assumptions, constraints, and evaluation protocols [59, 108, 82]. Practitioners face a confusing landscape with limited guidance on which approaches satisfy particular deployment requirements, and researchers are not able to assess whether observed improvements reflect genuine methodological advances or differences in experimental setup [14, 82]. Part of the problem is theoretical. The recourse literature lacks a shared parameterized formulation that subsumes existing methods as special cases, making it difficult to compare approaches at the level of formal objectives or to reason about what formally distinguishes them [59, 108]. Also, empirically, existing benchmarking infrastructure does not verify that published implementations actually reproduce their originally reported results, therefore comparisons across methods may conflate algorithmic differences with implementation-specific effects [82, 88]. Both problems point toward the same need, a shared foundation that makes the recourse field legible. Formally, through a unified mathematical structure, and empirically, through reproducible evaluation infrastructure that the community can build on.

4.1 A Unified Formalization

The algorithmic recourse literature has grown substantially since the foundational contributions of Wachter et al. [113] and Ustun et al. [107], expanding into causality, robustness, fairness, privacy, personalization, and multi-agent settings. Existing surveys have organized this landscape from complementary perspectives. Karimi et al. [59] introduce a taxonomy centered on the distinction between *contrastive explanations* and *consequential recommendations*, and provide unified definitions that have become the field's primary organizational vocabulary. Verma et al. [108] take a more empirical orientation, evaluating a broad set of algorithms against a rubric of desirable properties. However, neither work proposes a single parameterized formulation from which the existing methods can be derived as special cases. This absence makes it difficult to compare approaches at the level of their formal objectives, or to characterize precisely how two methods differ when both claim to address the same desideratum.

In practice, individual papers typically instantiate only a subset of the constraints and assumptions that characterize the recourse problem in its full generality, guided by the requirements of the target application or the limitations of the chosen solver. Methods that appear similar may therefore differ substantially in their formal guarantees, yet these differences are not visible without a common reference frame. Without such a frame, determining whether a method that claims improved robustness over a baseline is solving a strictly harder problem, or merely a differently parameterized version of the same one, requires careful manual inspection of each paper's formulation, a barrier to systematic comparison that grows with the size of the literature.

We propose a unified optimization template,

$$\begin{aligned} a^\star(x) \in \arg\min_{a \in A(x)} \; & c(a;\, x) \\ \text{s.t.} \quad & R(h, x, a;\, \mathcal{U}) \geq 0, \end{aligned} \tag{4.1}$$

where $c(a;\, x)$ denotes the cost of action a for individual x, $A(x)$ encodes feasibility constraints on the action space, h is the decision model, and $R(\cdot)$ is a success requirement parameterized by an uncertainty set $\mathcal{U}$ that indexes the deployment conditions under which validity is required. The major families of recourse methods correspond to distinct instantiations of these components. Vanilla proximity-based recourse [113, 107] sets $\mathcal{U} = \varnothing$ and $R = h(x + a) - \tau$, recovering the standard boundary-crossing objective used to compute c_{ij} in Chapter 2. Causal recourse [61, 60] replaces the post-action state $x+a$ with the interventional outcome $\phi_{\mathcal{M}}(x, a)$ under a structural causal model $\mathcal{M}$, grounding recommendations in the causal structure of the domain rather than in statistical proximity. Robust recourse enlarges

$\mathcal{U}$ to cover sources of instability. When $\mathcal{U}$ indexes plausible deployed models, R becomes $\inf_{h' \in \mathcal{U}} h'(x + a) \geq \tau$ [106]. When it indexes execution noise, R becomes a probabilistic constraint over the distribution of realized post-action states [89]. When it indexes temporal model change, validity is required to hold across future retraining events [95]. Fairness-aware recourse [72, 123, 31] adds a population-level disparity constraint $\text{Disparity}(C_\pi(X); S) \leq \delta$ to the formulation. The multi-agent formulation of Chapters 2 and 3 extends the template to the population level, replacing individual cost minimization with collective welfare maximization subject to provider capacity constraints.

The template exposes structural relationships among methods that are not apparent from their individual presentations. Vanilla recourse corresponds to the limiting case $|\mathcal{U}| = 0$ of the robust formulation, meaning that methods which do not explicitly address uncertainty implicitly assume a fixed, perfectly stable deployment environment, an assumption that is strong in practice and rarely made explicit. The cost function c and the success requirement R are separable components of the template, so any proximity metric is in principle compatible with any robustness model, yet this product space has been largely unexplored in the literature. The feasible action set $A(x)$ and the success requirement R operate in different spaces: $A(x)$ constrains the actions that may be taken, while R constrains the post-action states that are considered successful. Conflating the two, for instance, by encoding plausibility requirements in $A(x)$ rather than in R, is a common source of implementation inconsistency that the template renders immediately apparent. Having a common reference formulation allows a practitioner to characterize any recourse method by specifying which components it modifies, providing a precise basis for comparison that avoids the need to reconcile disparate notations.

To connect the earlier framework to the practical landscape, we summarize existing methods according to the main design choices they adopt. To provide a comprehensive synthesis of the algorithmic recourse literature, we further summarize existing methods in Table 4.1, which builds upon the systematic taxonomy of recourse methods introduced in prior survey work [59]. Table 4.1 augments the comparison dimensions to support benchmarking-oriented analysis including evaluation properties, tooling support, and reproducibility considerations. This synthesis provides a basis for understanding current capabilities and for identifying where new methodological and benchmarking efforts are most needed.

Several patterns emerge from this longitudinal view of the tables. Differentiable models are the most widely supported classifier family across all surveyed papers, reflecting the field's historical focus on gradient-based optimization, yet this creates a systematic gap relative to real-world deployment, where tree-based and rule-based classifiers dominate in high-stakes settings such as credit scoring and criminal justice [59, 107]. Conditional

actionability constraints, which respect feature dependencies and which Karimi et al. [59] identify as essential for consequential recommendations, have become more prevalent over time, reflecting a gradual internalization of the causal critique of naive feature independence. Density-based plausibility is the most frequently enforced plausibility criterion, while domain and prototypical consistency remain comparatively rare, suggesting that plausibility enforcement in practice is still largely shaped by computational convenience rather than domain-grounded requirements. Open-source code availability has improved substantially but remains uneven. As shown in the Code column of the tables, a significant fraction of methods do not provide public implementations, which Moreira et al. [82] identify as one of the primary barriers to reproducible evaluation. The diversity of solver approaches across the surveyed methods means that running methods under a common experimental protocol is not merely desirable but necessary before any comparison can be considered meaningful.

Table 4.1: An overview of recourse algorithms for consequential decision-making settings is presented. Ordered chronologically, we summarize the *goal*, *formulation*, *solution*, and *properties* of each algorithm. Symbols are used to indicate supported settings in the experimental section of the paper (●), settings that are natural extensions of the presented algorithm (●), settings that are partially supported (◐), and settings that are not supported (○). The models cover a broad range of tree-based (TB), kernel-based (KB), differentiable (DF), or other (OT) types. Actionability contraints (unconditional or conditional), plausibility constraints (domain-, density-, and prototypical-consistency), and additional constraints (diversity, sprasity) are also explored. While the primary datatypes used in consequential settings are tabular ▦ (involving a mix of numeric, binary, categorical, and ordinal variables), we also include additional works that generate recourse for non-tabular (images 📷 and document 🗎) datasets. Furthermore, papers that present analysis of such properties as optimality (opt.), coverage (cov.), and run-time complexity (rtm.) are specified in the table. Finally, we make note of those papers that provide *open-source* implementations of their algorithm.

Algorithm	Formulation														Solution					
	Model					Actionability		Plausibility			Extra			Data types			Tools	Access	Properties			Code
	TB	KB	DF	OT	uncond.	cond.	dom.	dens.	proto.	diver.	spar.	▦	📷	🗎					opt.	cov.	rtm.	
(2021.04) Nice [13]	○	○	○	●	●	○	○	●	●	○	●	●	○	○	SB	query, blackbox	●	●	○	○		
(2021.05) CCEC [27]	○	○	○	●	○	●	○	●	○	○	●	●	○	○	heuristic	query	●	○	●	●		
(2021.06) AGSA [109]	○	○	●	○	○	●	○	●	○	○	●	●	○	○	MB, heuristic	black box	●	●	○	●		
(2021.08) CARE [94]	○	○	●	○	○	●	○	●	○	●	●	●	○	○	GB	query	●	●	●	○		
(2021.09) Counternet [38]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	grad	○	○	●	○		
(2021.10) MCCE [98]	●	○	○	●	○	●	○	●	●	●	○	●	○	○	SB/heuristic	query	○	○	○	○		
(2021.10) CCDM [12]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	grad	○	○	●	●		
(2022.01) SECP [24]	○	○	○	●	○	●	○	●	○	○	●	●	○	○	SB	query	●	●	●	●		
(2022.01) RSCE [111]	○	○	○	●	●	○	●	○	○	○	●	●	○	○	heuristic	query	○	○	●	●		
(2022.03) PRRN [89]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	grad	●	●	●	○		
(2022.04) CAR [61]	○	○	●	○	○	●	○	●	○	○	●	●	○	○	GB	white box	●	●	○	●		
(2022.05) PARP [104]	○	○	●	○	○	●	●	○	○	○	●	●	○	○	MB	query	●	●	●	○		
(2022.06) RBR [84]	○	●	●	○	◐	●	○	●	○	◐	●	●	○	○	GB	query	●	●	●	●		
(2022.06) RTAA [45]	○	○	●	○	●	○	●	○	○	○	◐	●	○	○	MB, GB	white box	○	○	○	○		
(2022.07) FLDP [112]	○	○	●	○	○	●	○	●	○	●	○	●	○	○	GB	grad	○	○	●	○		
(2022.07) RCET [28]	●	○	○	○	●	○	○	●	○	○	○	●	○	○	heuristic	white box	○	○	○	○		
(2022.10) RAGUEL [41]	○	○	○	●	○	●	○	●	○	●	◐	●	○	○	SB	query	●	●	○	○		
(2022.10) ICRI [71]	○	○	○	●	○	●	●	○	○	○	●	●	○	○	MB	white box	●	●	●	○		
(2022.10) Faster-ce [100]	○	○	●	○	○	●	○	●	○	○	●	●	○	○	GB	grad, white box	○	○	●	●		
(2023.01) ARUF [101]	○	○	○	●	○	●	●	○	○	○	●	●	○	○	heuristc, SB	query, probability	●	●	●	○		
(2023.01) FRCERO [80]	○	○	○	●	●	○	●	○	○	○	○	●	○	○	verification based	white box	●	●	○	●		
(2023.04) DRRE [125]	○	○	○	●	●	○	○	●	○	○	○	●	○	○	SB	query	○	○	●	○		
(2023.05) RCEN [43]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	grad	○	○	●	○		
(2023.08) RecRec [110]	○	○	●	○	○	●	○	●	○	○	●	●	○	○	GB	grad	●	●	●	○		
(2023.08) SafeAR [120]	○	○	○	●	○	●	○	●	○	○	●	●	○	○	MB	query	●	●	●	●		
(2023.08) UGAR [122]	○	○	●	○	○	●	●	○	○	○	●	●	○	○	GB	grad	○	○	○	○		
(2023.08) ADAR [42]	○	○	●	○	●	○	●	○	○	○	●	●	○	○	GB	grad	●	●	●	○		
(2023.08) IRAS [30]	○	○	○	●	○	●	●	○	○	○	○	●	○	○	MB	query	●	●	●	○		
(2023.09) CELS [102]	○	○	○	●	○	●	○	●	○	○	●	●	○	○	SB	query	○	○	○	○		
(2023.09) FRCE [117]	○	○	○	●	●	○	●	○	○	○	●	●	○	○	verification based	query	●	●	●	●		
(2023.09) TraCE [20]	○	○	○	●	●	○	○	●	○	○	○	●	○	○	MB	query	○	○	●	○		
(2023.09) PRPC [52]	○	○	●	○	●	○	◐	●	○	○	○	●	○	○	GB	grad/query white box	●	○	●	●		
(2023.11) CVAR [16]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	query	●	●	●	●		
(2023.11) PAR [91]	○	○	○	●	●	○	○	●	○	○	●	●	○	○	SB	query	●	●	●	○		
(2023.11) CEEA [99]	○	○	○	●	○	●	○	●	○	●	○	●	○	○	SB	query	●	●	●	○		
(2023.12) RMMA [51]	○	○	○	●	●	○	●	○	○	○	○	●	○	○	heuristic	query	●	●	●	○		
(2023.12) PeRRL [47]	○	○	◐	●	○	●	●	◐	○	◐	○	●	○	○	MB, GB	white box/query	○	●	○	○		
(2023.12) PCRD [74]	○	○	●	●	○	●	○	●	○	●	○	●	○	○	heuristic, GB	grad, white box	○	●	○	●		
(2024.01) GTCE [25]	○	●	○	○	●	○	○	●	○	○	●	●	○	○	MB	query	●	○	○	●		
(2024.02) UFCE [103]	○	○	○	●	●	●	●	○	◐	○	●	●	○	○	heuristic	query	○	●	●	●		
(2024.02) LCCO [3]	○	○	○	●	●	○	○	●	○	◐	●	●	○	○	GB	grad	○	○	●	●		
(2024.03) RRCG [19]	○	○	○	●	●	○	●	○	●	●	○	○	○	●	heuristics	probability, query	○	○	○	○		
(2024.03) PFRO [124]	○	○	○	●	●	○	○	◐	●	○	◐	●	○	○	GB	query/white box	○	●	●	○		
(2024.03) Ijuice [68]	○	○	○	●	○	●	●	○	●	○	●	●	○	○	SB	query/white box	○	○	○	○		
(2024.04) IARC [53]	○	○	○	●	●	○	●	○	○	○	○	●	○	○	heuristic	white box	●	●	●	○		
(2024.05) RAR [65]	○	○	○	●	●	◐	●	○	○	○	○	●	○	○	MB , heuristics	query, probability	●	●	○	○		
(2024.05) RARM [44]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	grad	●	●	●	○		
(2024.06) CARMA [79]	○	○	●	○	○	●	○	●	○	○	●	●	○	○	MB	grad, white-box	●	●	●	●		
(2024.06) LDTF [56]	●	○	○	○	●	○	●	○	○	○	●	●	○	○	heuristic	white box	●	●	●	○		
(2024.06) AFCR [116]	○	○	●	○	○	●	●	○	○	○	●	●	○	○	GB	gradient	●	●	○	○		
(2024.06) CIRF [66]	○	○	●	○	○	●	○	●	○	○	○	●	○	○	GB	grad	●	○	●	●		

Continued on next page

Continued from previous page

(2024.09) LGBD [55]	●	○	○	○	●	○	●	○	○	○	●	●	○	○	heuristic	white box	●	●	●	○
(2024.09) MRCE [1]	○	○	○	●	●	○	○	●	○	○	◐	●	○	○	MB	query/blackbox	○	○	●	●
(2024.10) LRAR [62]	○	○	●	○	●	○	●	○	○	○	●	●	○	○	GB	query	●	●	●	○
(2024.10) PCIW [121]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	MB, GB	white box	●	○	●	●
(2024.10) FRSM [18]	○	○	○	●	○	●	○	●	○	○	○	●	○	○	SB	query	●	●	●	○
(2024.10) RDMC [119]	○	○	●	○	○	●	○	●	○	○	○	●	○	○	GB	grad	●	●	●	○
(2024.10) CERL [90]	○	○	●	○	○	●	○	●	○	○	○	●	●	○	GB	whitebox/blackbox/grad	●	●	●	○
(2024.12) LFCA [114]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	white box	●	●	●	○
(2025.02) FACP [76]	○	○	○	●	○	●	●	○	○	○	○	●	○	○	SB	white box	●	●	●	○
(2025.03) HARE [57]	○	○	○	●	○	●	◐	●	○	○	●	●	○	○	heuristic	query	●	●	○	●
(2025.03) EvenIf [63]	○	○	○	●	○	◐	○	●	●	○	○	●	●	●	SB	query	○	○	○	●
(2025.04) PLFCA [115]	○	○	●	○	●	○	○	●	○	○	●	●	○	○	GB	white box	●	●	●	●
(2025.04) DiCE-E [5]	○	○	●	○	●	○	○	●	○	●	◐	●	○	○	GB	white box/black box	●	○	●	●
(2025.05) SSGM [35]	○	○	○	●	●	○	○	●	○	○	●	●	○	○	MB	query	●	●	●	●
(2025.06) AERR [54]	○	○	○	●	●	○	●	○	○	○	○	●	○	○	heuristic	query	●	●	●	●
(2025.08) PDMC [118]	○	●	●	○	○	●	○	●	○	○	●	●	○	○	MB, GB	query	●	●	○	○
(2025.09) ElliCE [105]	○	○	●	○	●	○	●	○	○	○	●	●	○	○	GB	query	●	●	●	●
(2025.09) RLDA [17]	○	○	○	●	○	●	●	○	○	○	●	●	○	○	MB	query	●	●	●	○

4.2 A Reproducible Benchmarking Framework

The unified formulation of Section 4.1 provides a common language for describing recourse methods theoretically. However, theoretical unification also exposes how much remains implicit in practice. Although many methods can be expressed as instances of (4.1), their empirical evaluation introduces additional, largely unspecified degrees of freedom: choices related to dataset construction, preprocessing, predictive model training, randomness control, and metric definitions. These choices are rarely standardized and frequently under-reported, despite having substantial impact on observed outcomes. As a result, empirical comparisons across recourse methods often conflate algorithmic differences with experimental artifacts, making it difficult to determine whether reported improvements reflect genuine methodological advances [82].

Recourse evaluation is inherently multi-objective. Beyond validity, methods are expected to trade off cost, feasibility, plausibility, diversity, robustness, and fairness under competing constraints. These objectives are not independent. A more robust recommendation is typically more conservative and expensive, a more diverse set of recommendations may be less actionable, and a causally grounded recommendation may be sparser but harder to achieve, and without a shared evaluation protocol, it is impossible to reason about these trade-offs in a principled, cumulative way. Brughmans et al. [14] demonstrate this concretely, showing that commonly used evaluation metrics frequently disagree in their rankings of methods, meaning that which method appears best depends critically on which metric is chosen. Without fixing the metric suite and the experimental protocol, apparent progress across papers may be misleading.

Two benchmarking frameworks have been introduced in the recourse literature to address

this. CARLA [88] provides a unified interface for generating and evaluating counterfactual explanations and recourse across multiple datasets and model families, standardizing execution and measurement. ReLax [39] extended this line of work through JAX-based implementations, demonstrating that recourse benchmarking can scale to substantially larger datasets. However, both frameworks have important limitations. Neither provides a mechanism to verify that an implementation reproduces the results reported in the original paper. Reproduction fidelity in recourse is particularly sensitive to modeling assumptions, solver choices, and feature encodings, and without explicit validation it is impossible to distinguish algorithmic behavior from implementation artifacts [82]. Their method coverage is also narrow relative to the current state of the field, which has expanded well beyond early counterfactual methods to include causal, robust, fair, and generative approaches that existing frameworks have not incorporated. Even when evaluation metrics are shared, experimental protocols often are not, with differences in data splitting, randomness handling, and hyperparameter tuning can significantly influence empirical outcomes yet are rarely reported in sufficient detail to allow independent replication.

We have developed a benchmarking framework to address these limitations, available at `https://github.com/charmlab/recourse_benchmarks`. The framework explicitly separates predictive model training, recourse generation, and evaluation measurement. From the perspective of (4.1), this design makes explicit which components of the recourse problem are held fixed and which are varied, enabling comparisons that isolate the contribution of the recourse algorithm itself.

The framework is organized around four abstract components that together cover the full experimental pipeline. The *dataset* module handles loading, preprocessing, and the specification of feature properties with standardized preprocessing applied uniformly across all experiments. The *model* module wraps any binary classifier behind a unified prediction interface decoupled from its training framework all expose the same interface to every recourse method and can therefore be used interchangeably without modifying method code. The *method* module defines an abstract recourse generator. Any method that accepts a model interface and a factual instance and returns a recommended action satisfies the contract regardless of whether it uses gradient optimization, integer programming, a genetic algorithm, or any other approach internally. The *evaluation* module computes the full suite of recourse desiderata for any method–dataset–model combination and logs results in a unified format for downstream analysis and comparison.

A central design principle is treating reproducibility as a measurable, graded property rather than a binary outcome. For each recourse method included in the benchmark, a *reproducibility specification* is constructed that records the exact configuration, hyperparameters, preprocessing choices, constraint encodings, needed to reproduce the method's

originally reported results on its original paper's datasets and models. Before a method is admitted to the benchmark, its implementation is validated against this specification under explicit tolerance thresholds. Also we introduce a *graded reproducibility badge system* that communicates the degree of reproduction fidelity across datasets, models, metrics, and protocols, providing transparency about which aspects of reported behavior can be independently recovered and which cannot.

This validation process has itself produced findings of independent methodological interest. Subtle implementation differences, choices about how to encode feature actionability constraints, how to handle categorical features during optimization, what optimizer tolerance to use, and how to set random seeds, can produce meaningfully different recourse costs and validity rates for methods that are nominally identical. These differences are often larger than the performance gaps reported between methods in published comparisons, meaning that the apparent ranking of methods in prior work may reflect implementation choices as much as algorithmic merit. This finding provides concrete empirical grounding for the concern raised by Moreira et al. [82]: that without verified implementations, recourse benchmarks cannot be trusted to measure genuine methodological differences, and that reproduction fidelity must be treated as an experimental concern rather than an afterthought.

Experiments have been conducted across the standard recourse benchmarking datasets over multiple model families, producing a systematic comparison under controlled, reproducible conditions that goes substantially beyond the coverage of CARLA [88] in terms of both method breadth and verification rigor.

As algorithmic recourse research continues to evolve toward more complex settings, multi-agent interactions, performative effects, human-in-the-loop systems, and deployment at scale, standardized benchmarking will become increasingly important. The framework described here positions recourse evaluation on the same footing as other mature areas of machine learning research where shared benchmarks are the norm and progress is measured relative to a common, reproducible baseline. Unified evaluation infrastructure is not merely a supporting tool for the field; it is a prerequisite for the kind of cumulative scientific progress that makes it possible to understand what methods actually achieve, where they fail, and which open problems are genuinely unsolved.

Chapter 5

Summary and Future Work

Algorithmic recourse was conceived as an instrument of individual empowerment: a mechanism by which a person adversely affected by an automated decision could understand what changes would reverse that decision, and act on that understanding. For most of its history, the field has studied this problem from the perspective of a single individual interacting with a single model, in isolation from all others who face the same situation. We have argued that this framing, while analytically tractable, is structurally incomplete. When recourse is deployed at scale, individuals compete for access to providers whose capacity to grant favorable outcomes is finite. In this regime, the individually optimal recourse recommendation for each seeker is, in general, not simultaneously achievable for all seekers, and the resulting welfare loss cannot be recovered by any improvement to how individual recourse is computed. It can only be addressed by treating the allocation of recourse as a system-level problem.

5.1 Summary

The main contribution we have presented is the formal framework for multi-agent algorithmic recourse that takes this system-level view seriously. The starting point is the observation that the individual recourse recommendation may not remain valid when multiple seekers and providers interact in a system. The cost matrix $\mathbf{C} = [c_{ij}]_{n \times m}$ is an immediate output of any standard individual recourse method applied across all seeker–provider pairs, and it determines the weights of a bipartite graph over which coordination happens. This connection situates multi-agent recourse as a principled extension of the individual recourse

formulation rather than a separate problem, and makes the framework agnostic to the choice of underlying recourse method.

The welfare gap, defined as the difference between the individual welfare achievable under unconstrained optimization and the social welfare achievable under capacity constraints, is the central quantity around which the framework is organized. It is non-negative by construction and strictly positive whenever seekers' preferences are heterogeneous and provider capacity is limited, which is the realistic case. The framework addresses the welfare gap through three progressive layers of optimization. The first determines the optimal matching under fixed provider capacities, establishing the social welfare achievable given the current system configuration. The second identifies the capacity distribution that minimizes the welfare gap for a given total system capacity, admitting a closed-form solution via Algorithm 1. The third extends the second by penalizing large deviations from an initial capacity configuration, reflecting real-world constraints on how much providers can be asked to change their operating parameters. Together, the three layers reveal how much welfare is lost by the current system, how much can be recovered through redistribution alone, and at what practical cost.

Coordinated capacity redistribution turns out to be remarkably effective empirically. Across the Two-Moon synthetic benchmark and the real-world COMPAS and Credit datasets, evaluated over four provider model families, Equation (3.2) and Equation (3.3) consistently achieve social welfare above 94% of the individual welfare upper bound with minimal modifications to initial provider configurations. The welfare gap in most settings can be largely eliminated without acquiring additional resources, only by reallocating what already exists more intelligently.

The utilitarian matching objective maximizes aggregate welfare but is indifferent to how that welfare is distributed across seekers. Seekers with inherently high recourse costs may receive systematically worse outcomes even under the socially optimal matching. The inequality-averse extension addresses this by replacing the utilitarian objective with a concave social welfare function $\sum_i u_i^\alpha$, $\alpha \in (0, 1)$, which interpolates between the utilitarian sum ($\alpha \to 1$) and the Rawlsian max-min criterion ($\alpha \to 0$). Rather than imposing demographic constraints, this formulation bends the objective itself toward the worst-off seekers, in the spirit of the welfare-based fairness framework of Heidari et al. [46] and the Rawlsian view that just arrangements are those rational agents would choose from behind the veil of ignorance [96]. The experiments show that the efficiency–equity trade-off introduced by this extension is modest: substantial gains in the equity metric $E = \min_i u_i$ can be achieved at small losses in aggregate welfare, consistent with the bounded price-of-fairness results of Bertsimas et al. [9].

Taken together, these results position algorithmic recourse as a welfare allocation problem, connected to a body of theory and method in welfare economics, social choice, and resource allocation that has not previously been brought to bear on it. The welfare gap is not a failure of individual recourse methods. It is a structural consequence of the mismatch between what is individually optimal and what is collectively feasible under scarcity. System-level coordination is not an optional refinement; it is a prerequisite for recourse to function equitably at scale. Questions that welfare economics has studied for decades, how to aggregate individual utilities, how to characterize the efficiency–equity frontier, how to allocate scarce resources fairly across heterogeneous agents, are directly relevant to the design of recourse systems at scale, and much remains to be built on this foundation. Ultimately, recourse is meant to give people a realistic path to a better outcome. A system that ignores the constraints under which it operates and the interactions among the people it serves will systematically fail the individuals who need it most. It is what makes recourse a credible promise rather than an empty one.

5.2 Future Directions

The framework developed opens several concrete directions for future research. We organize these by proximity to the core contribution, beginning with extensions that follow most directly from the multi-agent recourse setting and moving toward broader open problems in the algorithmic recourse literature that remain relevant to welfare-aware system design.

Dynamic, Temporal, and Performative Recourse. The present framework treats the matching problem as a one-shot interaction: a static population of seekers and providers, fixed classifier parameters, and a central planner with full information about all recourse costs at the time of matching. Real-world recourse systems violate all three of these assumptions simultaneously. *Temporal recourse* refers to settings where individuals interact with a sequence of decisions over time — for example, a loan applicant rejected today who reapplies after taking qualifying steps, only to face an updated model or a different set of competing applicants. Planning recourse across multiple time steps requires treating each seeker's trajectory as an object of optimization, not just a single action [107, 60]. *Dynamic recourse* refers to settings where the population itself changes: new seekers enter, others exit, and provider capacities fluctuate. The welfare gap, as defined in Chapter 2, is a static quantity, but its dynamic analog — the cumulative welfare loss over a stream of arrivals under adaptive capacity constraints — is a natural object for future study, with structural connections to online bipartite matching [26] and bandit regret minimization [15].

Performative recourse refers to the feedback effect identified by König et al. [69] and Fonseca et al. [32]: when recourse recommendations are issued at scale, seekers who act on them collectively shift the data distribution, which may trigger model retraining and invalidate prior recommendations. In the multi-agent setting, this feedback is amplified — coordinated assignments under the matching framework could accelerate distribution shift more rapidly than uncoordinated individual actions. Future work should study how the optimal matching should adapt under repeated interactions, how to maintain welfare gap guarantees under retraining, and how to mitigate feedback-induced harms such as escalating recourse requirements or unequal adaptation across demographic groups [8, 34].

Game-Theoretic Extensions and Strategic Providers. The current framework assumes that providers passively accept the central planner's allocation. In practice, providers are strategic actors: a bank may prefer to serve seekers whose profile closely resembles its existing approved applicants, or may have incentives to attract seekers who are easy to approve rather than those with the greatest need. When providers act strategically, the matching problem becomes a mechanism design problem: the central planner must design an allocation rule that is not only socially optimal but also incentive-compatible, such that providers have no incentive to misreport their capacities or preferences. O'Brien et al. [85] identify this strategic dimension as a key gap in multi-agent recourse, noting that individually rational provider behavior can produce outcomes Pareto-dominated by coordinated solutions. The connection to two-sided matching theory [33] is also relevant here: stable matching mechanisms resist strategic manipulation by construction, and exploring whether stability and welfare optimality can be jointly achieved in the recourse setting is a natural direction. A further extension is to consider seekers who act strategically as well, potentially forming coalitions to improve their collective matching outcome, connecting to the *Algorithmic Collective Action* framework of Baumann et al. [6], in which coordinated feature manipulation can shift algorithmic outcomes in favor of organized groups.

A related practical concern is that providers in real-world settings are rarely passive participants in a centralized allocation. Competing providers, such as banks in a lending market, may have incentives to attract certain seekers and reject others, undermining the feasibility of centralized recourse recommendations. This connects to the mechanism design challenges already noted above, where the central planner must account for strategic provider behavior. A further empirical question concerns how distinct the decision models of providers are in practice. If providers in the same domain converge toward similar classifiers due to shared regulatory constraints or training data, the effective choice set available to seekers may be narrower than our framework assumes, reducing the welfare gains from coordination. Characterizing the degree of model heterogeneity across providers in real

deployments, and studying how it affects the welfare gap, is an important direction for future work.

Fairness-Aware Multi-Agent Recourse. The inequality-averse extension developed in Chapter 3 addresses distributional equity through the shape of the welfare objective, but does not explicitly condition on demographic group membership. A natural extension is to study whether the welfare gap is distributed equitably across groups. Perello et al. [92] show that recourse costs are systematically higher for disadvantaged groups even when classifiers are fair, and Bell et al. [7] demonstrate that this disparity compounds over repeated recourse attempts. In the matching framework, this means that seekers from disadvantaged groups may be disproportionately harmed by capacity constraints, not because the matching algorithm discriminates, but because their higher recourse costs make them less competitive for preferred providers. Future work should characterize how the welfare gap decomposes by demographic subgroup, and whether capacity redistribution strategies that minimize aggregate welfare loss also reduce or exacerbate group-level disparities. Ezzeddine et al. [31] and Yetukuti et al. [123] provide relevant frameworks for incorporating group fairness constraints into recourse generation that could be integrated into the matching layer. Resolving the tension between aggregate welfare maximization and group-level equity in capacity-constrained recourse systems is among the most pressing open problems in this area.

Toward Meaningful Improvement Beyond Boundary Crossing. The current framework defines recourse success as achieving a positive classification outcome under the assigned provider's model. This operationalization inherits a limitation shared by most recourse methods: it treats crossing the decision boundary as the goal, which may be a poor proxy for genuine improvement in the seeker's underlying situation [71]. A seeker who is matched to a provider with a low recourse cost may still receive recommendations that merely game the model without producing durable real-world benefit, for instance, a loan applicant who is advised to temporarily inflate their reported savings without addressing the structural factors that led to their rejection. Konig et la. [71] distinguish interventions that flip a model's prediction from those that produce genuine improvement in the domain, showing that these objectives can diverge substantially. Kanamori et al. [55] extend this perspective to long-term improvement, proposing recourse methods that account for the downstream consequences of actions over extended horizons rather than optimizing a single-step outcome. Integrating these notions of meaningful improvement into the cost function c_{ij}, so that the matching framework favors assignments whose recommendations have durable real-world impact, is an important direction for future work. Doing so would

require domain-grounded cost definitions and evaluation criteria that measure durability and actual impact, rather than solely relying on short-term boundary-crossing validity.

Risk-Aware Provider Capacity Models. A further simplifying assumption in the present framework is that provider capacity is modeled as a fixed scalar, a hard upper bound on the number of seekers a provider can serve. In practice, lenders do not operate with a rigid loan quota; rather, they make risk-adjusted decisions in which each applicant's creditworthiness influences approval probability, and the effective capacity is an emergent property of the portfolio rather than a fixed constraint [78]. Relaxing the scalar capacity assumption to incorporate risk-weighted or probabilistic capacity models, where a provider's willingness to serve a seeker depends jointly on recourse cost and estimated default risk, is a natural and important direction for future work.

References

[1] Model Reconstruction Using Counterfactual Explanations: Mitigating...

[2] Kenneth J Arrow. The theory of risk aversion. *Essays in the theory of risk-bearing*, 90:120, 1971.

[3] Alexander Asemota and Giles Hooker. Longitudinal Counterfactuals: Constraints and Opportunities, February 2024. arXiv:2403.00105 [cs].

[4] Anthony B Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263, September 1970.

[5] Volkan Bakir, Polat Goktas, and Sureyya Akyuz. DiCE-Extended: A Robust Approach to Counterfactual Explanations in Machine Learning, July 2025. arXiv:2504.19027 [cs].

[6] Joachim Baumann and Celestine Mendler-Dünner. Algorithmic Collective Action in Recommender Systems: Promoting Songs by Reordering Playlists. *Advances in Neural Information Processing Systems*, 37:119123–119149, December 2024.

[7] Andrew Bell, Joao Fonseca, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. Fairness in Algorithmic Recourse Through the Lens of Substantive Equality of Opportunity, January 2024. arXiv:2401.16088 [cs].

[8] Andrew Bell, Joao Fonseca, and Julia Stoyanovich. The Game Of Recourse: Simulating Algorithmic Recourse over Time to Improve Its Reliability and Fairness. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD '24, pages 464–467, New York, NY, USA, June 2024. Association for Computing Machinery.

[9] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The Price of Fairness. *Operations Research*, 59(1):17–31, February 2011.

[10] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. On the Efficiency-Fairness Trade-off. *Manage. Sci.*, 58(12):2234–2250, December 2012.

[11] Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. *Decision-Making in Computational Design & Technology eJournal*, December 2017.

[12] Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent Counterfactuals for Deep Models, October 2021. arXiv:2110.03109 [cs].

[13] Dieter Brughmans, Pieter Leyman, and David Martens. NICE: an algorithm for nearest instance counterfactual explanations. *Data Mining and Knowledge Discovery*, 38(5):2665–2703, September 2024.

[14] Dieter Brughmans, Lissa Melis, and David Martens. Disagreement amongst counterfactual explanations: how transparency can be misleading. *TOP*, 32(3):429–462, October 2024.

[15] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems, November 2012. arXiv:1204.5721 [cs].

[16] Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Coverage-Validity-Aware Algorithmic Recourse, January 2025. arXiv:2311.11349 [cs].

[17] Marina Ceccon, Alessandro Fabris, Goran Radanović, Asia J. Biega, and Gian Antonio Susto. Reinforcement Learning for Durable Algorithmic Recourse, September 2025. arXiv:2509.22102 [cs].

[18] Seung Hyun Cheon, Anneke Wernerfelt, Sorelle Friedler, and Berk Ustun. Feature responsiveness scores: Model-agnostic explanations for recourse. In *The Thirteenth International Conference on Learning Representations*, 2025.

[19] Jennifer Chien, Kevin Mckee, Jackie Kay, and William Isaac. Recourse for reclamation: Chatting with generative language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.

[20] Jeffrey Nicholas Clark, Edward Alexander Small, Nawid Keshtmand, Michelle Wing Lam Wan, Elena Fillola Mayoral, Enrico Werner, Christopher Bourdeaux, and Raul Santos-Rodriguez. TraCE: Trajectory Counterfactual Explanation Scores. In *Proceedings of the 5th Northern Lights Deep Learning Conference ({NLDL})*, pages 36–45. PMLR, January 2024.

[21] Richard Cole and Vasilis Gkatzelis. Approximating the Nash social welfare with indivisible items. *SIAM Journal on Computing*, 47(3):1211–1236, 2018.

[22] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 797–806, New York, NY, USA, August 2017. Association for Computing Machinery.

[23] Elliot Creager and Richard Zemel. Online Algorithmic Recourse by Collective Action, December 2023. arXiv:2401.00055 [cs].

[24] Giovanni De Toni, Bruno Lepri, and Andrea Passerini. Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis. *Machine Learning*, 112(4):1389–1409, April 2023.

[25] Javier Del Ser, Alejandro Barredo-Arrieta, Natalia Díaz-Rodríguez, Francisco Herrera, Anna Saranti, and Andreas Holzinger. On generating trustworthy counterfactual explanations. *Information Sciences*, 655:119898, January 2024.

[26] John Dickerson, Karthik Sankararaman, Kanthi Sarpatwar, Aravind Srinivasan, Kung-Lu Wu, and Pan Xu. Online Resource Allocation with Matching Constraints. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, January 2019.

[27] Tri Dung Duong, Qian Li, and Guandong Xu. Causality-based counterfactual explanation for classification models. *Knowledge-Based Systems*, 300:112200, September 2024.

[28] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust Counterfactual Explanations for Tree-Based Ensembles. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5742–5756. PMLR, June 2022.

[29] Ahmad-Reza Ehyaei, Ali Shirali, and Samira Samadi. Collective Counterfactual Explanations: Balancing Individual Goals and Collective Dynamics, September 2025. arXiv:2402.04579 [cs].

[30] Andrew Estornell, Yatong Chen, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Incentivizing Recourse through Auditing in Strategic Classification. *IJCAI*, January 2023.

[31] Fatima Ezzeddine, Obaida Ammar, Silvia Giordano, and Omran Ayoub. Fair Recourse for All: Ensuring Individual and Group Fairness in Counterfactual Explanations, January 2026. arXiv:2601.20449 [cs].

[32] Joao Fonseca, Andrew Bell, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. Setting the Right Expectations: Algorithmic Recourse Over Time. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11, October 2023. arXiv:2309.06969 [cs].

[33] D. Gale and L. S. Shapley. College Admissions and the Stability of Marriage. *The American Mathematical Monthly*, 69(1):9–15, January 1962.

[34] Ruijiang Gao and Himabindu Lakkaraju. On the Impact of Algorithmic Recourse on Social Segregation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10727–10743. PMLR, July 2023.

[35] Prateek Garg, Lokesh Nagalapatti, and Sunita Sarawagi. From Search To Sampling: Generative Models For Robust Algorithmic Recourse, May 2025. arXiv:2505.07351 [cs].

[36] GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), April 2016. Legislative Body: EP, CONSIL.

[37] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, September 2024.

[38] Hangzhi Guo, Thanh H. Nguyen, and Amulya Yadav. CounterNet: End-to-End Training of Prediction Aware Counterfactual Explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 577–589, New York, NY, USA, August 2023. Association for Computing Machinery.

[39] Hangzhi Guo, Xinchang Xiong, Wenbo Zhang, and Amulya Yadav. ReLax: An Efficient and Scalable Recourse Explanation Benchmarking Library using JAX. October 2023.

[40] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024.

[41] Aparajita Haldar, Teddy Cunningham, and Hakan Ferhatosmanoglu. RAGUEL: Recourse-Aware Group Unfairness Elimination. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 666–675, New York, NY, USA, October 2022. Association for Computing Machinery.

[42] Jenny Hamer, Nicholas Perello, Jake Valladares, Vignesh Viswanathan, and Yair Zick. Simple Steps to Success: A Method for Step-Based Counterfactual Explanations, November 2024. arXiv:2306.15557 [cs].

[43] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, pages 12351–12367. PMLR, July 2023.

[44] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust Algorithmic Recourse Under Model Multiplicity With Probabilistic Guarantees. *IEEE Journal on Selected Areas in Information Theory*, 5:357–368, 2024.

[45] Ian Hardy. *Towards Improved Recourse Trade-offs with Adaptive Adversarial Training*. PhD thesis, UC Santa Cruz, 2023.

[46] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[47] Dat Hong and Tong Wang. Personalized Path Recourse for Reinforcement Learning Agents, November 2024. arXiv:2312.08724 [cs].

[48] Safwan Hossain, Evi Micha, and Nisarg Shah. Fair Algorithms for Multi-Agent Multi-Armed Bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 24005–24017. Curran Associates, Inc., 2021.

[49] Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 535–545, New York, NY, USA, January 2020. Association for Computing Machinery.

[50] I-Cheng Yeh. Default of Credit Card Clients, 2009.

[51] Junqi Jiang. Recourse under Model Multiplicity via Argumentative Ensembling. *New Zealand*, 2024.

[52] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. In *Proceedings of the 15th Asian Conference on Machine Learning*, pages 582–597. PMLR, February 2024.

[53] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Interval Abstractions for Robust Counterfactual Explanations. *Artificial Intelligence*, 336:104218, November 2024. arXiv:2404.13736 [cs].

[54] Junqi Jiang, Antonio Rago, Francesco Leofante, and Francesca Toni. Argumentative Ensembling for Robust Recourse under Model Multiplicity, June 2025. arXiv:2506.20260 [cs].

[55] Kentaro Kanamori, Ken Kobayashi, and Takuya Takagi. Learning Gradient Boosted Decision Trees with Algorithmic Recourse. October 2025.

[56] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Learning Decision Trees and Forests with Algorithmic Recourse, June 2024. arXiv:2406.01098 [cs].

[57] Sai Srinivas Kancheti, Rahul Vigneswaran, Bamdev Mishra, and Vineeth N Balasubramanian. Hare: Human-in-the-loop algorithmic recourse. *Transactions on Machine Learning Research*.

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

[59] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.*, 55(5):95:1–95:29, December 2022.

[60] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 265–277. Curran Associates, Inc., 2020.

[61] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Towards Causal Algorithmic Recourse. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 139–166. Springer International Publishing, Cham, 2022.

[62] Kshitij Kayastha, Vasilis Gkatzelis, and Shahin Jabbari. Learning-Augmented Robust Algorithmic Recourse. October 2024.

[63] Eoin M. Kenny, Weipeng Huang, Saugat Aryal, and Mark T. Keane. Explaining and Auditing with "Even-If": Uses for Semi-factual Explanations in AI/ML. In Ireneusz Czarnowski, Robert J. Howlett, and Lakhmi C. Jain, editors, *Intelligent Decision Technologies*, pages 135–145, Singapore, 2025. Springer Nature.

[64] Zahra Khotanlou, Kate Larson, and Amir-Hossein Karimi. From Individual to Multi-Agent Algorithmic Recourse: Minimizing the Welfare Gap via Capacitated Bipartite Matching, August 2025. arXiv:2508.11070 [cs].

[65] Dongwhi Kim and Nuno Moniz. Relevance-aware Algorithmic Recourse, May 2024. arXiv:2405.19072 [cs].

[66] Hee-Dong Kim, Yeong-Joon Ju, Jung-Ho Hong, and Seong-Whan Lee. CIRF: Importance of related features for plausible counterfactual explanations. *Information Sciences*, 678:120974, September 2024.

[67] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[68] Alejandro Kuratomi, Ioanna Miliou, Zed Lee, Tony Lindgren, and Panagiotis Papapetrou. Ijuice: integer JUstIfied counterfactual explanations. *Machine Learning*, 113(8):5731–5771, August 2024.

[69] Gunnar König, Hidde Fokkema, Timo Freiesleben, Celestine Mendler-Dünner, and Ulrike von Luxburg. Performative Validity of Recourse Explanations, November 2025. arXiv:2506.15366 [stat].

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

[71] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. Improvement-Focused Causal Recourse (ICR). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):11847–11855, June 2023.

[72] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the Fairness of Causal Algorithmic Recourse. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9584–9594, June 2022.

[73] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. ProPublica COMPAS Analysis, 2016.

[74] Francesco Leofante and Nico Potyka. Promoting Counterfactual Robustness through Diversity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21322–21330, March 2024.

[75] David Lewis. *Counterfactuals*. Blackwell, Malden, Mass., 1973.

[76] Benton Li, Nativ Levy, Brit Youngmann, Sainyam Galhotra, and Sudeepa Roy. Fair and Actionable Causal Prescription Ruleset. *Proc. ACM Manag. Data*, 3(3):156:1–156:28, June 2025.

[77] Zhuoshu Li, Kelsey Lieberman, William Macke, Sofia Carrillo, Chien-Ju Ho, Jason Wellen, and Sanmay Das. Incorporating Compatible Pairs in Kidney Exchange: A Dynamic Weighted Matching Model. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pages 349–367, New York, NY, USA, June 2019. Association for Computing Machinery.

[78] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed Impact of Fair Machine Learning. pages 6196–6200, 2019.

[79] Ayan Majumdar and Isabel Valera. CARMA: A practical framework to generate recommendations for causal algorithmic recourse at scale. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1745–1762, New York, NY, USA, June 2024. Association for Computing Machinery.

[80] Donato Maragno, Jannis Kurtz, Tabea E. Röber, Rob Goedhart, Ş. İlker Birbil, and Dick den Hertog. Finding Regions of Counterfactual Explanations via Robust Optimization. *INFORMS Journal on Computing*, 36(5):1316–1334, September 2024.

[81] Kleopatra Markou, Dimitrios Tomaras, Vana Kalogeraki, and Dimitrios Gunopulos. A Framework for Feasible Counterfactual Exploration incorporating Causality, Sparsity and Density. In *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, pages 254–261, May 2024. ISSN: 2473-3490.

[82] Catarina Moreira, Yu-Liang Chou, Chihcheng Hsieh, Chun Ouyang, João Pereira, and Joaquim Jorge. Benchmarking Instance-Centric Counterfactual Algorithms for XAI: From White Box to Black Box. *ACM Comput. Surv.*, 57(6):145:1–145:37, February 2025.

[83] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

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

[85] Andrew O'Brien and Edward Kim. Toward Multi-Agent Algorithmic Recourse: Challenges From a Game-Theoretic Perspective. *The International FLAIRS Conference Proceedings*, 35, May 2022.

[86] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. Achieving Fairness in the Stochastic Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 22(174):1–31, 2021.

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

[88] Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. June 2021.

[89] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse, October 2023. arXiv:2203.06768 [cs].

[90] Paraskevas Pegios, Aasa Feragen, Andreas Abildtrup Hansen, and Georgios Arvanitidis. Counterfactual Explanations via Riemannian Latent Space Traversal, November 2024. arXiv:2411.02259 [cs].

[91] Sikha Pentyala, Shubham Sharma, Sanjay Kariyappa, Freddy Lecue, and Daniele Magazzeni. Privacy-Preserving Algorithmic Recourse, November 2023. arXiv:2311.14137 [cs].

[92] Nicholas Perello, Cyrus Cousins, Yair Zick, and Przemyslaw Grabowicz. Discrimination Induced by Algorithmic Recourse Objectives. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 1653–1663, New York, NY, USA, June 2025. Association for Computing Machinery.

[93] John W. Pratt. Risk Aversion in the Small and in the Large. In Georges Dionne and Scott E. Harrington, editors, *Foundations of Insurance Economics: Readings in Economics and Finance*, pages 83–98. Springer Netherlands, Dordrecht, 1992.

[94] Peyman Rasouli and Ingrid Chieh Yu. Care: Coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, 17(1):13–38, 2024.

[95] Kaivalya Rawal and Himabindu Lakkaraju. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. In *Advances in Neural Information Processing Systems*, volume 33, pages 12187–12198. Curran Associates, Inc., 2020.

[96] John Rawls. *A theory of justice*. Belknap Press of Harvard Univ. Press, Cambridge, Mass, rev. ed., 5.- 6. printing edition, 2003.

[97] David Rea, Craig Froehle, Suzanne Masterson, Brian Stettler, Gregory Fermann, and Arthur Pancioli. Unequal but Fair: Incorporating Distributive Justice in Operational Allocation Models. *Production and Operations Management*, 30(7):2304–2320, July 2021.

[98] Annabelle Redelmeier, Martin Jullum, Kjersti Aas, and Anders Løland. MCCE: Monte Carlo sampling of valid and realistic counterfactual explanations for tabular data. *Data Mining and Knowledge Discovery*, 38(4):1830–1861, July 2024.

[99] Giorgio Sacchi. *Computationally Efficient Explainable AI: Bayesian Optimization for Computing Multiple Counterfactual Explanantions*. 2023.

[100] Shubham Sharma, Alan Gee, Jette Henderson, and Joydeep Ghosh. FASTER-CE: Fast, Sparse, Transparent, and Robust Counterfactual Explanations. In Ilias Maglogiannis, Lazaros Iliadis, John Macintyre, Markos Avlonitis, and Antonios Papaleonidas, editors, *Artificial Intelligence Applications and Innovations*, pages 183–196, Cham, 2024. Springer Nature Switzerland.

[101] Manan Singh, Sai Srinivas Kancheti, Shivam Gupta, Ganesh Ghalme, Shweta Jain, and Narayanan C. Krishnan. Algorithmic Recourse based on User's Feature-order Preference. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, CODS-COMAD '23, pages 293–294, New York, NY, USA, January 2023. Association for Computing Machinery.

[102] Edward A. Small, Jeffrey N. Clark, Christopher J. McWilliams, Kacper Sokol, Jeffrey Chan, Flora D. Salim, and Raul Santos-Rodriguez. Counterfactual Explanations via Locally-guided Sequential Algorithmic Recourse, September 2023. arXiv:2309.04211 [cs].

[103] Muhammad Suffian, Jose M. Alonso-Moral, and Alessandro Bogliolo. Introducing User Feedback-Based Counterfactual Explanations (UFCE). *International Journal of Computational Intelligence Systems*, 17(1):123, May 2024.

[104] Giovanni De Toni, Paolo Viappiani, Stefano Teso, Bruno Lepri, and Andrea Passerini. Personalized Algorithmic Recourse with Preference Elicitation, January 2024. arXiv:2205.13743 [cs].

[105] Bohdan Turbal, Iryna Voitsitska, and Lesia Semenova. ElliCE: Efficient and Provably Robust Algorithmic Recourse via the Rashomon Sets. October 2025.

[106] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable Algorithmic Recourse, July 2021. arXiv:2102.13620 [cs].

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

[108] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *ACM Comput. Surv.*, 56(12):312:1–312:42, October 2024.

[109] Sahil Verma, Keegan Hines, and John P. Dickerson. Amortized Generation of Sequential Algorithmic Recourses for Black-Box Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8512–8519, June 2022.

[110] Sahil Verma, Ashudeep Singh, Varich Boonsanong, John P Dickerson, and Chirag Shah. Recrec: Algorithmic recourse for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4325–4329, 2023.

[111] Marco Virgolin and Saverio Fracaros. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, 316:103840, March 2023.

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

[113] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, 2017.

[114] Haotian Wang, Hao Zou, Xueguang Zhou, Shangwen Wang, Long Lan, Wenjing Yang, and Peng Cui. Learning Feasible Causal Algorithmic Recourse without Prior Structure. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 805–813, December 2024. ISSN: 2375-9259.

[115] Haotian Wang, Hao Zou, Xueguang Zhou, Shangwen Wang, Wenjing Yang, and Peng Cui. Learning feasible causal algorithmic recourse: A prior structural knowledge free approach. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 45074518, New York, NY, USA, 2025. Association for Computing Machinery.

[116] Shuang Wang and Yongkai Wu. Achieving Fairness through Constrained Recourse. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2024. ISSN: 2161-4407.

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

[118] Zimu Wang, Hao Zou, Han Yu, Shaohua Fan, Haotian Wang, Yue He, and Peng Cui. PDMC: Generating Feasible Algorithmic Recourse via Perturbation Data Manifold Constraint. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pages 3111–3121, New York, NY, USA, August 2025. Association for Computing Machinery.

[119] Joshua Nathaniel Williams, Anurag Katakkar, Hoda Heidari, and J. Zico Kolter. Rethinking Distance Metrics for Counterfactual Explainability, October 2024. arXiv:2410.14522 [cs].

[120] Haochen Wu, Shubham Sharma, Sunandita Patra, and Sriram Gopalakrishnan. SafeAR: Safe Algorithmic Recourse by Risk-Aware Policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15915–15923, March 2024. arXiv:2308.12367 [cs].

[121] Yueqing Xuan, Kacper Sokol, Mark Sanderson, and Jeffrey Chan. Perfect counterfactuals in imperfect worlds: modelling noisy implementation of actions in sequential algorithmic recourse. *Machine Learning*, 114(8):187, July 2025.

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

[123] Jayanth Yetukuri, Ian Hardy, Yevgeniy Vorobeychik, Berk Ustun, and Yang Liu. Providing Fair Recourse over Plausible Groups. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21753–21760, March 2024.

[124] Jayanth Yetukuri, Ian Hardy, Yevgeniy Vorobeychik, Berk Ustun, and Yang Liu. Providing fair recourse over plausible groups. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[125] Songming Zhang, Xiaofeng Chen, Shiping Wen, and Zhongshan Li. Density-based reliable and robust explainer for counterfactual explanation. *Expert Systems with Applications*, 226:120214, September 2023.

APPENDICES

A.1 Proof of Theorem 3.4.1

We prove that:

$$\lim_{\gamma\to\infty}\frac{\mathrm{SW}(\gamma)}{\mathrm{IW}(\gamma)}=\frac{\max_{z\in\mathcal{Z}}\sum_{i\in S_{\min}}\sum_{j\in J_i^*}z_{ij}}{|S_{\min}|}$$

Notation. Let $c_{\min}=\min_{i,j}c_{ij}$ denote the globally smallest recourse cost across all seeker-provider pairs. Let $c_i^*=\min_j c_{ij}$ denote the best achievable cost for seeker i, and define the set of seekers with globally minimal best cost:

$$S_{\min}=\{i:c_i^*=c_{\min}\}.$$

For each seeker i, define the set of minimum-cost providers:

$$J_i^*=\{j:c_{ij}=c_i^*\}.$$

The individual welfare and social welfare are:

$$\mathrm{IW}(\gamma)=\sum_{i=1}^{n}e^{-\gamma c_i^*},\qquad \mathrm{SW}(\gamma)=\max_{z\in\mathcal{Z}}\sum_{i=1}^{n}\sum_{j=1}^{m}e^{-\gamma c_{ij}}z_{ij},$$

where $\mathcal{Z}$ is the finite set of feasible binary matchings under the capacity and matching constraints of Equation (**??**). Since $\mathcal{Z}$ is finite, the maximum in $\mathrm{SW}(\gamma)$ is always attained.

Step 1: Factor out the dominant exponential term. Factor $e^{-\gamma c_{\min}}$ from both numerator and denominator:

$$\frac{\mathrm{SW}(\gamma)}{\mathrm{IW}(\gamma)}=\frac{\max_{z\in\mathcal{Z}}\sum_{i=1}^{n}\sum_{j=1}^{m}e^{-\gamma(c_{ij}-c_{\min})}z_{ij}}{\sum_{i=1}^{n}e^{-\gamma(c_i^*-c_{\min})}}.$$

Since $c_{ij} \geq c_i^* \geq c_{\min}$ for all i, j, every exponent in both numerator and denominator is non-negative, so all terms lie in $[0, 1]$ for $\gamma > 0$.

Step 2: Limit of the denominator. For each seeker i:

$$e^{-\gamma(c_i^* - c_{\min})} \xrightarrow{\gamma \to \infty} \begin{cases} 1 & \text{if } i \in S_{\min}, \\ 0 & \text{if } i \notin S_{\min}. \end{cases}$$

since $c_i^* - c_{\min} = 0$ for $i \in S_{\min}$ and $c_i^* - c_{\min} > 0$ otherwise. By dominated convergence (the sum is finite):

$$\sum_{i=1}^{n} e^{-\gamma(c_i^* - c_{\min})} \xrightarrow{\gamma \to \infty} |S_{\min}|.$$

Step 3: Limit of each term in the numerator. For any fixed feasible matching $z \in \mathcal{Z}$ and any seeker-provider pair (i, j):

$$e^{-\gamma(c_{ij} - c_{\min})} \xrightarrow{\gamma \to \infty} \begin{cases} 1 & \text{if } c_{ij} = c_{\min}, \\ 0 & \text{if } c_{ij} > c_{\min}. \end{cases}$$

Now $c_{ij} = c_{\min}$ is possible only when $i \in S_{\min}$ and $j \in J_i^*$, since for $i \in S_{\min}$ we have $c_i^* = c_{\min}$. Therefore, for any fixed $z \in \mathcal{Z}$, since each seeker is matched to at most one provider ($\sum_j z_{ij} \leq 1$):

$$\sum_{i=1}^{n} \sum_{j=1}^{m} e^{-\gamma(c_{ij} - c_{\min})} z_{ij} \xrightarrow{\gamma \to \infty} \sum_{i \in S_{\min}} \sum_{j \in J_i^*} z_{ij},$$

which counts the number of seekers in $S_{\min}$ matched to one of their minimum-cost providers under z.

Step 4: Pointwise limit of $F(z, \gamma)$. For each fixed $z \in \mathcal{Z}$, define:

$$F(z, \gamma) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} e^{-\gamma(c_{ij} - c_{\min})} z_{ij}}{\sum_{i=1}^{n} e^{-\gamma(c_i^* - c_{\min})}}.$$

By Steps 2 and 3, both numerator and denominator converge as $\gamma \to \infty$, and the denominator converges to $|S_{\min}| > 0$. Therefore:

$$F(z, \gamma) \xrightarrow{\gamma \to \infty} \frac{\sum_{i \in S_{\min}} \sum_{j \in J_i^*} z_{ij}}{|S_{\min}|}.$$

Step 5: Exchange limit and maximum. Since $\mathcal{Z}$ is a *finite* set, we can exchange the limit and the maximum:

$$\lim_{\gamma\to\infty}\frac{\mathrm{SW}(\gamma)}{\mathrm{IW}(\gamma)}=\lim_{\gamma\to\infty}\max_{z\in\mathcal{Z}}F(z,\gamma)=\max_{z\in\mathcal{Z}}\lim_{\gamma\to\infty}F(z,\gamma)=\frac{\max_{z\in\mathcal{Z}}\sum_{i\in S_{\min}}\sum_{j\in J_i^*}z_{ij}}{|S_{\min}|}.$$

The right-hand side is the maximum fraction of seekers in $S_{\min}$ who can simultaneously be matched to one of their minimum-cost providers under the fixed capacity vector $\hat{k}$. It is determined by $\hat{k}$ and the structure of the minimum-cost edges $\{J_i^*\}_{i\in S_{\min}}$, and is independent of non-minimal cost magnitudes. □

A.2 Provider Classification Performance

Table A.1 reports accuracy, precision, and recall for all 15 providers used in the large-scale Credit and COMPAS experiments, evaluated on the held-out test set using each provider's tier-specific classification threshold. The table confirms heterogeneity in predictive behavior across providers, particularly in recall, which reflects the varying rates at which providers reject applicants and consequently determines the difficulty of obtaining recourse from each provider.

Table A.1: Classification performance of the 15 providers on the held-out test set (large-scale DiCE setting). Predictions use each provider's tier-specific threshold.

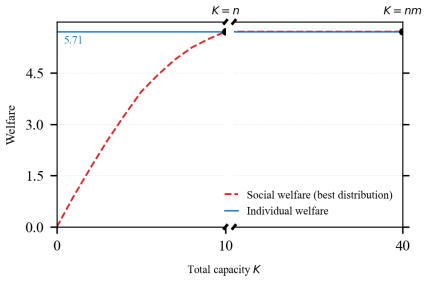
Provider	**Accuracy**	**Precision**	**Recall**
LR_1	0.733	0.883	0.758
Tree_2	0.703	0.893	0.703
RF_3	0.699	0.891	0.698
MLP_4	0.669	0.899	0.648
LR_5	0.709	0.887	0.718
Tree_6	0.703	0.893	0.703
RF_7	0.652	0.900	0.622
MLP_8	0.550	0.917	0.465
LR_9	0.419	0.907	0.283
Tree_10	0.221	0.000	0.000
RF_11	0.755	0.881	0.793
MLP_12	0.735	0.877	0.767
LR_13	0.733	0.883	0.758
Tree_14	0.796	0.851	0.895
RF_15	0.745	0.874	0.785

A.3 Additional Experiments

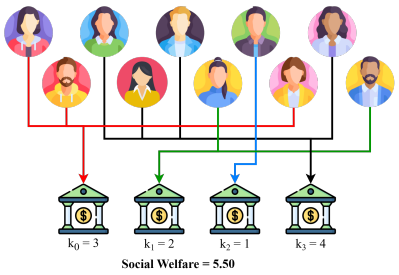
To test robustness to the distance function, we repeated all experiments with the ℓ_1 norm instead of the ℓ_∞ norm. The datasets, provider models, and optimization procedure were unchanged, except for minor parameter adjustments. We kept the setup and γ the same but increased β to 0.05. As shown in Figure A.1, the framework still achieves near-optimal welfare with minimal adjustments, and quantitative results for all datasets are summarized in Table A.2.

$$\mathbf{W} = \begin{bmatrix} 0.256 & 0.104 & 0.256 & 0.353 \\ 0.383 & 0.096 & 0.366 & 0.441 \\ 0.179 & 0.148 & 0.129 & 0.902 \\ 0.415 & 0.064 & 0.432 & 0.800 \\ 0.187 & 0.237 & 0.168 & 0.694 \\ 0.299 & 0.499 & 0.178 & 0.393 \\ 0.165 & 0.214 & 0.198 & 0.171 \\ 0.211 & 0.021 & 0.230 & 0.251 \\ 0.509 & 0.049 & 0.535 & 0.727 \\ 0.368 & 0.830 & 0.314 & 0.395 \end{bmatrix}$$

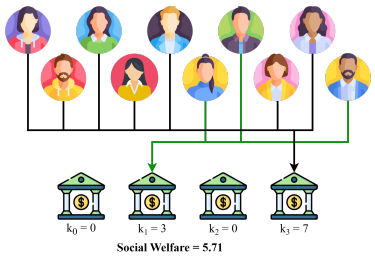
(a) Illustration of Algorithm 1 on the MACE-generated weight matrix.



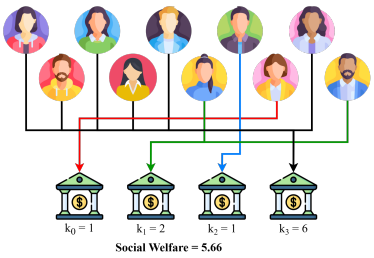
(b) Social welfare and individual-welfare plotted as the total capacity K grows from 0 to $n * m$.



(c) Optimal seeker-to-provider assignments obtained by solving Equation (3.1) under the initial capacity vector.



(d) Provider capacities distributed via Equation (3.2), eliminating the welfare gap.



(e) Final matching (Equation (3.3)) balancing welfare against capacity deviation costs.

Figure A.1: Illustration of our proposed framework: Weight matrix and individually optimal matches (a), Welfare curve showing the gap between social and individual optima. (c–e) Optimization process: transitioning from initial assignments (c) to welfare-maximizing allocations (d), and balancing welfare gains against capacity adjustment costs (e).

Table A.2: Comparative results of the three optimization formulations. Each cell shows social welfare (SW), capacity vector $\mathbf{k}$, and the percentage of the individual-welfare upper bound (IW) attained in each optimization problem.

	Two-Moon			**Credit**			**COMPAS**		
	Capacity	SW	% IW	Capacity	SW	% IW	Capacity	SW	% IW
Equation (3.1)	$(3, 2, 1, 4)$	5.50	96.32%	$(3, 2, 6, 1)$	7.49	87.04%	$(3, 8, 1, 3)$	12.03	95.74%
Equation (3.2)	$(0, 3, 0, 7)$	5.71	100%	$(4, 0, 3, 5)$	8.61	100%	$(11, 1, 3, 0)$	12.57	100%
Equation (3.3)	$(1, 2, 1, 6)$	5.66	99.03%	$(3, 0, 6, 3)$	8.44	98.03%	$(5, 6, 1, 3)$	12.26	97.53%