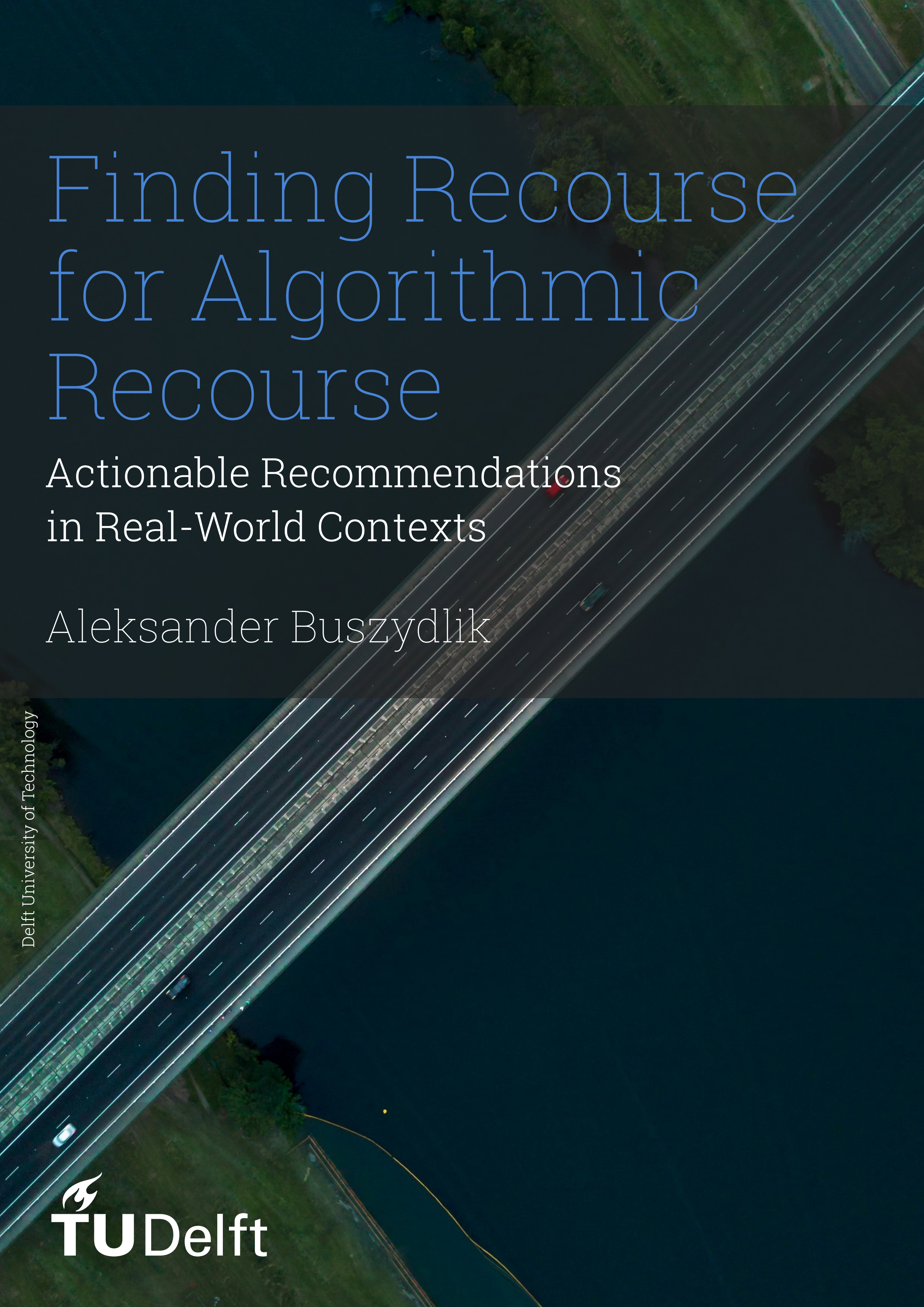

Finding Recourse for Algorithmic Recourse

Actionable Recommendations in Real-World Contexts

Aleksander Buszydlik

Delft University of Technology

This page is unintentionally left blank.

Finding Recourse for Algorithmic Recourse

Actionable Recommendations in Real-World Contexts

by

Aleksander Buszydlik

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Wednesday, November 20, 2024 at 15:00.

Student number:	5090458	
Project duration:	November 20, 2023 – November 20, 2024	
Thesis committee:	Dr. Ir. Cynthia C. S. Liem,	TU Delft, supervisor
	Dr. Ir. Roel I. J. Dobbe,	TU Delft, supervisor
	Patrick Altmeyer,	TU Delft, supervisor
	Dr. Pradeep K. Murukannaiah,	TU Delft, committee member

Cover: “*Cars Passing Through Bridge*” by Mudassir Ali via Pexels under the Pexels License

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

“C’est ce goût de l’amour, ce goût donc qui m’a poussé aujourd’hui
à entreprendre une construction mécanique, mais demain qui sait?
Peut-être simplement à me mettre au service de la communauté,
à faire le don... le don de soi.”

It is this taste for love, it pushed me to work on a mechanical construction today, but tomorrow, who knows? It may, perhaps, lead me to serve the community, to give the gift... the gift of myself.

– Edouard Baer as Otis in Asterix & Obelix: Mission Cleopatra

Preface

Twelve months have passed since I started working on this project. As I think about this time, I recognize that it has been one of the most formative periods of my life – especially in terms of my academic interests – and I believe this is reflected on the pages of this thesis in the values that it attempts to promote and in the multidisciplinary approach that it takes. I found the process that culminated in this draft to be a challenging but incredibly rewarding puzzle. Many people contributed a piece (or several) to the final image. I want to extend my deepest gratitude to all of them.

First, I am immensely grateful to my supervisors – Dr. Cynthia C. S. Liem, Dr. Roel I. J. Dobbe, and Patrick Altmeyer – whose guidance helped me find focus in this significant project but also left enough space to satisfy my personal curiosity.

Cynthia, you are a big inspiration for me, both academically and personally. It has been a privilege to pursue several projects under your supervision; each of them helped me better define my envisioned career path. I appreciate your sage advice, which elevated this project to its fullest potential.

Roel, I profoundly admire your ability to link concepts across different fields and communicate them in an insightful, precise, and engaging way. Your vision for AI systems had a transformative impact on my thoughts on this topic, and through that, it was instrumental in shaping this thesis.

Patrick, I am thankful for all of your great ideas, insightful feedback, and for ensuring that I feel supported at all stages of this work. Your work ethic sets the standards I aspire to reach.

I also want to recognize several people who allowed me to tap into their expertise and shared many brilliant suggestions; while some did not find their way into the current draft, I hope to revisit them in the follow-up works. I wish to extend my sincere thanks to:

- Dr. Pradeep K. Murukannaiah, who encouraged me to think about the values that influence the design of artificial intelligence systems and graciously agreed to join my thesis committee;
- Dr. Luciano Cavalcante Siebert, who helped me situate my research within the broader field of meaningful human control and motivated me with his keen interest in my work;
- Prof. Ibo van de Poel, who discussed with me the ethical dilemmas of algorithmic recourse and suggested how to reason about such intangible concepts;
- Dr. Ujwal Gadiraju, who highlighted the connections between my research and the fascinating problem of artificial intelligence supply chains;
- Andrew Demetriou and Dr. Bernd Dudzik, who guided me through the process of carrying out a literature review and made this task a little bit less daunting;
- All members of the STEAD Systems Group and especially Íñigo Martínez de Rituerto de Troya and Sem Nouws, who provided me with valuable feedback at various stages of my research;
- Dr. Richard Grimes, who offered invaluable assistance in preserving the artifacts of this research and navigating the complex licensing issues;
- Jurriaan Parie, who provided me with some initial ideas on the legal landscape of my topic;
- and several other people who volunteered their time to talk to me and to whom confidentiality was extended in this thesis, most notably the experts that informed Chapter 7.

Next, I am indebted to Dr. Gosia Migut and Dr. Tom Viering. I am fortunate to have been mentored by them throughout my university years. Gosia, I am grateful for your trust and confidence, and for checking in on me whenever I have been overloaded with work. Tom, I deeply appreciate your enthusiasm for my studies, and the warm and welcoming atmosphere you always create.

Lastly, I want to express my sincere thanks to several people who are very important in my life.

To Karol, for a friendship that has continued from high school through all of our studies in Delft. Thank you for the mutual learning, your bright ideas, your inquisitive attitude, and your thoughtful feedback. I am also thankful for all the experiences that we have shared along the way.

To Eva, for your unwavering support, uplifting encouragement, and very important reminders (and initiative) to take breaks. I am also grateful for your many insightful suggestions about this draft.

To Lucia, for your unshakable optimistic attitude, your willingness to put up with my thesis musings, and your ability to distract from them with memes and reels.

To Aleksandra, to Paula, and to Shruthi, for many heartfelt chats and for your emotional support, but also for many fun activities. Your friendship has been a constant source of comfort.

To all the other amazing friends I made throughout my five years in The Netherlands: in my classes, as a teaching assistant, through the Faculty Student Council, and while climbing at Beest Boulders HS.

To my parents, Elżbieta and Jerzy, and to my grandmas, Jadwiga and Marianna, for your love and care, for showing great interest in my work, and for unconditionally believing in me. Dziękuję.

Aleksander Buszydlik
Den Haag, November 2024

Abstract

The aim of algorithmic recourse (AR) is generally understood to be the provision of "actionable" recommendations to individuals affected by algorithmic decision-making systems in an attempt to present them with the capacity to take actions that would guarantee more desirable outcomes in the future. Over the past few years, the literature has predominantly focused on the development of *solutions* to generate "actionable" counterfactual explanations that further satisfy various desiderata, such as diversity or robustness. We believe that algorithmic recourse, by its nature, should be seen as a practical challenge: real-world decision-making systems are complex dynamic entities involving various actors – end users, domain experts, system owners, etc. – engaging in social and technical processes. Thus, research on algorithmic recourse should account for the characteristics of systems where such mechanisms could be implemented. This necessitates a rich understanding of the *problem* space of AR but, as we observe, it remains largely uncharted in the existing literature.

We focus on algorithmic recourse in real-world contexts, applying Design Science Research methods to bridge the gap between its technical affordances and the social constraints of real-world decision-making systems where it could be applied. First, we conduct a systematized literature review of 127 publications to learn about the authors' perception of the problem. Next, we consider a case study of a risk profiling model developed to support the authorities of a major Dutch city in the detection of welfare fraud. We employ a desk research approach to learn about the system, reinforce our understanding of the requirements for algorithms in public administration settings through interviews with experts, and make use of accident analysis methodologies to theorize about the value of AR interventions in this setting. We draw on these insights to propose a conceptual framework for the evaluation of AR in real-world contexts and provide its proof-of-concept instantiation as a simulation tool that facilitates the study of such mechanisms within decision-making processes. Finally, we design and prove an algorithm to generate actionable recommendations in expert systems. These are commonly used in public administration systems but overlooked in existing research.

On the example of our endeavor, we learn about the ways to strengthen the connections between the problem space and the solution space of algorithmic recourse. We argue that AR can be discussed on three levels of complexity: (1) as actionable recommendations, (2) as the process of improving outcomes, or (3) as the task of developing mechanisms to support end-users in this process. We advocate for computer science authors to focus on the final, broadest meaning of the challenge to improve the applicability of their solutions in real-world contexts. We also encourage researchers from other fields to contribute their perspectives and for practitioners to support further research by building upon our approach to reason about the place for AR solutions in their domains of expertise.

Contents

Appendix **108**

List of Figures

List of Tables

List of Algorithms

Glossary of acronyms

General

ABM	Agent-based model(ing)
ADM	Algorithmic decision-making
AI	Artificial intelligence
AR	Algorithmic recourse
CE	Counterfactual explanation
DEVS	Discrete-Event System Specification
DSR	Design Science Research
DTSS	Discrete-Time System Specification
GAN	Generative adversarial network
ML	Machine learning
SCM	Structural causal model
STAMP	Systems-Theoretic Accident Model and Processes
STPA	Systems-Theoretic Process Analysis
VAE	Variational autoencoder

Institutions

CBS	Statistics Netherlands (*Centraal Bureau voor de Statistiek*)
CSO	Civil society organizations (*maatschappelijke organisaties)*
DPA	Data Protection Authority (*Autoriteit Persoonsgegevens*)
OBI	Team Research and Business Intelligence (*Onderzoek en Business Intelligence*)
SVB	Social Insurance Bank (*Sociale Verzekeringsbank*)
SZW	Ministry of Social Affairs and Employment (*Ministerie van Sociale Zaken en Werkgelegenheid*)
T&T	Team Testing and Monitoring (*Toetsing en Toezicht*)
THO	Team Reinvestigations (*Team HetOnderzoeken*)
UWV	Employee Insurance Agency (*Uitvoeringsinstituut Werknemersverzekeringen*)
W&I	Work & Income (*Werk & Inkomen*)

Legal acts

Wet SUWI	W&I Implementation Agencies Structure Act (*Wet Structuur Uitvoeringsorganisatie W&I*)
GALA (*Awb*)	General Administrative Law Act (*Algemene wet bestuursrecht*)
GDPR (*AVG*)	General Data Protection Regulation (*Algemene Verordening Gegevensbescherming*)
Wmo	Social Support Act (*Wet maatschappelijke ondersteuning*)
WW	Unemployment Act (*Werkloosheidswet*)

Introduction 1

1.1 Motivation

Algorithmic decision-making (ADM) – the use of algorithmic tools to make automated decisions or support the decisions of humans – is increasingly used to improve the "throughput", "stability", or "objectivity" of decisions in high-stakes domains such as healthcare [89, 160], justice system [8, 209], or public administration [91, 255]. While regulatory frameworks, including the **General Data Protection Regulation (GDPR)** and the **Artificial Intelligence Act (AI Act)**, ban fully-automated decisions and require mechanisms for human oversight, it is not clear what form they ought to take.

Art. 22(1) GDPR: *"The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."*

Art. 14(1) AI Act: *"High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use."*

At the same time, human oversight procedures were not enough to prevent multiple high-profile failures of algorithmic decision-making systems. The Netherlands alone has experienced the fallout of the childcare benefits scandal (toeslagenaffaire), System Risk Indication (SyRI), and automated risk profiling for social assistance in several municipalities, including Rotterdam. These highlight the importance of contestability throughout model lifecycles, such as procedures to ensure the agency of the affected end-users.

A solution that has recently gained traction in computer science research is algorithmic recourse (AR), which is the focus of this thesis. **AR aims** to provide (non-expert) model users with *"actionable"* recommendations on how to modify the prediction outcomes. For example, a person who unsuccessfully applied for social welfare assistance could receive a recommendation like *"if you had met with the case worker more often, you would have qualified for the benefits"*.

While there is some disagreement about its practical goals, for the purposes of this research **we rely on the following operational definition of algorithmic recourse:** AR involves the provision of recommendations aligned with the preferences of non-expert users in an attempt to help them improve outcomes in an ADM setting.

Still, AI systems rely on more than just the technical components; they involve various stakeholders engaged in decision-making processes. By extension, this also applies to algorithmic recourse. It is an inherently practical problem that resembles a bureaucratic complaint process: an individual unhappy with some decision engages with a representative of the issuing organization in an attempt to overturn it. Yet, we observe that much of the existing work is highly theoretical, with little consideration of whether it could be applied in organizational settings [see also 34]. Deploying AR in real-world systems without analyzing its mechanics in a broader context and without knowing what types of dynamics are expected to arise is bound to lead to unanticipated outcomes. Many of them will be undesirable and even potentially unsafe, and impossible to validate with respect to a set of requirements because the requirements for AR are *necessarily* socio-technical.

This thesis attempts to provide a more nuanced perspective on the problem. We engage in Design Science Research to learn about the potential value of AR in the example setting of social assistance re-investigations under the Participation Act. We apply qualitative methods – including a literature review, expert interviews, and desk research – as well as quantitative computational experiments and algorithm design to provide a comprehensive evaluation of algorithmic recourse mechanisms. Consequently, we explore how algorithms can be responsibly integrated into a highly consequential setting that directly concerns vulnerable populations.

1.2 Contributions

Algorithmic recourse has seen an explosion of interest over the past several years, with dozens of methods proposed to generate "actionable explanations" of **black-box model** decisions. Our overarching goal in this thesis is to look beyond this hype and evaluate the practical value of algorithmic recourse in real-world contexts. Specifically, we make the following four contributions to the state-of-the-art.

We fully acknowledge the ACM call for more inclusive language in computing and its suggestion to discuss "opaque" and "clear-box" models [21]. We subscribe to the **"black-box" and "white-box" distinction** as it is the prevailing nomenclature in algorithmic recourse literature. We follow all other recommendations of ACM, including the use of singular "they/them" in examples.

First, we engage in a systematized review (meta-research) of the literature on algorithmic recourse to learn about the practical considerations that underlie existing contributions (Chapter 4). This allows us to identify five key shortcomings that are pervasive in the existing research and likely to diminish its value outside the lab.

Second, we look at algorithmic recourse through the lens of a concrete real-world socio-technical system of social assistance benefits in the Netherlands. We look at a case study of a risk profiling model previously used by the Work & Income department of a major Dutch city that discriminated against vulnerable groups and decide to what extent algorithmic recourse could have helped address the hazards inherent to this decision-making process (Chapters 5-7).

Third, relatedly, we develop a novel conceptualization of algorithmic recourse as a control mechanism. We evaluate the case study using the tools of system safety, which provides a framework to theorize about the added value of algorithmic recourse in a system and preemptively analyze the hazards it could bring about (Chapter 8).

Fourth, finally, we design three artifacts to connect theoretical research on algorithmic recourse with the practical requirements of the social welfare domain. We explain how to evaluate algorithmic recourse solutions in real-world contexts (Chapter 9), develop a simulation framework to address these evaluation challenges and apply it on the case study (Chapter 10), and propose a provably-correct algorithm to generate actionable recommendations in expert systems that dominate public administration settings (Chapter 11).

1.3 Structure of this thesis

This document spans 12 chapters. Chapter 1, the current chapter, is the introduction. Next, in Chapter 2, we explain the background of our research, including the challenges of explainability in machine learning. Then, Chapter 3 motivates our approach and introduces the research questions. The following six chapters explain the subsequent steps in our research. In Chapter 4, we contribute a systematized literature review of the field of algorithmic recourse, focusing on the authors' grasp of the problem. Chapter 5 briefly explains the social welfare processes in the Netherlands as a backdrop for the case study, which is the focus of Chapter 6. Next, in Chapter 7, we discuss the interviews that we have conducted with experts in (algorithmic) decision-making to learn about the social requirements for algorithms in public administration. Chapter 8 looks at our case study through the lens of system safety to assess the potential value of algorithmic recourse as a safety mechanism. Further, in Chapter 9, we explain the challenges for the evaluation of algorithmic recourse in real-world contexts, and we provide a solution to address them in Chapter 10. In Chapter 11, we take a brief detour to take on a shortcoming of existing research with respect to expert systems. Finally, Chapter 12 concludes this thesis.

Background | 2

Our research focuses on the provision of *automated and actionable recommendations* to help individuals improve their outcomes when *machine learning models* are employed in a *public administration setting*. We will break down the previous sentence in this chapter, **summarizing the background concepts important for the rest of the document**. We begin by introducing the goals of machine learning technologies in Section 2.1. Then, in Section 2.2, we discuss some approaches that have been proposed to achieve explainability in machine learning. Finally, in Section 2.3, we look at the characteristics of algorithmic decision-making in public administration.

Parts of the discussion in this chapter are adapted from our unpublished paper *Grounding and Validation of Algorithmic Recourse in Real-World Contexts: A Systematized Literature Review* that was drafted as part of this thesis work. The rest of the paper forms Chapter 4

2.1 On machine learning

Machine learning (ML) is a field of artificial intelligence concerned with the development of statistical methods to extract information from observed (training) data and generalize it to unseen (test) data. A flavor of machine learning particularly relevant to our work is supervised learning, where the training data consists of a set of features (measurements; X) and labels (ground-truth outcomes; y). Supervised models learn a mapping $f : X \mapsto y$ from observations; subsequently, f can be applied to infer the labels of unseen instances. Notably, y can take the form of categorical outcomes in classification tasks or continuous outcomes in regression tasks. Then, the performance of a model can be evaluated by comparing the ground-truth outcomes y with the predicted outcomes $\hat{y}$. Typically, the more training data is available, the better f can become [254]. However, quantity is not the only desideratum: models require sufficiently high-quality training data to learn meaningful patterns.

[254]: Viering and Loog (2023), 'The Shape of Learning Curves: A Review'

If a model is relatively simple, the mapping that it has learned from the data can be readily interpreted by its operators. For example, a linear regression model associates a single weight with each feature, yielding a mapping $f_{LR} : \hat{y} = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$, where $x_1, x_2, \ldots, x_n \in X$. Thus, a change of Δx_i in feature x_i influences the outcome by exactly $\beta_i \cdot \Delta x_i$ units. Nonetheless, as the complexity of a model increases, it becomes more and more difficult to interpret its decision-making logic. This may happen because:

- the number of features increases;
- the number of parameters per feature increases;
- the relationships between parameters become non-linear;
- the outcomes hold as statistical truths (model is probabilistic).

In particular, methods such as neural networks (which may rely on billions of parameters) or tree ensembles (which may aggregate outcomes from thousands of individual models) are at the lowest end of the interpretability spectrum. Yet, the complexity of models used in the industry keeps growing by as much as four to five times per year, measured in computing resources [69]. While deep learning models may be characterized by their (perceived) high performance, they are likely to put the end-users – the individuals subjected to algorithmic decisions and the decision-makers operating on these decisions – in a precarious position where they are unable to understand the grounds of a prediction, act on it, or trust it [272].

[69]: Epoch AI (2023), *Key Trends and Figures in Machine Learning*

[272]: Weld and Bansal (2019), 'The Challenge of Crafting Intelligible Intelligence'

2.2 On explainability in machine learning

"Interpretability" and *"explainability"* are often used interchangeably in machine learning research, but we follow the distinction of [152] who carried out a review of literature related to the concepts and proposed that interpretability is a property of a particular model (e.g., generalized additive models [104] or self-explaining neural networks [17]). In contrast, explainability refers to post-hoc solutions that may improve the legibility of decision logic across a variety of models (e.g., Shapley additive explanations [145] or counterfactual explanations [261]). On the one hand, this means that explainability techniques can be applied to models that would otherwise be completely opaque. On the other hand, post-hoc methods attract well-founded objections: they are *"likely to perpetuate bad practice and can potentially cause great harm to society"* [204].

[152]: Marcinkevičs and Vogt (2023), 'Interpretable and explainable machine learning: a methods-centric overview with concrete examples'

[204]: Rudin (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'

2.2.1 Counterfactual explanations

Counterfactual explanations (CEs) are of particular interest to our work. They attempt to explain why the model made a *specific* prediction for a *specific* instance of data without offering "global" insights about the decision-making logic of the model. In natural language, CEs can be represented in the form of conditional statements such as *"if the value of feature x_i was $\boldsymbol{a}$ instead of $\boldsymbol{b}$, the model would have predicted class y_i instead of y_j"*. Counterfactual explanations have been introduced by [261] who claimed that they can improve the understanding of model decisions *"without opening the black box"*: access to the model prediction function generally suffices to compute (multiple possible) explanations. The same authors argued that CEs are a psychologically grounded way to (1) help decision subjects understand an algorithmic decision, (2) provide them with the information needed to contest it, and (3) inform them about actions that could be taken to overturn it. Indeed, humans tend to reason about events through counterfactual statements [156].

[261]: Wachter, Mittelstadt, and Russell (2017), 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR'

[156]: Miller (2019), 'Explanation in artificial intelligence: Insights from the social sciences'

Since the publication of [261] in 2017, dozens of methods to generate counterfactual explanations have been proposed in the literature. These attempt to accommodate a variety of desiderata such as "proximity" (minimal distance between a factual instance and its corresponding counterfactual), "diversity" (reasonable coverage of one factual instance with multiple different counterfactuals), or "robustness" (counterfactual instances that are guaranteed to work even if the data is not perfectly stable). We point the interested readers to [93], [230], or [251] for reviews of CE generation methods.

[93]: Guidotti (2022), 'Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking'

[230]: Stepin, Alonso, Catala, and Pereira-Fariña (2021), 'A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence'

[251]: Verma, Boonsanong, Hoang, Hines, Dickerson, and Shah (2022), 'Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review'

2.2.2 Algorithmic recourse

We briefly discuss CEs in the previous section because they are seen as the go-to method for algorithmic – or actionable, individual – recourse, which was introduced in [244] as *"the ability of a person to change the decision of the model through actionable input variables"*. Thus, recourse is distinct from the "explanation" or "justification" of algorithmic decisions, and more closely related to the notion of contestability of AI systems [9] in that it aims not only to improve the trust in the algorithm but also increase **human agency** [250].

[244]: Ustun, Spangher, and Liu (2019), 'Actionable Recourse in Linear Classification'

[9]: Alfrink, Keller, Kortuem, and Doorn (2022), 'Contestable AI by design: Towards a framework'

[250]: Venkatasubramanian and Alfano (2020), 'The Philosophical Basis of Algorithmic Recourse'

Consider someone whose application for a bank loan was rejected; they could be provided with an AR recommendation of the form *"if you had requested $5000 less, you would have qualified for this loan"*. The key consideration for AR is "actionability", which entails that the recipient of a recommendation should be able to implement it. If the person had been informed *"if you had been 10 years younger, you would have qualified for the loan"*, they would still receive a valid CE, *but not* AR. More recently, [122] has recast the problem as reasoning about minimal interventions on the structural causal model. This formulation (at least theoretically) addresses an important shortcoming of "correlational" recourse: without accounting for the downstream causal effects of actions, an individual may exert more effort than necessary and still fail to achieve the target outcome. Indeed, counterfactuals are an inherently causal concept [181].

Agency may be defined as the capacity to take intentional actions [212], which underwrites *autonomy*, or the ability to self-govern [37]. It remains undecided whether AR can best promote agency by providing people with as many options as possible or with *meaningful* options. If the latter, it is unclear what makes a *meaningful* recommendation although we pose that it relates to its "quality" and/or "actionability".

[122]: Karimi, Schölkopf, and Valera (2021), 'Algorithmic Recourse: From Counterfactual Explanations to Interventions'

[181]: Pearl (2009), *Causality*

We note that problems similar to AR have been studied under a variety of different names: *actionable knowledge discovery* [e.g., 3], *action rules mining* [e.g., 189], *inverse classification* [e.g., 6], *why not questions* [e.g., 105], or *actionable feature tweaking* [237]. These alternative formulations have generally focused on "business" knowledge rather than individual recommendations, but ultimately, the goal of all these approaches is to extract information from a model that allows the user – an individual or a decision-maker – to act.

We highlight these formulations to emphasize that recourse does not have to be achieved through CEs. Instead, they should be seen as *one of the means* to achieve recourse, particularly promising in that they do not require expert-level understanding of the model to be useful. Because of this, **we look at algorithmic recourse as a *task* rather than a set of *methods* in this work**. We believe that the trend of narrowing down the definition of recourse to smaller and smaller subsets of techniques – culminating in [122] equating recourse with (only) causal counterfactuals – distracts from the spirit of the problem: ensuring that humans negatively affected by algorithmic decisions have the tools to react to these decisions. In other words, in our interpretation, algorithmic recourse does not even require a black-box model; settings that rely on "transparent" models may still benefit from algorithmic recourse solutions.

Existing research has generally considered AR in simplistic settings that are far removed from real-world socio-technical decision-making systems, where it would be implemented as a process. For example, such systems are dynamic [192, 243], must support the implementation of AR at scale [14, 169], and involve various stakeholders beyond the end-users [32, 261]. Moreover, if the intended goal of AR is to help individuals subjected to algorithmic decisions in an effective manner, research must entail a rich understanding of "actionability" to account for the differences between them [250]. This inherent complexity of algorithmic recourse suggests that its operationalization in real-world contexts requires multi-dimensional analyses tailored to specific settings, motivating our work.

2.3 On (algorithms in) public administration

In this thesis, we look at algorithmic recourse through the lens of social security, a process of public administration. While there is no agreement among scholars on what exactly constitutes public administration, its primary goal is recognized as the realization of a government's goals through the implementation of its policies [103].

The concept has originated already in antiquity, but it took until the early 20$^{\text{th}}$ century for a rigorous study of public administration to emerge through the writings of Max Weber on bureaucracy [43]. Weber identified three components of a successful administration in the public sector: (1) the clear division of duties among offices, (2) the establishment of regulated chains of command, and (3) the employment of qualified officials [270]. He postulated that a well-organized (bureaucratic) administration *"produces an optimal efficiency for precision, speed, clarity, command of case knowledge, continuity, confidentiality, uniformity, and tight subordination"* [270].

Many of these advantages are also ascribed to algorithmic decision-making systems, so it seems unsurprising that bureaucracies turn to algorithms in their pursuit of efficiency. While the application of ADM tools by public administration organizations is hardly a new phenomenon, their burden on society keeps growing along with their pervasiveness [139]. Indeed, the Netherlands has been "algorithmicizing" its national and local governments for decades, but only recent advances in AI have raised major concerns about the impacts of ADM tools on fundamental rights [191].

[103]: Hasan (2018), 'Governance and Public Administration'

[43]: Chapman, Page, and Mosher (2024), *Public administration*

[270]: Waters and Waters (2015), *Weber's rationalism and modern society: New translations on politics, bureaucracy, and social stratification*

[139]: Levy, Chasalow, and Riley (2021), 'Algorithms and decision-making in the public sector'

[191]: Rathenau Instituut (2021), *Governing algorithmic decision-making in government. The role of the Senate.*

Many challenges for algorithmic decision-making systems have been described in the literature, including the questions of safety (unexpected dynamics in operational environments), security (risks of attacks and abuse), privacy (retrieval of information stored in the training data or model parameters), fairness (equal treatment of individuals), or explainability [40]. As algorithms in public administration contexts are likely to operate on sensitive data (e.g., demographics) to decide on consequential problems (e.g., social security) at a large scale (e.g., nationwide), they ought to be kept to **exceptionally high standards** with respect to these values.

Regulatory frameworks such as the GDPR or the AI Act pay special attention to algorithmic tools that produce legal effects for citizens, but it is worth noting that the "complexity" of a model influences the required safeguards. To illustrate this point we can look at rule-based systems used to decide on the eligibility of citizens for social assistance – while they trigger the protection of the GDPR, they are unaffected by the provisions of the AI Act [68]. However, all forms of data-driven modeling solutions may discriminate against marginalized groups [249]. For example, digital surveillance methods have the strongest influence on vulnerable populations (*cf.* automated profiling) [72, 73]. As such, ADM tools aggravate the power asymmetries inherent to public administration: they create distance between *"those who shape a system"* and *"those affected by a system"*, with few accountability mechanisms allowing the latter to achieve meaningful control over the outcomes [147]. We view algorithmic recourse as a task that may help address this problem.

[40]: Castelluccia and Le Métayer (2019), *Understanding algorithmic decision-making: Opportunities and challenges*

We can argue about these **standards** with the concept of *entitlement* [273], e.g., residents of a country contribute to the social security system with their taxes, and so they have the "positive right" to social welfare assistance when in need but also the "negative right" not to have their privacy infringed.

[68]: Enqvist (2024), 'Rule-based versus AI-driven benefits allocation: GDPR and AIA legal implications and challenges for automation in public social security administration'

[249]: Veale and Binns (2017), 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data'

[72]: Eubanks (2012), *Digital dead end: Fighting for social justice in the information age*

[73]: Eubanks (2014), 'Want to Predict the Future of Surveillance? Ask Poor Communities'

[147]: Maas (2023), 'Machine learning and power relations'

Approach 3

In this chapter, we explain the goals of our research and motivate the approach selected to address these goals. First, in Section 3.1, we outline the objective of our research, and in Section 3.2, we discuss the questions that allow us to address this objective. Lastly, in Section 3.3, we explain the Design Science Research paradigm for practice-oriented research that was applied in this work.

3.1 Research objective

Computer science authors interested in algorithmic recourse tend to operate in the solution space, without **barely having explored and understood** the problem space. Namely, dozens of techniques to generate algorithmic recourse have already been proposed in the literature; however – **to the best of our knowledge** – the question *"is it useful to be able to generate algorithmic recourse?"* has not been answered yet. While this question may seem provocative, being able to decide on the value of algorithmic recourse, the requirements that it can fulfill, or the challenges that it cannot address on its own is needed to progress the field and allow it to leave CS labs.

We discuss the extent of the **existing understanding** of problem space in Chapter 4; especially Section 4.2.7 highlights some ideas that are more strongly grounded in reality.

Several interesting publications **emphasize the ethical value of AR** [e.g., 23, 136, 250]. Even if recourse is "morally good", it is still not necessarily useful.

Thus, our central research objective is **to determine what types of interventions are required to connect the technical affordances of algorithmic recourse with the social constraints and needs of the domains where it could be applied**. We will achieve this **by appraising AR mechanisms from the point of view of a realistic socio-technical system**. Through that, we aim to advance the understanding of the **problem space** of algorithmic recourse to facilitate further developments in its **solution space**.

The **problem space** can be characterized by concepts such as the *needs* (task at hand), *goals* (ways to address the task), *requirements* (constraints of the environment), and *stakeholders* (agents that inform the other parts) [149].
The **solution space** is simply the set of artifacts that could solve the problem.

3.2 Research questions

We define eight questions to help us address the research objective. We introduce and explain them in this section, describe research carried out to explore them in Chapters 4 through 8, develop three (proof-of-concept) artifacts in Chapters 9 through 11, and finally answer the questions in Chapter 12 (Section 12.1). For every question, we state the research methods that will be applied to approach it; Section 3.3 explains how these methods relate to each other.

The first two questions aim to help us understand the current state of the research on algorithmic recourse. RQ 1 and RQ 2 look at the four aspects of the problem space – the needs, the goals, the requirements, and the stakeholders – as envisioned by researchers.

(**RQ 1**) How are the goals and tasks of algorithmic recourse defined and understood by researchers in the field?

Objective: learn about the authors' comprehension of the problem that their solutions (e.g., algorithms or analyses) aim to tackle.

Focus: primarily definitions, but also certain operational aspects such as stakeholders that receive and implement recommendations.

Methods: literature review (**inductive coding**) in Chapter 4.

Codes are derived from the data.

(**RQ 2**) What types of practical considerations are recognized and neglected in the literature on algorithmic recourse?

Objective: identify the connections between theoretical research and practical applications of recourse discussed in the literature.

Focus: primarily operational aspects.

Methods: literature review (inductive coding) in Chapter 4.

The next two questions focus on the actual problem space. Of course, we look at only one potential application of algorithmic recourse, but RQ 3 and RQ 4 should still enable us to contrast the research practices with (some) requirements of the prospective domains.

(**RQ 3**) How can algorithmic recourse complement the existing safety mechanisms for ADM tools in public administration?

Objective: learn about the requirements for algorithms in public administration settings and the ways to improve their reliability.

Focus: current (best) practices in real-world systems.

Methods: semi-structured interviews with experts in Chapter 7, supported by the case study (desk research) in Chapter 6.

(**RQ 4**) What needs of public administration could be addressed by algorithmic recourse but have not been explored yet?

Objective: explore if there exist any potential scenarios (beyond the improvement of outcomes) where algorithmic recourse solutions could support algorithmic decision-making in public administration settings that have not been identified in the literature.

Focus: characteristics, opportunities, threats for ADM in this setting.

Methods: semi-structured interviews with experts in Chapter 7 triangulated against the literature review in Chapter 4.

RQ 5, RQ 6, and RQ 7 aim to reason about the implementation of an AR mechanism in a real-world context. To support responsible uptake, they focus on the ways to evaluate its potential opportunities and threats *before* launching pilot experiments with humans.

(**RQ 5**) How can the authorities explore the potential value of algorithmic recourse before implementing it in a system?

Objective: present an approach to decide if algorithmic recourse could, in theory, contribute to the safe operation of a system.

Focus: system safety perspective on algorithmic recourse.

Methods: example System-Theoretic Process Analysis in Chapter 8 (architecture analysis) based on the case study from Chapter 6.

(**RQ 6**) What are the ways to evaluate the quality of algorithmic recourse recommendations in practical settings?

Objective: propose a theoretical framework to evaluate algorithmic recourse mechanisms attending to its **abstract emergent properties**.

These include system-specific values, e.g., **"actionability" or "robustness"**.

Focus: development of an analysis model.

Methods: artifact of Design Science Research process in Chapter 9 based on the joint analysis from Chapters 4, 7, and 8.

(**RQ 7**) To what extent do "digital twin" solutions allow for the reliable exploration of potential dynamics of algorithmic recourse before implementing it in a system?

Objective: develop a simulation model of a real-world system and evaluate its ability to describe the dynamics of algorithmic recourse.

Focus: long-term multi-agent dynamics of algorithmic recourse.

Methods: computer simulations and computational experiments in Chapter 10 based on the theoretical analyses in Chapters 8 and 9.

Finally, as a way to "close the (design) loop", RQ 8 aims to explore the ways to adapt our findings to other domains and, through that, strengthen the connections between the problem space(s) and the solution space(s) of algorithmic recourse in other applications.

(**RQ 8**) What are the ways to align research on algorithmic recourse with the requirements of realistic domains?

Objective: compare existing research practices to the requirements for algorithmic recourse in real-world contexts identified on the example of a complex socio-technical system from the case study.

Focus: unexplored and underexplored directions of research.

Methods: artifact of Design Science Research process in Chapter 12 based on the joint analysis of all research chapters in this thesis.

3.3 Design Science Research

Design Science Research (DSR) is a research methodology that originated in the field of information systems, combining the objectives of behavioral science (i.e., the analysis of social and organizational factors in a system) and design science (i.e., the development of engineering artifacts to support that system) [110]. In effect, DSR is particularly suitable to develop solutions for practical challenges, especially ones that take the form of *wicked problems* [109] – problems characterized by the lack of clear definition, ill-defined and unstable requirements, a large number of stakeholders with conflicting interests, no ideal solution, and so on [201]. Given the aforesaid complexity of introducing algorithmic recourse into real-world systems, our objective clearly yields itself to this approach.

[110]: Hevner, March, Park, and Ram (2004), 'Design Science in Information Systems Research'

[109]: Hevner and Chatterjee (2010), 'Design Science Research in Information Systems'

[201]: Rittel and Webber (1973), 'Dilemmas in a general theory of planning'

All complete applications of Design Science Research consist of three interlinked cycles [108]. First, the "relevance cycle" gathers problems and requirements from the environment. Second, the "rigor cycle" contributes theoretical and practical grounding from existing knowledge bases. Third, the "design cycle" develops artifacts to address the challenge from the environment based on the information from knowledge bases. Finally, the design artifacts are evaluated in the environment through the "relevance cycle" and contribute further knowledge through the "rigor cycle" [108]. Thus, DSR is, by nature, iterative as it resembles a (guided) search process: the form of final design artifacts cannot be anticipated upfront. These artifacts can take various forms – models, constructs, or methods – and should be provided along with a proof-of-concept instantiation [271].

[108]: Hevner (2007), 'A Three Cycle View of Design Science Research'

[271]: Weigand, Johannesson, and Andersson (2021), 'An artifact ontology for design science research'

Seven guidelines for effective Design Science Research have been formulated by [110]. Most importantly, the DSR process should **(1)** contribute a design artifact **(2)** developed to solve a specific problem. This artifact should be **(3)** purposeful (i.e., its evaluation should prove its utility), **(4)** innovative (i.e., address a novel problem or an old problem in a better way), and **(5)** well-defined, rigorous, consistent. It should also **(6)** follow from a clear search process. Finally, **(7)** DSR and its artifacts must be communicated effectively.

We make our best effort to respect these guidelines. The structure of this document roughly follows our complete search process; we further map all steps in our research (and the corresponding chapters) against the structure of DSR inquiries in **Figure** 3.1.

All figures in this thesis employ a color palette accessible for people with color blindness of [279] and have been checked for sufficiently high contrast.

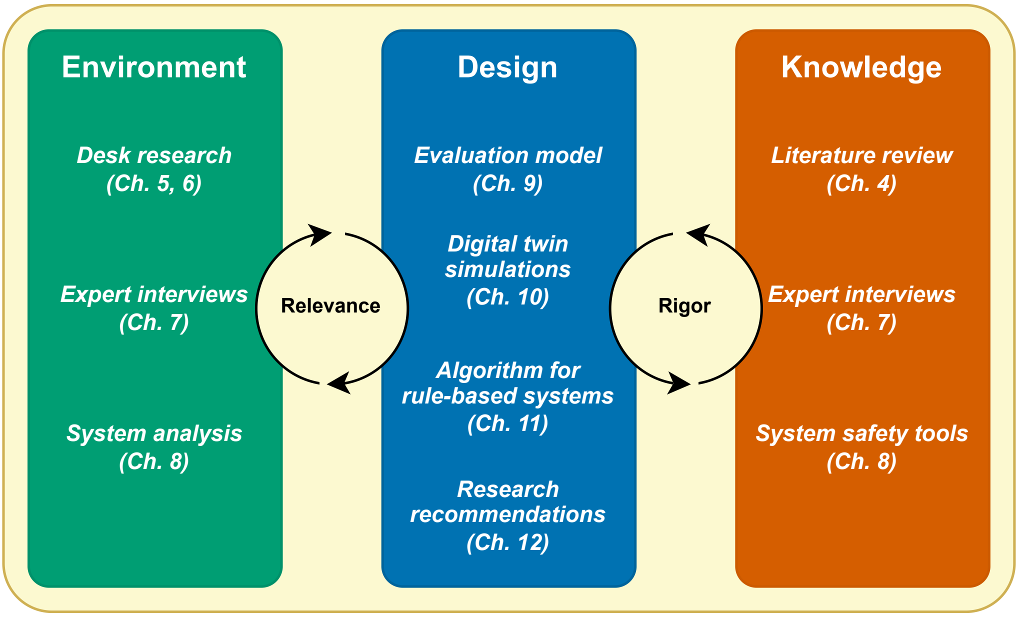
Environment
Desk research (Ch. 5, 6)
Expert interviews (Ch. 7)
System analysis (Ch. 8)
Relevance
Design
Evaluation model (Ch. 9)
Digital twin simulations (Ch. 10)
Algorithm for rule-based systems (Ch. 11)
Research recommendations (Ch. 12)
Rigor
Knowledge
Literature review (Ch. 4)
Expert interviews (Ch. 7)
System safety tools (Ch. 8)

Figure 3.1: Representation of the Design Science Research process as it applies in our work. The complete application of DSR spans Chapters 4 through 12, i.e., the rest of this manuscript.

4 Systematized literature review

We begin the exploration of algorithmic recourse with an analysis of the existing knowledge. This will allow us to verify if our assumption in Section 3.1 that the problem space of AR is underexplored is correct. We focus on the authors' understanding of the concept, which distinguishes our work from the previous literature reviews of the field that have focused on the solutions. First, [121] attempted to unify the definitions and formulations of algorithmic recourse but otherwise looked at the technical aspects. Second, [251] developed a rubric to compare counterfactual explainers (equated with AR) and identified 21 research challenges. While these remained mostly technical, several of them are relevant to our work, for example, CEs *"as an interactive service to the applicants"* or reinforcing *"the ties between machine learning and regulatory communities"*.

Specifically, we contribute a **systematized review** of 127 publications that address the goals of algorithmic recourse and we evaluate to what extent they incorporate various practical considerations. We discuss our approach in Section 4.1, introduce our results in Section 4.2, derive five recommendations on how to improve the practicality of research on algorithmic recourse in Section 4.3, and finally address the limitations of our work in Section 4.4.

We characterize the form of this review as **systematized** because we follow a fully systematic approach to the collection of records, but their selection is not necessarily exhaustive [90]. Many impactful ideas in computer science are published only in the form of preprints, but these were not collected for the purposes of this review.

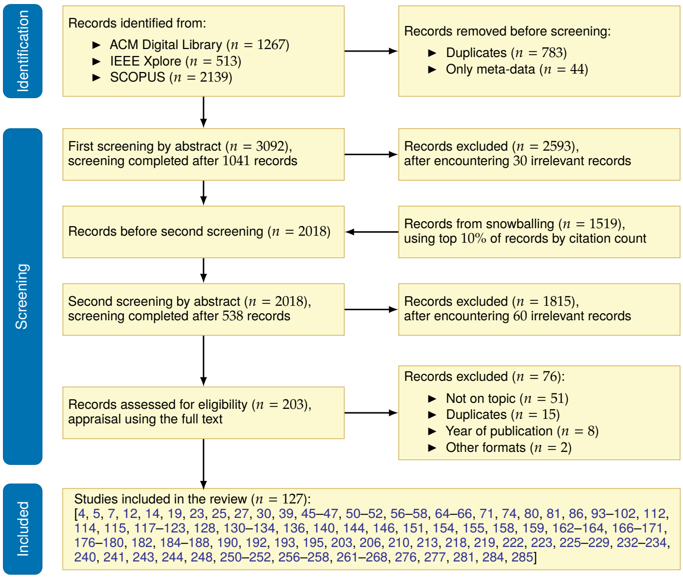
Identification
Screening
Included
Records identified from:
► ACM Digital Library ($n = 1267$)
► IEEE Xplore ($n = 513$)
► SCOPUS ($n = 2139$)
Records removed before screening:
► Duplicates ($n = 783$)
► Only meta-data ($n = 44$)
First screening by abstract ($n = 3092$), screening completed after 1041 records
Records excluded ($n = 2593$), after encountering 30 irrelevant records
Records before second screening ($n = 2018$)
Records from snowballing ($n = 1519$), using top 10% of records by citation count
Second screening by abstract ($n = 2018$), screening completed after 538 records
Records excluded ($n = 1815$), after encountering 60 irrelevant records
Records assessed for eligibility ($n = 203$), appraisal using the full text
Records excluded ($n = 76$):
► Not on topic ($n = 51$)
► Duplicates ($n = 15$)
► Year of publication ($n = 8$)
► Other formats ($n = 2$)
Studies included in the review ($n = 127$):
[4, 5, 7, 12, 14, 19, 23, 25, 27, 30, 39, 45–47, 50–52, 56–58, 64–66, 71, 74, 80, 81, 86, 93–102, 112, 114, 115, 117–123, 128, 130–134, 136, 140, 144, 146, 151, 154, 155, 158, 159, 162–164, 166–171, 176–180, 182, 184–188, 190, 192, 193, 195, 203, 206, 210, 213, 218, 219, 222, 223, 225–229, 232–234, 240, 241, 243, 244, 248, 250–252, 256–258, 261–268, 276, 277, 281, 284, 285]

Figure 4.1: Identification of studies via databases and snowballing

4.1 Setup and protocol

To carry out the literature review, we follow the SALSA – Search, Appraisal, Synthesis, Analysis – framework introduced in [90]. The first three steps are covered in this section, the analysis follows jointly in sections 4.2 and 4.3. Figure 4.1 above presents our process in the form of a PRISMA flow diagram [175].

[90]: Grant and Booth (2009), 'A typology of reviews: an analysis of 14 review types and associated methodologies'

4.1.1 Search

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

We adapt this initial **query** to account for the semantic differences between the search engines, see Appendix A.

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

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

To facilitate the screening process, we employ the open-source ASReview tool, which makes use of an active learning approach to re-order the set of collected publications, such that the ones deemed most relevant are always "at the top of the stack" [247]. We run ASReview on the **default settings** . The researchers behind the tool suggest employing a stopping rule measured in the number of consecutive irrelevant records, which we set to 30, or 1% of the entire dataset. We accept all papers that focus on algorithmic recourse and counterfactual explanations, completing the screening after evaluating 1040 abstracts (33.67% of the dataset), leading to 504 (16.30%) records, among which we identify further 5 duplicates to remove. This results in the reported number of 499 relevant records.

The **settings** entail feature extraction using TF-IDF, Naive Bayes classifier, mitigation of class imbalance with double dynamic resampling, and maximum (certainty-based) query strategy.

We observe that **some important publications may be missing** from our results. For instance, [261] was published in the Harvard Journal of Law & Technology that is not indexed by computer science search engines. Thus, we decide to augment the set of records by applying snowballing, which has been shown as a good alternative to databases in systematic reviews in software engineering [278].

[261]: Wachter, Mittelstadt, and Russell (2017), 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR'

[278]: Wohlin (2014), 'Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering'

We decide to make use of citation counts as a proxy for impact. Due to the lack of a suitable tool that would provide unbiased citation counts for *all* papers in our dataset, we collect them from Google Scholar. Unfortunately, its citation counts tend to be inflated, but we make use of snowballing purely to enrich the dataset, so this does not impact the validity of our study. We manually collect the citation counts for all 499 results from the first screening on January 27th and 28th, order them descendingly, and collect references for the top 50 (10%) "most impactful" papers, yielding 1519 new records.

As expected, we observe that **[261] (mentioned above)** is referenced by 39 of the 50 publications used for snowballing, highlighting its importance.

While this strategy introduces several pre-prints into our result set [specifically: 97, 114, 133, 166, 192, 251], we decide not to exclude them. Our review remains primarily concerned with peer-reviewed work. Here, we also note that [195], which we collected as a pre-print, has been published between the search and appraisal. As such we decided to evaluate its published version and refer to it in the text.

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

4.1.2 Appraisal

We were able to retrieve all of the remaining 203 documents. For each document, we require that the authors explicitly cite recourse as the center of interest, or **look at** **(1)** explanations **(2)** provided for individual instances **(3)** with the goal of acting upon them **(4)** in an attempt to modify the predictions **(5)** of a classification model. We exclude 51 publications that are not on topic, primarily because they focus on CEs for the sake of explanation. Five works in this category look at (what they call) recourse but extend the problem to settings beyond the scope of this review: recommender systems [59, 84, 253], text classification [67], and anomaly detection [54]. Further 15 publications are duplicates, typically pre-prints of other documents that were included in the review. Next, 8 documents were published before [261] that sparked the research on AR, and thus we exclude them as well. These look at the alternative formulations discussed earlier in Section 2.2.2. Finally, two documents are not publications: one is an abstract of a talk, and the other is a student poster. For each document, we answer a number of questions relating to the practical considerations introduced by the authors.

We apply this **operational definition** based on the scoping review as we expect some publications to discuss the goals of AR without explicitly naming this field of research.

4.1.3 Synthesis

To compile the results, we carry out a thematic content analysis following the approach presented in [78]. First, we explore the data extracted from the set of publications relevant to each question to find the commonalities, which allows us to create the initial set of codes. We evaluate the documents against these codes and keep track of any other considerations. If these appear in multiple documents, we create new codes for them. Afterward, we re-evaluate all documents against the new code. As the coding exercise is carried out by one author, they do a third pass over all documents to double-check for potential errors. Finally, where relevant, we cluster the codes into larger themes. In the analysis, we only look at the explicit statements provided by the authors; we do not attempt to infer their understanding of the problem. Thus, the numbers provided in Section 4.2 should be understood as describing how algorithmic recourse is *discussed* in the literature.

[78]: Friese, Soratto, and Pires (2018), 'Carrying out a computer-aided thematic content analysis with ATLAS.ti'

4.2 Thematic content analysis

The following eleven sections introduce the results of the thematic analysis. For brevity, we focus our discussions on the main themes, but we still highlight specific publications if we observe that the authors introduce novel, highly relevant considerations that do not fit into other themes. We begin with more general points such as definitions in Sections 4.2.1 to 4.2.3. Then, in Sections 4.2.4 to 4.2.7 we investigate the social components of AR research. Finally, in Sections 4.2.8 to 4.2.11, we look at the aspects relevant to practitioners.

4.2.1 What forms of contributions do authors choose to make to the research on algorithmic recourse?

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

Table 4.1: Forms of contributions.

Code	Records
Propose methods	69.3%
Develop theoretical frameworks	15.7%
Theoretical or empirical analyses	11.8%
Applications	11.8%
Benchmarks	3.9%
Reviews	2.4%
Total publications in analysis	127

4.2.2 What are the criteria covered in the authors' definitions of algorithmic recourse?

We also evaluate what is understood as the problem to be addressed by AR mechanisms. In particular, what are the criteria to satisfy authors' definitions of recourse. A similar question was posed by [121] who combined six definitions into *"recourse can be achieved by an affected individual if they can understand and accordingly act to alleviate an unfavorable situation, thus exercising temporally-extended agency"*, but this approach was far from systematic.

[121]: Karimi, Barthe, Schölkopf, and Valera (2022), 'A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations'

Instead, we are interested in the underlying concepts. While 74 (58.3%) publications explicitly define AR, 16 (12.6%) mention it but do not include a definition, and 37 (29.1%) do not mention AR, even though they align with its goals. The most common theme is *overturning undesirable decisions* present in 47 definitions (63.5% of all definitions), but specifically *overturning algorithmic decisions* is mentioned only 43 (58.1%) times. It is generally understood that *AR is provided to affected individuals* (44, or 59.5%) but 4 (5.4%) definitions *consider stakeholders* more broadly. *Actionability* as a requirement for recourse is noted in 39 (52.7%) definitions. Then, 20 (27.0%) publications specifically mention CEs as means to AR, while 26 (35.1%) include other technical considerations in the definitions, such as *"changes to actionable input variables"* or *"desired classes"*.

Table 4.2: Definitions of AR.

Code	Records
Overturning undesirable decisions	63.5%
Overturning algorithmic decisions	58.1%
AR provided to affected individuals	59.5%
AR provided to other stakeholders	5.4%
AR requires actionability	52.7%
Technical considerations	35.1%
AR relies on counterfactual explanations	27.0%
AR requires explanation, justification, or understanding	24.3%
Future-orientation or other temporal aspects	13.5%
"Consequential settings"	12.2%
AR as an ability	10.8%
AR requires accounting for the preferences of recipients	2.7%
Total publications in analysis	74

We also point to several themes that are, interestingly, underrepresented. Only 18 (24.3%) documents mention *explanation, justification, or understanding of a decision* as the pre-requisite for AR. Next, 10 (13.5%) highlight *future-orientation or other temporal aspects* of the provided recommendations. Although *"consequential settings"*, typically **bank lending**, are given as examples in nine (12.2%) definitions, they are never explicitly mentioned as the scenarios where recourse ought to be provided, which may be akin to the "*enjoyment of recourse*" as defined by [250] where people are aware that there exists a way to reverse undesirable decisions. 8 publications (10.8%) promote *AR as an ability*. Finally, only 2 (2.7%) publications require that recourse accounts for the *preferences* of its recipients.

Financial domain dominates the evaluations as well, with 90 of 116 evaluations on non-synthetic data making use of at least one finance-related dataset, most commonly `German Credit Data` [111] with 51 uses, see Section 4.2.10.

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

As we observe, "actionability" is a concept that underpins AR but we discover that, in general, its understanding is limited. 91 (71.6%) publications attempt to define what it means (for a CE) to be actionable. Most commonly, in 48 (52.7%) out of 91 definitions, it is understood as *acting only on directly-mutable features*, 6 (6.6%) distinguish that *features may be indirectly-mutable* but still not actionable, while 22

(24.2%) also highlight that *feature values may need to be constrained*. Next, 19 (20.9%) definitions rely on a tautology that actionability means *people can take actions*, 11 (12.1%) emphasize that these *actions must be successful or lead to change*, and 3 (3.3%) further require that they are *aligned with people's real-world objectives*. Only 14 (15.4%) definitions put users at the center stage, indicating that actionability *depends on the user or their preferences*, while 2 (2.2%) highlight the *importance of the context* [252, 266], for instance, that the ability to act on a recommendation may change over time. Importantly, ethical considerations are never mentioned as the pre-requisite for actionability, but we find some broader discussions about this [e.g., 250].

Table 4.3: Definitions of actionability.

Code	Records
Acting on directly mutable features	52.7%
Features may be indirectly mutable but not actionable	6.6%
Feature values may need to be constrained	24.2%
Actionability is when people can take actions	20.9%
Actionability depends on a user or their preferences	15.4%
... that are successful or lead to change	12.1%
... that are aligned with their real-world objectives	3.3%
Actionability depends on the context	2.2%
Total publications in analysis	91

4.2.4 What is the role of end users? What other stakeholders are envisioned in the recourse process?

Given that AR is to be implemented in socio-technical systems that include a variety of actors, we are interested in the types of stakeholders acknowledged in the literature. A total of 105 publications provide explicit consideration of this type. In general, end-users subject to algorithmic decisions are envisioned to be the recipients of AR, but this is not always the case: it may also be provided to experts [e.g., 46, 47, 134] or organizations [e.g., 118, 128, 256], which highlights that in some cases AR may be carried out on behalf of the affected individuals. In any case, 47 (44.8%) publications in the subset agree that end users should inform actionability, but it is rarely clear *how* these preferences should be specified. User-friendly (interactive) interfaces are a consideration in only 14 (13.3%) documents. A total of 29 (27.6%) publications envision domain experts as a group that informs the recourse process. They are either expected to inform actionability in the AR system or provide other forms of knowledge, typically in the form of a causal structure. Besides the experts, authors of 35 (33.3%) papers have discussed a variety of stakeholders. Most commonly system owners [e.g., 45, 64, 71, 164], but also auditors [e.g., 244, 268], data scientists [e.g., 56, 154], developers [e.g., 47, 232], practitioners [e.g., 178, 266], regulators [e.g., 56, 219], or even potential attackers [180].

Table 4.4: Stakeholders in AR process.

Code	Records
End-users should inform actionability	44.8%
... through user-friendly (interactive) interfaces	13.3%
Domain experts should inform actionability	27.6%
Other stakeholders mentioned	33.3%
Total publications in analysis	105

4.2.5 What types of real-world considerations motivate or underlie existing research?

With the multitude of challenges that stand ahead of real-world AR, we are interested in the considerations that motivate existing work. The main theme we find is *ensuring proper individual actionability*, which is addressed in 46 (37.4%) of 123 publications relevant to this question. This is typically achieved with the encoding of user preferences as constraints, but other ideas have been proposed, e.g., providing diverse CEs. *Tackling specific desiderata for AR* (beyond actionability) is the second largest area of research with 28 (22.8%)

publications. Various *other technical challenges*, such as integrating background knowledge [e.g., 30, 115, 119, 176] or incorporating feature importance [e.g., 5, 7, 171, 206] are considered in 24 (19.5%) documents. We also find 19 (15.4%) publications that discuss the problem of *communicating recourse to the end users*. 16 (13.0%) focus on the *dynamics of real-world systems*, typically addressing the robustness of algorithmic recourse [e.g., 132, 166, 168, 243], while 14 (11.4%) look at its mechanisms in *multi-agent systems*. This also relates to *performance considerations* emphasized in 15 (12.2%) of documents. *Causality* drives research in 14 (11.4%) cases. We also find several themes that are under-emphasized: only 9 (7.3%) publications are directly *motivated by research in psychology*, while *ethics of AR* are emphasized in only 7 (5.7%) documents.

Table 4.5: Real-world considerations.

Code	Records
Ensuring proper individual actionability	37.4%
Tackling specific desiderata for AR	22.8%
Other technical challenges	19.5%
Communicating recourse to the end users	15.4%
Dynamics of real-world systems	13.0%
Performance considerations	12.2%
Recourse in multi-agent systems	11.4%
Causality	11.4%
Motivation from psychology	7.3%
Motivation from ethics	5.7%
Total publications in analysis	123

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

While the previous section looked at the considerations that drive existing research, in this section we distill the recommendations for *future* research going beyond the improvement of own work, which are provided in 74 documents. *Causality* is highlighted as a challenge in 22 (29.7%) of them, while *other technical considerations* are given in 20 (27.0%) cases. These range from robustness [e.g., 96, 210, 243], support for categorical features [e.g., 66, 267], or distinguishing between valid CEs and adversarial examples [177]. Next, 19 (25.6%) documents highlight the importance of *ensuring proper individual actionability*, which also relates to *communicating recourse to the end users* (9, or 12.2%) and *supporting realistic cost functions* (8, or 10.8%). *Ethics of AR* are highlighted in 11 (14.9%) publications, for example, that AR research may detract from other obligations of model owners [136, 234]. The same number of publications emphasize the need to (1) *ground research in user studies*, and (2) accommodate for the *dynamics of real-world systems*. *Privacy or security* is highlighted in 10 (13.5%) documents, while the *abuse of recourse*, such as strategic behaviors, surfaces in 7 (9.4%) papers. Other challenges include improving *performance* (8, or 10.8%), considering *multi-agent systems* (4, or 5.4%), and developing *legal frameworks* (4, or 5.4%) for recourse. We also highlight several challenges particularly relevant to our work: (the usefulness of) recourse is perceived as difficult to evaluate in practice [80, 112, 203], it must account for individual, contextual, societal, and even cultural factors [222], which further means that engagement with recourse mechanisms and the likelihood of its implementation are context-dependent [e.g., 7, 81, 228].

Table 4.6: Real-world challenges.

Code	Records
Causality	29.7%
Other technical challenges	27.0%
Ensuring proper individual actionability	25.6%
Dynamics of real-world systems	15.4%
Ethics of AR	14.9%
Ground research in user studies	14.9%
Privacy or security	13.5%
Communicating recourse to the end users	12.2%
Supporting realistic cost functions	10.8%
Performance considerations	10.8 %
Mitigating abuse of recourse	9.4%
Recourse in multi-agent systems	5.4%
Developing legal frameworks	5.4%
Total publications in analysis	74

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

Real-world systems entail the implementation of recourse by multiple agents, which may introduce group-level dynamics. However, out of 119 documents relevant to this question, 93 (78.2%) seem to understand algorithmic recourse as a purely individual phenomenon. Among the remaining 26 documents we find considerations for several different group-level effects. Various perspectives on the problem of *fairness*, covering both individual and group formulations are addressed by [19, 66, 97, 218, 219, 232, 258, 264]. Next, [14] shows that the implementation of AR on a large scale may lead to **domain and model shifts**, which introduce unexpected costs for the stakeholders. In [81], the authors focus on another negative consequence of AR at scale, showing that it may reinforce social segregation. The impact of the "right to be forgotten", where data deletion requests trigger model retraining that may invalidate existing recourses is addressed in [132]. Then, [169] develop a game-theoretic framework for AR in multi-agent settings, attempting to optimize for "social welfare" rather than the profits of individual agents. We find two further similar perspectives on recourse: [71] proposes auditing and subsidies to minimize the risks of strategic behaviors in a multi-agent setting, while [241] attempts to incentivize actual improvement for a population of agents. Finally, [118] provides a framework that generates transparent and consistent recommendations for a sub-population. Two other lines of research account for the remaining papers with group-level considerations. First, in a causal setting [e.g., 124, 130] *sub-populations* are necessary to estimate the interventional effects on individuals. Second, some works highlight the importance of *global insights* into the data [47, 80, 86, 140, 187, 193, 262], e.g., recourse summaries [140, 193].

Such **endogenous dynamics** were postulated earlier in the first version of [192] dated December 22nd 2020, but the authors have completely removed this discussion from the subsequent versions of the pre-print.

Table 4.7: Group-level dynamics.

Code	Records
Fairness	8
Global insights	7
Sub-population considerations	5
Dynamics of real-world systems	3
Game-theoretic approaches	3
Total publications in analysis	26

4.2.8 What are the requirements for proposed methods?

While we do not attempt to repeat the excellent analyses of the characteristics of methods as provided in [93], [121], and [251], we emphasize that strict requirements for the application of methods may be an important obstacle to the broader adoption of AR by practitioners. Out of 88 publications that propose methods for AR, 70 (79.5%) *focus on the generation of individual CEs*. Among the other 18 works, we find, for example, [140, 193] that offer methods to generate global insights ("summaries" of recourses), or [227] that attempts to visualize CEs against the decision boundary of a model.

Notably, we find that as many as 25 CE generators (35.2% of all CE generation methods) are either *model agnostic* or can be naturally extended into a model-agnostic setting. This follows from the large popularity of formulations relying on *genetic algorithms* (8, or

[93]: Guidotti (2022), 'Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking'

[121]: Karimi, Barthe, Schölkopf, and Valera (2022), 'A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations'

[251]: Verma, Boonsanong, Hoang, Hines, Dickerson, and Shah (2022), 'Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review'

11.4%), *linear or quadratic programming* (7, or 10.0%), and *satisfiability approaches* (4, or 5.7%). Purely gradient-based formulations remain the single largest category with 13 (18.5%) methods. We also find 4 (5.7%) methods based on case-based reasoning and one (1.4%) formulation using only greedy heuristics [86]. Slightly over half of the generators – 37 (52.9%) – rely on *other or mixed approaches*. The former includes graphs [25, 182], program synthesis [58, 186], Monte Carlo sampling [195], or deep reinforcement learning [46]. The latter includes, e.g., gradients with heuristic refinement [284] or gradients with kernel-density estimation [168]. We find that 16 (22.5%) methods are designed for *specific model types*, most commonly trees or tree ensembles. Next, 13 (18.3%) methods require the practitioner to *train a new model*, for instance, a residual GANs [163], or different types of VAEs [65, 81, 123, 151, 179]. Finally, 10 (14.3%) methods require the specification of the *full SCM or a causal graph*.

Table 4.8: Requirements.

Code	Records
Generation of individual CEs	70
Other goals	18
Fully model-agnostic	25
Multiple model types	29
Fully model-specific	16
Gradient-based formulations	13
Genetic algorithms formulations	8
Linear or quadratic programming formulations	7
Case-based reasoning approaches	4
Satisfiability approaches	4
Heuristic approaches	1
Other and mixed formulations	37
Training a new model	13
Full SCM or a causal graph	11
Total publications in analysis	88

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

We now explore the different forms of "real-world" evaluations, going beyond quantitative experiments, which are present in 51 publications. Most commonly, in 28 (54.9%) of these, the authors make use of *case studies* presenting the methods in an end-to-end manner. Among them, the application of recourse in the `Hired.com` marketplace goes furthest in simulating real-world conditions for AR [164], but the recommendations are still not evaluated with humans in the loop. Further, 9 (17.6%) documents include other forms of *short walk-through examples*. We also identify 14 (27.5%) papers that evaluate the methods with *user experiments*, 10 of which involve non-expert users and 4 involve expert users. While we do not observe any interviews with non-expert users, we find 1 (2.0%) publication where *experts are interviewed* [47]. *Other involvement of non-experts* applies to [206], where they inform the development of methods. *Other involvement of experts* is featured in two documents where they evaluated the outputs of methods [51, 233]. Altogether, end users were involved in 17 publications, which is only 13.3% of all publications covered in our study, even more striking than the 21% of CE methods evaluated with user studies as reported in [126].

Table 4.9: Realistic evaluation.

Code	Records
Case studies	54.9%
Walk-through examples	17.6%
Experiments with non-expert users	19.6%
Experiments with expert users	7.8%
Interviews with experts	2.0%
Other involvement of experts	3.9%
Other involvement of non-experts	2.0%
Total publications in analysis	51

[126]: Keane, Kenny, Delaney, and Smyth (2021), 'If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques'

4.2.10 Which datasets are used in the computational experiments with the proposed methods?

We argue that datasets used in evaluation can help us understand domains where the authors envision their methods would be applied. Unfortunately, we find that the referencing standards for datasets in algorithmic recourse literature are sub-par. For instance, the `Default of Credit Card Clients` dataset introduced by [283] is referred to under six different names. To resolve such differences, we work backward, following the references to a primary source. This poses another problem: we find that authors do not always refer to the primary sources for the datasets used in the evaluation. As one example, [192] and [193] make use of a dataset referred to as `Bail`, both referring to another publication of one of the authors. When cross-referenced with the target publication we find that it is a tertiary source. Relatedly, we find many publications simply referring to dataset repositories rather than individual datasets.

Table 4.10: Non-synthetic datasets.

Code	Records
Statlog (German Credit Data)	77.6%
South German Credit Data	3.4 %
Adult	37.1%
COMPAS Recidivism Risk Scores	20.7 %
Default of Credit Card Clients	16.4%
Home Equity Line of Credit	16.4%
Proprietary datasets	10.3 %
Other datasets	70.7%
Total publications in analysis	116

We find 100 different datasets used in 116 publications that report evaluations with non-synthetic data. Financial domain dominates the landscape, with 90 (77.6%) publications using at least one related dataset. These include `Statlog (German Credit Data)` [111] in 51 (44.0%) papers and further 4 (3.4%) using its improved version `South German Credit` [242], `Adult` [26] in 43 (37.1%) papers, or `Default of Credit Card Clients` [282] and `Home Equity Line of Credit` [75], both referenced 19 (16.4%) times. Another popular choice is the `COMPAS Recidivism Risk Scores` dataset [135] used 24 (20.7%) times. Proprietary datasets were used in 12 (10.3%) papers, **nine of which focus on the application** of AR. In 82 (70.7%) publications the evaluation relies on at least one other dataset.

This means that **60% of papers that apply AR** cannot be reproduced.

4.2.11 What are the open source and documentation practices in the research on algorithmic recourse?

Finally, we note that the lack of availability of well-documented open-source code may be an important obstacle to the application of AR in real-world systems. For all 116 publications that involve some form of computational experiments, we verify whether the source code is publicly available. If the authors do not explicitly link to their code in the paper, we attempt to find it independently. Ultimately, we collect open-source implementations for 64 (55.2%) publications. Then, for each of them, we evaluate the quality of documentation. The *instructions on the general usage* (installation and workflow) are provided with 27 (41.5%) repositories, while *instructions on the reproduction of results* in 23 (35.4%). In 19 (29.2%) cases we find *walk-through tutorials*, typically in the form of Jupyter Notebooks, although we note that they differ in quality. For instance, five repositories include code-only notebooks with no further textual explanation that could guide the practitioner. Implementations for four papers include more “professionalized” *documentation* [14, 159, 178, 266]. The latter sets a golden standard as it further includes a tutorial video and a live demo. We do not find *any* additional materials for practitioners for 13 (20.0%) of the available implementations.

Table 4.11: Documentation practices.

Code	Records
Open-source code available	55.2%
Open-source code not available	44.8%
Instructions on general usage	27
Instructions on reproduction	23
Walk-through tutorials	19
Documentation	4
No materials for practitioners	13
Total publications in analysis	116

4.3 Discussion of the literature review

Here, we summarize the review. First, in Section 4.3.1, we explain why reframing algorithmic recourse as a socio-technical problem is necessary to allow for its adoption in real-world contexts. Second, we provide other researchers with five recommendations to help them address this challenge in Section 4.3.2.

4.3.1 Algorithmic recourse as a socio-technical challenge

Regardless of whether AR can be normatively expected or not [136], many systems can genuinely benefit from such mechanisms, especially when the interests of the system owner and the end users are aligned [128], such as in the public administration contexts (our case study), in the healthcare system to improve the well-being of patients [134, 171, 265], or on the online platforms that aim to enhance the experience of their users [164, 237]. Nonetheless, the values and norms underlying recourse – trust, agency, fairness, safety, and so on – are emergent properties of systems where AR mechanisms would be introduced. Such norms can only be understood and evaluated when accounting for the technical, social, and institutional components of a system [63], but the latter two remain largely unexplored in the recourse literature.

Recourse is not inherently safe or unsafe, *but* its (incorrect) implementation may lead to the emergence of unsafe dynamics, such as the unexpected costs to stakeholders as discussed by [14] or the reinforcement of social segregation addressed in [81]. While it may be too challenging to provide accurate system-level evaluations at this stage of research, authors can still expand the boundaries of their analyses to account for global effects or look at the position of recourse mechanisms in the broader context of a complete socio-technical AI system [62]. As AR is a "reality-centric AI" problem [208] by its nature, working towards its integration into existing systems will require a design-oriented approach, potentially with *specific* systems in mind. The "Abstraction Traps" discussed by [215] in the context of research on fair machine learning apply here: that technical solutions designed for one social context cannot be directly repurposed for another application, that values to which they are expected to adhere to cannot be captured with mathematical formulas, that their insertion into an existing process will impact its behavior, or that the best solutions may not necessarily be technical.

It is perhaps most telling that only 12% of surveyed publications attempt to apply recourse in realistic settings. We will discuss two of these settings to highlight the stark differences in system properties. Most of the applications included in our review focus on the provision of actionable individual recommendations to students [4, 5, 50, 188, 226, 240, 276]. In this relatively low-stakes

[136]: Leben (2023), 'Explainable AI as Evidence of Fair Decisions'

[63]: Dobbe and Wolters (2024), 'Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context'

[62]: Dobbe, Krendl Gilbert, and Mintz (2021), 'Hard choices in artificial intelligence'

[208]: Schaar and Rashbass (2023), *The case for Reality-centric AI*

[215]: Selbst, boyd, Friedler, Venkatasubramanian, and Vertesi (2019), 'Fairness and Abstraction in Sociotechnical Systems'

domain almost any recommendation will be actionable in that following a personalized set of learning activities does not require resources other than time. Even then, the system involves multiple actors – students, teachers, parents – whose interactions will impact the process, e.g., because students may fail to benefit from learning activities without additional support. Conversely, we find several publications where authors attempt to provide recourse in the high-stakes medical domain [134, 171, 265]. Here, recommendations must be tailored to the preferences, resources, or lifestyles of patients to have a chance of being actionable. Moreover, certain aspects of their implementation fully depend on other actors, such as a clinician prescribing the medications. It may also happen that recourse does not exist if the health outcomes of a patient cannot be improved.

4.3.2 Recommendations for future research

We distill our findings from the review into five recommendations. First, in Sections 4.2.2, 4.2.3 we observed that *operational* definitions for recourse are still unavailable. Second, Sections 4.2.4 and 4.2.9 underlined insufficient consideration for people involved in recourse processes. Third, Sections 4.2.5, 4.2.6 highlighted the overwhelmingly technical approaches to recourse. Fourth, Section 4.2.7 stressed the lack of group-level analyses. Fifth, from Sections 4.2.9 to 4.2.11 we learned about the missing consideration of practitioners. Our recommendations are primarily targeted at other researchers, but they can also inspire practitioners to think about algorithmic recourse in their domains. Hence, for each recommendation, we derive examples of two questions to encourage thinking about the meaning of algorithmic recourse in specific domains.

1. Broadening the scope of research.
AR is generally seen as a service for affected individuals, but this formalization may be unnecessarily limiting. In fact, in many systems, these individuals may be unable to directly act on recommendations [see 250]. Instead, we propose to operationalize the aim of AR as the provision of recommendations *aligned with the preferences* of *non-expert users* in an attempt *to help them improve outcomes* in an *ADM setting*, which emphasizes *easy to understand* and *individually actionable* recommendations as the key research problem.

Example questions for practitioners:

- *Who would be responsible for implementing recommendations?*
- *How would the preferences of end-users be collected?*

2. Engaging end-users, affected individuals, and communities. Algorithmic recourse solutions are rarely evaluated with humans. Instead, they attempt to satisfy a variety of desiderata formulated by authors and assessed in an automated manner. Sparsity or proximity are far from perfect proxies for individual actionability. For AR to be useful, it must satisfy the preferences of its end users. Research is also necessary to learn about the needs of the affected individuals, and to validate the potential contributions and inherent limitations of AR. Authors may benefit from the rich literature on human-computer interaction [e.g., 18, 48] or psychology.

Example questions for practitioners:

- *What should be the measures for the quality of an AR mechanism?*
- *How would the recommendations be communicated to end-users?*

3. Accepting a socio-technical perspective. A pervasive assumption in the literature is that all challenges of AR require purely technical solutions. For instance, many authors emphasize the importance of causal modeling to guarantee recourse, but the models that aim to be explained are themselves *not* causal. Similarly, to improve the performance of CE generators many authors turn to deep generative models [65, 81, 114, 123, 151, 163, 179]. Not only do they explain the data rather than the model [16], but more importantly they shift the problem from improving the trust in non-interpretable models, to attempting to trust non-interpretable explainers. Although a socio-technical perspective on AR brings its own challenges, such as accounting for the roles of stakeholders involved in the provision of recourse, it creates important opportunities. For example, developing “recourse contracts” [64, 74] or designing feedback processes to account for imperfect robustness.

Example questions for practitioners:

- *What (types of) procedures are already in place?*
- *What (types of) stakeholders would interact with the AR mechanism?*

4. Accounting for emergent effects. Decision-making systems involve multiple individuals that may have competing interests. From the onset, research on AR should consider group-level effects, such as external costs or fairness, to explore these dynamics. This requires expanding the boundaries of analysis, but it is necessary to ensure the safe operation of the system. Dynamics may also emerge due to multi-system interactions: changes implemented by an individual to improve their outcomes in one system will affect them in other contexts [see also 23].

Example questions for practitioners:

- *What (types of) dynamics could put the system in hazardous states?*
- *What would be the ways to mitigate these unsafe dynamics?*

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

Example questions for practitioners:

- *Which existing AR solutions could be applicable in the system?*
- *What are the ways to tailor them to the system's specific constraints?*

4.4 Limitations of the literature review

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

Again, the impossibility of performing **an exhaustive search** follows from the specificities of CS publishing. Some other pertinent publications pointed out by a reviewer of our NeurIPS submission include [70, 129, 238] on human-in-the-loop preference elicitation and [28, 76] on multi-agent effects.

Social welfare in the Netherlands 5

Having investigated the knowledge base of algorithmic recourse, we move on to explore the environment of our problem in the coming few chapters. We start by introducing the background for our case study, further discussed in Chapter 6. We explain the forms of social assistance available to the residents of the Netherlands in Section 5.1, with a focus on *bijstanduitkering* under *Participatiewet* in Section 5.2. Next, in Section 5.3, we elaborate on other international, national, and local laws that shape the functioning of social assistance. Finally, Section 5.4 motivates the case study by evaluating the interest of Dutch municipalities to employ algorithms in this domain based on the publicly available information from the Algorithm Register.

5.1 Instruments of social assistance

The Ministry of Social Affairs and Employment (*Ministerie van Sociale Zaken en Werkgelegenheid*; *SZW*) is responsible for the national policy on **social security**, including social assistance, in the Netherlands. Depending on the specific type of assistance, various administrative units are involved in the implementation of insurances:

All **links in this chapter** have been accessed on September 29, 2024, and saved in the Internet Archive Wayback Machine to ensure traceability.

Where not explicitly cited, the links refer to the information provided by the central government on the pages of **Rijksoverheid.nl** [199].

- **Employee Insurance Agency** (*Uitvoeringsinstituut Werknemersverzekeringen*; *UWV*) responsible for unemployment benefits under *Werkloosheidswet* (*WW*) and schemes for people with disabilities, including *WAO*, *WIA* and *Wajong*.
- **Social Insurance Bank** (*Sociale Verzekeringsbank*; *SVB*) responsible for long-term insurances such as pension schemes (*AOW*), care for vulnerable elderly people (*Wlz*), and child benefits.
- **Individual municipalities** are responsible for certain other forms of benefits that aim to ensure the residents' income does not fall below the social assistance standard. Most importantly for this thesis, municipalities are responsible for the implementation of the **Participation Act** (*Participatiewet*), but also the benefits for people who become unemployed close to retirement age (*IOW*, *IOAW*, *IOAZ*), and different forms of support under the Social Support Act (*Wmo 2015*) to ensure that people can live in their own homes as long as possible.

Financial aid under Participation Act is known as *bijstand* or *bijstanduitkering*, directly translated as "assistance".

Our case study focuses on a machine learning model deployed to support the municipality of Rotterdam in the implementation of regulations laid out in the Participation Act. Hence, we will focus on this form of social assistance in Section 5.2.

5.2 Legal framework for *bijstand*

The Participation Act entered into force on January 1st 2015, defining a form of assistance for all residents of the Netherlands who are able to work but cannot find employment [174]. People who become unemployed are initially covered under the *WW-uitkering* after registration as a job seeker with the *UVW* [274]. The length of coverage depends on the employment history, ranging from three months to two years [245]. After this period, people are transferred to the care of their municipality under the Participation Act.

According to the data from Statistics Netherlands (*Centraal Bureau voor de Statistiek*; *CBS*), roughly five million people in the Netherlands receive some form of social support, but the number of people insured under *Participatiewet* differs greatly between municipalities. [42]. While in a majority of municipalities the density is roughly 10 to 15 per 1000 inhabitants, in the top 10 most populated municipalities it averages 44 per 1000 inhabitants. In particular, in Rotterdam, 6.1% of residents receive *bijstand*, the highest percentage in the country.

To be entitled to the *bijstanduitkering* individuals must:

(a) be 18 years of age or older;
(b) have a Dutch address;
(c) have a Dutch citizenship or valid residence permit;
(d) have income below the social assistance standard;
(e) not be eligible for another form of social assistance;
(f) not be in prison or a detention center.

Additionally, recipients must remain registered as job seekers and accept any employment that is offered to them, and they may be obligated by the municipality to carry out **volunteer work** [183]. The latter may take various forms of socially useful tasks, e.g., helping in shelters for unhoused people or supervising playgrounds.

Similar **"quid pro quo"** obligations are enforced in IOAW and IOAZ benefits.

The **maximum amount of *bijstand*** is tied to the gross statutory minimum wage, which is adjusted twice every year. For a particular resident, it further depends on their marital (or co-habitation status), as well as the number of adult cost-sharers in the household.

As one example, the minimum wage standard on September 29, 2024 was €2,133.60 and the **maximum amount of *bijstand*** for a single person younger than retirement age was €1,308.45.

As the implementation of *Participatiewet* is the responsibility of the individual municipalities, there exist some differences with regard to the methods employed to ensure the eligibility of recipients. All municipalities must **periodically re-examine** the recipients of *bijstand* to ensure that benefits are duly granted. Nonetheless, they may employ **various tools to nominate residents** for re-examination, such as random, expert-driven, or algorithm-driven selections [35]. This also means that different municipalities may approach the re-examinations with different levels of strictness (see Chapter 7).

There are **certain exceptions** in the Participation Act. For instance, people may be re-examined only once every two years and only after having received assistance for more than six months.

Many municipalities additionally run web forms or phone numbers for anonymous **reporting of welfare fraud**.

5.3 Other relevant laws

Apart from the *Participatiewet*, several other legislative acts are relevant for the implementation of municipal social assistance. In this section, we do not aim to be exhaustive; instead, we highlight several of the **most important acts** to reflect on the complexity of the legislative landscape and inform the further analysis of the system.

This section, and specifically the inclusion of GDPR, Awb, and Wet SUWI **was initially informed by the analysis of the Algorithm Audit**; we refer the reader to Appendix A in [11].

European Convention on Human Rights (ECHR)
Article 8 of ECHR defines the *right to respect for private and family life*, according to which public authorities must not interfere with the private and family life, home, and correspondence of citizens unless it is *"necessary in a democratic society (...) for the prevention of disorder or crime"*. Notably, **Article 8 is understood to impose a requirement for public authorities to ensure a fair balance** between the rights of an affected individual and the community as a whole.

This **interpretation** has previously led The Hague District Court to determine that the fraud detection "System Risk Indication" (SyRI) model was unlawful and its outcomes not binding [194].

General Data Protection Regulation (GDPR)
Several provisions of the GDPR are relevant for this setting. First, with regard to algorithm-driven nominations, Article 4 defines profiling as *"any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person"*. Additionally, Article 5 outlines the principles relating to the processing of personal data, including the requirement to process data *"lawfully, fairly and in a transparent manner"* and collect it for *"specified, explicit and legitimate purposes"*. Next, **Articles 13-15 discuss the right of a data subject to access their personal data**, specifically requiring the data controllers to inform about *"the existence of automated decision-making"* and to provide *"meaningful information about the logic involved"*. Finally, Article 22 bestows the right *"not to be subject to a decision based solely on automated processing, including profiling"*, although its phrasing indicates the article is only applicable in settings where humans are not involved at any step of the decision process (i.e., decisions are fully automated).

Same provisions are codified for different types of systems, and thus **relevant rules** include 13(2)f, 14(2)g, and 15(1)h.

Artificial Intelligence Act (AI Act)
While the AI Act entered into law only on August 1st 2024 and thus its provisions are not directly relevant to our case study, we highlight that according to Article 86 *"any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system"* should be able to receive a *"clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken"*, which specifically spells out the right to explanation whose status in GDPR was debated [87, 214, 260]. Moreover, Annex III outlines different types of high-risk AI systems, specifically including *"AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for essential public assistance benefits and services"*, and thus introducing a variety of other requirements for AI systems that may be deployed in this domain in the future.

Algemene wet bestuursrecht (Awb)

The Dutch General Administrative Law Act (GALA) establishes general rules for the relationships between public authorities and individual Dutch citizens (residents). First, GALA enforces the **duty of care requirement** in Section 3.2, according to which administrative bodies must act to prevent consequences disproportionate to the objectives of decisions. Second, Section 3.7 and specifically Article 3:47 require that the reasoning behind decisions, including the legal basis, is provided along with these decisions. Third, all subordinate legislation must comply with the regulations of GALA.

We additionally point to the discussion of Algorithm Audit who note that factors such as the **selection of features for an ML model could fall under the duty of care requirement** [11].

Wet Structuur Uitvoeringsorganisatie W&I (Wet SUWI)

Roughly translated to the "Work and Income Implementation Agencies Structure Act", the law establishes the structure of public authorities relevant to the implementation of social assistance as discussed in Section 5.1 and the relationships between them. For instance, Article 9 informs the cooperation between *UWV*, *SVB*, and **municipalities**, Article 30c outlines the administrative processing of applications for benefits, and Article 62 discusses the mutual exchange of data between all involved authorities.

More specifically, the College of Mayor and Aldermen, which is the **executive board of municipalities**.

Municipal rules and regulations

Various local regulations are relevant to the problem. As explained in Section 5.2, individual municipalities have a certain degree of freedom in the implementation of the Participation Act. As such, these local regulations differ between municipalities. In the case of Rotterdam, pertinent legal acts include the *Verordening maatregelen en handhaving* on the enforcement of obligations under Participation Act, IOAW, and IOAZ, or the *Nadere regels voorzieningen Participatiewet* on the local implementation of the Participation Act [173].

5.4 Algorithm Register

Finally, we examine the Algorithm Register, which is under development as a database of algorithms employed by public authorities in the Netherlands [157]. Our goal is to understand how Dutch municipalities use algorithmic decision-making in the social welfare domain. We export **the complete list of algorithms** registered at the end of June 2024 and find that 25 out of 411 algorithms are directly relevant to the tasks of social welfare, and further 29 are employed in related domains such as work reintegration or *Wmo* support.

Most notably, we find the re-examination model of Rotterdam which will be the subject of our case study, and a similar model registered by the municipality of Utrecht (both marked as out-of-use) [172]. We also note the "Rights of Rotterdam residents" algorithm used by the municipality to determine citizens' eligibility for *bijstand* [172].

[157]: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2023), *Handreiking Algoritmeregister. Aan de slag met het Algoritmeregister.*

The **registration of algorithms** is not (yet) legally mandated and many descriptions therein are marked by parsimony. Thus, the numbers in this section are likely to be underestimated.

Building on the point above, **Nissewaard also used a machine learning model** to detect social assistance fraud [259]. However, this algorithm cannot be found in the Algorithm Register.

6 Case study: risk profiling in Rotterdam

To *ground and validate* algorithmic recourse in the real world, we decide to look at it through the lens of a case study. We turn our attention to the process of algorithm-driven selection for re-examinations under the Participation Act that has previously been in place in the municipality of Rotterdam. We start this chapter by introducing the case study in Section 6.1. Then, in Section 6.2, we look at the direct and indirect stakeholders in the system.

6.1 Context

To the best of our knowledge, the re-examination model employed by the Work & Income department of Rotterdam was initially disclosed in a case study by the Rotterdam Court of Audit (*Rekenkamer*) as part of the "*Coloured Technology*" report from 2021 that looked at the ethical use of algorithms by the city authorities, following principles such as responsibility, transparency, or fairness [196]. Later, in 2023, this specific model became the centerpiece of an **investigation coordinated by Lighthouse Reports** that focused on the "*suspicion machines*" – fraud detection systems employed in social welfare across Europe [83]. As acknowledged by the journalists *"out of dozens of cities we contacted, [Rotterdam] was the only one willing to share the code behind its algorithm"*. The investigation not only unveiled that the model was characterized by **poor technical performance** but also that it was likely to discriminate against people from vulnerable backgrounds, e.g., migrants or single mothers, recommending them for re-examinations much more often than in expectation [35].

This investigation was a collaborative effort between **Lighthouse Reports**, WIRED, VPRO, The Pulitzer Center, Follow The Money, Vers Beton, Open Rotterdam, and The Center for Artistic Inquiry and Reporting, with additional consulting of researchers from nine research institutes and NGOs.

Although the model was used outside of the lab, it has **never officially left the pilot stage**. Rotterdam decommissioned the model in 2021, two years before it became the subject of the Lighthouse Reports investigation [35].

While the aforementioned case studies focused on the system **as it was**, we instead "re-design" and evaluate the system **as it could have been** if algorithmic recourse had been a risk mitigation strategy employed by the city. Indeed, there are three compelling reasons to study the model employed by the Rotterdam Work & Income department, hereafter called the "Rotterdam case".

First, the case is a perfectly cromulent setting for the application of algorithmic recourse. The goals of the municipal authorities and the affected benefits recipients are generally aligned: people who do not have enough income should be supported with *bijstand*, and they should not be unnecessarily subjected to **invasive re-examinations** provided that they continue to fulfill the eligibility requirements. Moreover, if a machine learning model was trained on features relevant to the pursued task, we would expect the municipality to appreciate the possibility of supplying its citizens with recommendations that would help them lower the associated risk scores, especially if the outcome of a re-examination does not

These **re-examinations** may even include *"home visits where fraud investigators reportedly sift through laundry and count toothbrushes"* [36].

change the status of a welfare recipient. As previously explained in Section 5.2, Rotterdam has the highest density of residents who rely on *bijstand* in all of the Netherlands. In 2021, the Work & Income department of Rotterdam supported **≈35,000 individual welfare recepients** and engaged in ≈6,000 investigations each year [196]. Thus, algorithmic recourse mechanisms could relieve the municipality of unnecessary expenditures in the long term.

Based on the most recent data from CBS, we would expect this number to have grown to **over 40,000** in 2023 [42].

Second, as we identified in Chapter 4, previous case studies on the application of algorithmic recourse tend to either be relatively simplistic or rely on proprietary data and models unavailable to the general public. Conversely, the Rotterdam case involves a reasonably complex decision-making system where we can evaluate the technical, social, and organizational processes owing to the vast trove of documents unveiled in earlier investigations. In particular, Lighthouse Reports published *"the holy trinity of algorithmic accountability: the training data, the model file and the code for [the] system"* in their GitHub repository [35]. We can also draw from the analyses of the Rotterdam Court of Audit that discussed the organizational processes behind the model [196] and Algorithm Audit that looked at the quality of **features used by the model** [11].

We will introduce this **model** in more depth in Chapter 10. For now, it suffices to say that Rotterdam used a gradient-boosting machine trained on 315 features of social welfare recipients.

Finally, we highlight the societal significance of this case study. While in the context of engaging with corporations (e.g., banks that are frequently mentioned in research on algorithmic recourse) individuals generally have the choice to withdraw themselves from a particular decision-making process (e.g., move their account to a different bank), this is not the case with governmental processes. As such, the implementation of machine learning models for any purposes of public administration requires particularly strong safety mechanisms. Moreover, people who rely on social welfare are *by definition* in a vulnerable position. Thus, a re-examination is likely to have destructive impacts on their livelihood. As one example, WIRED.com cites the following case of a resident of Rotterdam who was **subjected to re-examination twice**: *"She was questioned and lost her benefits for a month. 'I could only pay rent,' she says. She recalls the stress of borrowing food from neighbors and asking her 16-year-old son, who was still in school, to take on a job to help pay other bills."* [36].

Reflecting on the second investigation that was launched two years later, the woman stated: *'The atmosphere at the meetings with the municipality is terrible'* and *'It took me two years to recover from this. I was destroyed mentally.'* [36].

6.2 Stakeholder analysis

In this section, we aim to identify and introduce the main stakeholders that played a role in the design, development, and use of the algorithmic decision-making system of Rotterdam. For ease of exposure, we split this overview into two parts, separating the internal and the external stakeholders. **We put in bold** the stakeholders that are also the actors; they are important for the analysis because they influence the behavior of the system during operation.

Internal stakeholders

These stakeholders exist in the Work and Income (W&I) department.

(a) **Team Reinvestigations**
(*Team HetOnderzoeken; THO*)
THO is responsible for the investigation of benefit recipients to ensure that *bijstanduitkering* is duly granted. Its employees are concerned with the actual evaluations; the selection process is handled by another team (see below) [196].

[196]: Rekenkamer Rotterdam (2021), *gekleurde technologie. verkenning ethisch gebruik algoritmes.*

(b) **Team Testing and Monitoring**
(*Toetsing en Toezicht; T&T*)
T&T provided domain expertise during the development of the model. They were also co-responsible for the operation of the model (along with its maintainers), focusing on the selection of the benefits recipients for re-examination [196].

(c) **Consultants** (*Consulenten*)
Consultants are the client-oriented professionals in the W&I department. They are responsible for interviewing the benefit recipients and storing information about them [196]. Thus, consultants (co-)created the dataset used by the model.

(d) **Team Complaints** (*Klachten*)
While Rotterdam has a municipality-wide complaints office, a designated unit within the W&I department is responsible for the handling complaints about welfare benefits, e.g., for customers who have not been treated fairly or have not received sufficient information.

(e) Client Council (*Cliëntenraad*)
The Client Council is an independent body within W&I that is tasked with representing the interests of all benefit recipients.

(f) **Concern Management** (*Concerndirectie*)
The municipality of Rotterdam is organized into six clusters, including the Work & Income cluster, which itself consists of **six units** [173]. The "concern manager" of W&I would generally bear ultimate responsibility for all matters related to the cluster, including algorithmic decision-making systems. Nonetheless, the model was developed with the support of another cluster – Management and Corporate Support (*BCO*) – and so the assignment of responsibility was not clear [196]. In Rotterdam, Concern Management additionally includes three independent organizational units; among these, **Concern Auditing**, which carries out the audits of systems, processes, and programs of the municipality.

We were not able to find reliable information about the **second-level structure** of the W&I department. We are also not certain about the specific unit responsible for *bijstand* (though there is reason to believe it falls under the auspices of the Income unit). Nevertheless, this information is not necessary to properly evaluate the system.

External stakeholders

This group of stakeholders exists outside of W&I Rotterdam.

(a) **Benefit recipients** (*Bijstandsontvangers*)
People who receive *bijstanduitkering*. Their eligibility may be re-evaluated every two years. If selection was always random, a typical benefit recipient in Rotterdam would expect a re-examination roughly once every six years.

(b) Accenture
A multinational tech consulting company that was contracted by Rotterdam to develop the initial version of model [35].

(c) **Team Research and Business Intelligence**
(*Onderzoek en Business Intelligence; OBI*)
OBI is the research team in the municipality *BCO* department. Employees of OBI complimented the Accenture team during development. Later, the team took over for the maintenance and operation of the system [196].

(d) Municipal Council (*Gemeenteraad*)
The assembly of elected representatives that sets municipal policies (including, for example, the local implementation of the Participation Act) and oversees their execution.

(e) **College of Mayor and Aldermen**
(*College van Burgemeester en Wethouders*)
The "executive board" of a municipality that is selected by the *Gemeenteraad*. Its tasks include the establishment of the structure of the municipal organization (thus also the W&I department) and the nomination of the *Concerndirectie*. In Rotterdam, the Alderman responsible for W&I coordinated the schedule for re-examinations with the department [196].

(f) Employee Insurance Agency
(*Uitvoeringsinstituut Werknemersverzekeringen; UWV*)
To receive *bijstand*, individuals must register as job seekers and submit their application to the *UWV*. As such, the agency fulfills a supporting role in the process. The mutual provision of information between the municipalities and the *UWV* is regulated by *Wet SUWI* (see Section 5.3).

(g) Ministry of Social Affairs and Employment
(*Ministerie van Sociale Zaken en Werkgelegenheid; SZW*)
Ministry *SZW* establishes the general regulations for social welfare, including the Participation Act, which are then interpreted and implemented by individual municipalities.

(h) **Privacy Officers *and* Data Protection Officer**
These stakeholders oversee the handling and protection of personal data by the municipality. While the former are employed by the *BCO* concern, the latter is an independent officer of the municipality that focuses on compliance with the GDPR and the "Police Data Act" (*Wet politiegegevens*; *WPG*).

(i) Data Protection Authority
(*Autoriteit Persoonsgegevens*)
The Dutch Data Protection Authority is an independent administrative body (*zelfstandig bestuursorgaan*) that supervises the processing of personal data in accordance with the GDPR in the Netherlands. Thus, it is the institution where people can lodge a complaint about data processing at the municipality.

(j) National Ombudsman *and* Ombudsman Rotterdam-Rijnmond
The National Ombudsman is an independent counsel that represents the interests of Dutch residents in relation to the public authorities. Moreover, the Rotterdam-Rijnmond region has also established a similar Ombudsman office. Benefit recipients unhappy with the decisions of the municipality can file a complaint at either institution.

(k) Civil Society Organizations (CSOs)
CSOs, such as Bits of Freedom, are the final stakeholders under consideration. CSOs can indirectly influence the system through advocacy efforts or policy initiatives.

Next steps

Our analysis in this chapter was entirely based on desk research methods. To further double-check its quality and learn about other requirements of the domain, we decide to conduct several interviews. These are discussed in Chapter 7, the next chapter. Afterward, we return to the case study in Chapter 8, interpreting the Rotterdam case through the lens of system engineering and evaluate the value of algorithmic recourse as a safety mechanism in this setting. We make use of these insights to inform our evaluation strategy in Chapter 9, and other design artifacts in Chapters 10 and 11.

Expert interviews 7

To verify the requirements for algorithmic recourse in the specific domain of social welfare, we planned to organize a series of interviews with decision-makers in several municipalities in the Netherlands, including, among others, Rotterdam. At the explicit request of the municipalities, they were approached through online contact forms. Our invitation for the interviews emphasized that the aim of the interviews would be to learn about the processes of social welfare in Work & Income departments and discuss the opportunities and barriers for algorithmic decision-making in this setting, focusing on algorithmic recourse mechanisms. The invitation letter that we submitted to the municipalities is available in Appendix C.

Even though several municipalities initially followed up on our request, informing us that it had been transferred to the responsible department, ultimately, none of them accepted the invitation. In fact, only **Rotterdam sent an email denying the interview**. Other municipalities (five total) did not provide us with a response.

The **decision-making processes in W&I Rotterdam are especially important for this research** given the case study. After the department turned down our request for the interview, we informed the officials in writing that the research would instead be entirely based on publicly available documents, including the outcomes of previous investigations, and the municipality would not be contacted again for further information or comments. Our response was left unacknowledged.

Ultimately, we distilled all necessary information from relevant national laws, internal regulations, and documents published by journalists associated with Lighthouse Reports. Thus, we decided that the interviews would still suit the goal of this thesis if they approached the topic broadly, rather than focusing solely on social welfare. On recommendation from several sources, we contacted **12 experts** in the field of (algorithmic) decision-making in public administration, of whom four graciously agreed to a meeting. One further expert was invited to jointly participate in an interview by their colleague. Thus, five experts shared their ideas with us.

Identities of the experts are known to authors; we approached academics and researchers, municipal employees and executives, and privacy officers.

We explain the process of conducting the interviews in Section 7.1. Then, in Section 7.2, we introduce and explain eleven themes that we observed in the interviewees' responses. Finally, we discuss the impact of their responses on our research in Section 7.3.

7.1 Setup and protocol

We conducted four interviews of one hour with a total of five experts in July and August 2024. Three of them were held online, and one was held in person. The interviewees were provided with the informed consent form several days before the scheduled date of the interview. At the latest, the form was signed at the beginning of the meeting and evaluated by the researcher in case the interview setup would require adaptation due to disagreement with explicit consent points, but this was never needed.

All interviews were recorded for subsequent transcription with a Zoom H1 voice recorder without access to the Internet. Additionally, the responsible researcher took notes to mitigate the impacts of possible recording failure. Our interviews followed a semi-structured format, with a small number of questions relating to the opportunities and barriers for algorithmic decision-making in public administration, and the potential value of AR mechanisms posed to all experts – other questions built upon topics brought up by the experts or corroborated findings from earlier interviews.

The recordings were manually transcribed using the **intelligent verbatim method** in three passes. In the first pass, the utterances were transcribed to the best of our abilities. Then, we added timestamps to individual statements and fixed any transcription mistakes. The final pass aimed to verify that the transcript was complete and (virtually) mistake-free. A very small number of words were inaudible in the recordings, they were marked as such in the transcripts. Our recordings were deleted after transcription and analysis.

We **removed filler words and repeated phrases, and fixed minor grammatical errors** where the speaker's intent was clear but otherwise directly transcribed the recordings.

The complete transcripts were archived on TU Delft servers in an anonymized form; they are accessible only to the research team. Our findings in Section 7.2 are distilled through a thematic content analysis. Although no statements can be traced back to the interviewees, it should be understood that the interviewees shared their own opinions based on their expertise in the field, and these opinions do not necessarily reflect the views of their employers. Furthermore, we asked the interviewees not to share any details that could be considered confidential by their organizations. Finally, a "99%" draft of this chapter was submitted to the interviewees so that they could **approve** the outputs and provide any necessary corrections within at least two working weeks.

All interviewees approved the draft without having requested any changes.

Our complete process, including the setup of the interviews and the management of data artifacts, has been verified by the EEMCS Faculty Data Steward and approved by the Institutional Review Board on March 24th 2024 with the "minimal risk" designation.

7.2 Thematic content analysis

The complete transcripts count slightly over 19000 words. To analyze them, we employ a standard thematic content analysis. We read through the set of complete transcripts and find recurring themes in the statements of the interviewees. After defining 11 themes, we re-read the transcripts and categorize (groups of) statements to corresponding themes whenever applicable; some statements relate to more than one theme, in which case they are assigned to all relevant themes. In the following section we summarize our analysis by employing two criteria of importance: we quote all sentiments voiced by multiple interviewees and additionally highlight considerations that are particularly pertinent to this work.

7.2.1 Perceived advantages of ADM systems

When asked broadly about the potential advantages of algorithmic decision-making systems, all interviewees emphasized their high efficiency and efficacy compared to human decisions. Further, two interviewees mentioned fairness of outcomes, at least to the extent that the same inputs will lead to the same outputs, with the caveat that biases may enter the system already at the level of data.

7.2.2 Perceived barriers to the adoption of ADM systems

The interviewees brought up several problems associated with the use of algorithms. Maybe the most relevant consideration for this document is that the model outputs require careful interpretation by decision-makers, for instance, because they may be based on faulty data or hold only as statistical truths. One expert pointed out that the application of algorithms is not always preceded by an analysis of whether they are needed in a task. Another noted that algorithms cannot account for contextual factors often important in public administration. Finally, a third expert explained that *"everyone is waving big flags about AI anyways. I think we are more attuned to problems and risk regarding AI than to* ***simple systems****"*.

Such **simple systems** have also been excluded from regulation under the EU's AI Act. Recital 12 informs us that *"the definition [of AI] should be based on key characteristics of AI systems that distinguish it from simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations."*

7.2.3 Characteristics of ADM systems in governance

Multiple interviewees confirmed that algorithms in the governance context tend to take the form of **if-then statements** rather than machine learning models. Many problems in governance have *"standard solutions"* that encourage the use of algorithms; moreover, many processes would not be possible without such algorithms due to the amount of data that the government needs to parse. This does not mean expert systems operate without issues. One interviewee pointed out that even interpretable if-then statements can be nested so densely that they become challenging to understand (e.g., the tax system), or their implementation may be incorrect. Finally, another interviewee highlighted that ADM systems used in governance need to fulfill strict requirements related to **fundamental rights**, which can be evaluated with the FRAIA/IAMA framework

Expert systems are a natural choice in this setting because "if-then" rules follow directly from relevant legal acts.

The AI Act obliges certain organizations to carry out **fundamental rights** impact assessments. Recital 96 states that *"In order to efficiently ensure that fundamental rights are protected, deployers of high-risk AI systems that are bodies governed by public law, or private entities providing public services and deployers of certain high-risk AI systems listed in an annex to this Regulation, such as banking or insurance entities, should carry out a fundamental rights impact assessment prior to putting it into use."*

7.2.4 Characteristics of processes in social welfare domain

Three topics stand out in this theme. First, as one expert explained: *"We have to focus on the human (...). Make the system human, see the people, know the people (...)"*, highlighting the importance of contextual factors that may not be readily embedded into algorithms. Second, while national laws shape the social welfare benefits, individual municipalities have some freedom regarding their practical implementation. Third, the work of the consultants has a discretionary character: they may deviate from municipal regulations if needed. Relatedly, a consultant's decisions with regard to one provision (e.g., lowering the amount of assistance) may have downstream impacts on various other duties of the municipality (e.g., ensuring that citizens do not have arrears).

7.2.5 Considerations of involved stakeholders

As recognized by one expert, *"developing an algorithm is actually a non-technical question"* because it involves perspectives from ethics, law, experts (e.g., client-oriented professionals), or the management of organizations. When asked about involving citizens, the expert noted *"Naturally, you get citizens who are interested in AI or algorithms that will come to those focus groups, and obviously that is not really the group you are actually looking for"*. This was recognized by another interviewee who confirmed that some municipalities have tried to co-create algorithms with their residents, but these efforts have been mostly unsuccessful. As an alternative approach, a third expert suggested that non-governmental or civil society organizations may *"play part of that role of citizens"*.

7.2.6 Outlook on black-box/opaque/complex models

We observe two types of sentiments. One group of experts recognized that public administration in the Netherlands has become very cautious of opaque models, especially since laws such as the *Awb* demand a justification for administrative decisions. The other group believed that opaque models are permissible if their operators have the tools and the know-how to interpret their predictions.

7.2.7 Outlook on algorithmic recourse mechanisms

While our experts recognized the potential of algorithmic recourse as a safety mechanism, they also noted that its practical value would be heavily context-dependent. Two additional considerations are worth highlighting. One expert mentioned that recourse could be a way to give *"citizens voice to have their own input in the process"*, emphasizing the problem of power asymmetries intrinsic to public administration. Another expert pointed out that the usefulness of AR will depend on the ability to generate *"tailor-made recommendations"*, linking to the importance of individual actionability.

7.2.8 Considerations of transparency in decision-making

The interviewees mentioned three goals of transparency. First, it allows for a justification of decisions: *"If you use an algorithm and use the insights to make a decision about a person, I think you should be able to explain why that decision is made"*. Second, it is a way to ensure that the decision-makers can be held accountable. Third, it may enable the model owners to learn from the algorithm. At the same time, transparency can be described on a scale – two experts mentioned that a large degree of transparency might be overwhelming to one stakeholder (e.g., a citizen) but the same amount of information will be useful to another stakeholder (e.g., an auditor). While multiple interviewees agreed that a large degree of transparency could involve potential risks for some applications (e.g., opening them up to misuse), one expert pointed out that in many cases *"rules are spelled out in legislation"* so being secretive about models that implement them *"does not make any sense"*.

7.2.9 Considerations of agency of end-users

The most important insight pertaining to this theme is that contestability requires a high level of procedural awareness from citizens, which manifests itself in two distinct ways. On the one hand, as one expert explained: *"We have a lot of institutions where you can complain, but the threshold is very high"*. Another interviewee concurred: *"We in the government have much more power (...) than the citizens. (...) it is a big step to complain to us"*. On the other hand, many people may not even know **how and where to complain**: *"[W]e also have in the Netherlands the Complaint Law – Klachtrecht – but how many people are using the Klachtrecht?"*. As recognized by a different expert, every received complaint signals many more complaints that were never filed; this requires appropriate procedures from governmental organizations.

In practical terms, the *Awb* distinguishes between **appeals** (*beroepen*) to a court, and **complaints** (*klachten*) lodged directly against the administrative body that issued the decision.

7.2.10 Considerations of oversight and audit capabilities

Many different stakeholders were listed as having some responsibility for overseeing algorithmic decision-making systems, including model developers, data scientists and other algorithm experts, internal audit units within organizations, external agencies such as Autoriteit Persoonsgegevens, Rekenkamers, and Directie Coordinatie Algoritmes, or civil society organizations. Three experts were asked about the theoretical value of having citizens audit ADM systems in public administration. They agreed that such mechanisms would be beneficial and, as one expert noted, *"it would also provide the municipalities with a lot of valuable information"*. In any case, multiple experts agreed that algorithmic systems require strong, explicit practices to monitor them throughout their lifespans.

7.2.11 Considerations of other principles and values

The interviewees discussed a variety of principles and values relevant to the domain of interest. They brought up the values of accountability, moral correctness, privacy, service, and trust (in other humans and in algorithmic decisions). They also highlighted three important guiding principles: respect for human rights and fundamental rights, respect for identity, and political neutrality.

7.3 Discussion of the interviews

While our interviews approached the topic of algorithmic decision-making in public administration relatively broadly, several issues brought up by the experts have strong ties to the problem of operationalizing algorithmic recourse in such domains.

7.3.1 Algorithmic recourse in expert systems

As recognized by the experts, the vast majority of algorithmic decision-making systems in public administration use cases are based on simple if-then statements. These are far removed from machine learning models that have been the focus of research on algorithmic recourse – rules are not automatically inferred from data but rather follow directly from legislation.

At the same time, expert systems may still pose a variety of risks. For example, even simple algorithms may be implemented incorrectly. One expert recalled a case in an unnamed municipality where the system used to evaluate if citizens have paid off their debts relied on the strict equality between the amount of debt and the amount transferred by the citizen. Thus, if a citizen transferred one cent less or one cent more, they were marked as if having arrears with the municipality. Naturally, the risk of faulty implementation increases with the complexity of a system. While the latter is bound by legislation, expert systems may still become so complex that they are challenging to interpret, such as in the case of the tax system.

In Chapter 4, we have presented a survey of algorithmic recourse solutions. It is important to emphasize that none of them have been developed for expert systems. Nonetheless, there exist use cases where actionable recommendations may be useful for citizens in public administration contexts. We propose an algorithm to generate algorithmic recourse for expert systems in Chapter 11.

7.3.2 Encouraging contestation among end-users

The interviewees emphasized that contesting administrative decisions may be prohibitively difficult for citizens, especially as in domains such as social welfare, they are in a vulnerable position. Algorithmic recourse, but *not* explanations, provided on an opt-out basis may be a good way to encourage contestation of outcomes.

There are at least two merits of algorithmic recourse in this context. First, if the success of interventions can be guaranteed, it becomes a basis for the affected people to **exercise their rights**. Second, providing a citizen with a suggestion on how to improve their outcome is an implicit acknowledgement that the government *wants* their outcomes to be positive. Although administrative decisions are always provided with a justification why they have been taken over alternatives – such as the legal basis or the relevant factors – an actionable recommendation may go a step further by encouraging the citizen to *act* on a negative outcome.

From AI Act Recital 171: *"**Affected persons should have the right to obtain an explanation** where a deployer's decision is based mainly upon the output from certain high-risk AI systems that fall within the scope of this Regulation and where that decision produces legal effects or similarly significantly affects those persons in a way that they consider to have an adverse impact on their health, safety or fundamental rights. **That explanation should be clear and meaningful and should provide a basis on which the affected persons are able to exercise their rights.**"*

7.3.3 Broader landscape of (safety) interventions

The experts emphasized that the value of algorithmic recourse will depend on its specific application: *"The context, the environment, the service that you want to provide, but also the people"*. Indeed, various other measures to mitigate algorithmic risks have been mentioned: fundamental rights impact assessments, procurement guidelines for algorithmic systems, complaints offices, or high standards for the registration of algorithms in the Algorithm Register.

Hence, the addition of algorithmic recourse into a system should be preceded by the stock-taking of solutions that are already in place. As one example, if an organization does not have strong procedures to deal with complaints, these may be a better investment of initial resources. As recognized by [136], algorithmic recourse **should not be considered a panacea**, but it is also essential to recognize that safety interventions need not be mutually exclusive.

[136]: Leben (2023), 'Explainable AI as Evidence of Fair Decisions'

It has also been argued, e.g., that post-hoc explainability methods on their own **do not guarantee sufficient transparency** to establish the compliance of black-box models with the EU non-discrimination laws [246].

7.3.4 Discretional character of decision-making

Social welfare processes allow for discretionality from municipal officers. When asked about a person's ability to predict the outcomes of their applications for assistance, an expert stated *"For 99.99% times I am sure it will have the outcome you predict, but it is not 100%"*; this may extend to recourse. For example, if the citizen is asked to submit 10 job applications to show their engagement, but they manage to submit only 9 applications, a consultant may still recognize it as a requirement duly fulfilled and grant a better outcome.

Thus, the actionability desideratum of algorithmic recourse must not only account for personal characteristics of the citizen, but also the discretional mandate of the municipal officers. In principle, this could be achieved by implementing fuzzy logic or similar technical solutions, but it will remain a (weak) approximation of contextual factors that are difficult to capture in numbers. Instead, it may be one of the settings where the goals of algorithmic recourse need to be achieved by relying on social or organizational interventions.

7.3.5 Democratic audit of models

Finally, we propose an additional value of algorithmic recourse and counterfactual explanations – they may become a simple tool to allow citizens to audit governmental models, provided that the explanations and recommendations are sufficiently faithful [16]. If a citizen is provided with an explanation such as *"Our algorithm suggests you are at a high risk of misusing allowances. Unless you increase the frequency of your meetings with a case worker by X, we will launch a reexamination effort within Z weeks"* the important factor behind the model's decision would likely be perceived as relevant to the problem and accepted by the citizen. If, instead, they are **recommended to improve their appearance or move to a different city district**, recourse unveils the explicit or implicit biases in the model. This function of algorithmic recourse benefits from the scale at which public administration models are applied – if the model provides a prediction on tens of thousands of citizens, even "weak" biases may be reflected in hundreds of recommendations, and in turn, there is a chance that at least one citizen will inform the model owner about the failure of the model. It is also a setting where providing the citizen with *both* an explanation (i.e., the direct reasons behind a decision) and an algorithmic recourse recommendation (i.e., the solution to improve the decision) may be of value. While democratic auditing of models through CEs and AR may be an important tool in the "participatory AI" toolbox [33], to the best of our knowledge, **this line of work has not been explored** yet.

[16]: Altmeyer, Farmanbar, Deursen, and Liem (2024), 'Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals'

The judgment of physical appearance and the district of residence were **among 315 features used by the Rotterdam black-box reexamination model**.

[33]: Birhane, Isaac, Prabhakaran, Diaz, Elish, Gabriel, and Mohamed (2022), 'Power to the People? Opportunities and Challenges for Participatory AI'

Other authors have recognized the value of AR as a method to gain insights into the model and the data (see Section 4.2.7), and CEs have been postulated as useful for tasks such as model debugging [e.g., 1, 220]. Our main point here is about *who* performs the audit.

8 System analysis

In this part, we combine the insights from Chapters 5 – 7 to look at algorithmic recourse through the lens of a real-world system where it could be employed. Following our own recommendations from Section 4.3, we expand the boundary of the analysis to include social and organizational components of the Rotterdam case.

Although the model used by W&I Rotterdam had significant shortcomings, we note that its **technical rollout was, overall, reasonable**: the development was guided by the city's own data scientists and domain experts, the department organized multiple pilot stages with an increasing number of participants, and when the final model still performed below expectation it was entirely decommissioned. Consequently, we do not analyze a specific incident. Instead, we reflect on the operating process and functional requirements of the re-examinations in Section 8.1. Then, we develop the process models of human and automated controllers in Section 8.2 and decide if there was any potential for inadequate control in Section 8.3. Finally, we consider the value of algorithmic recourse as a mechanism to improve the safe operations of the model in Section 8.4.

We must also commend Rotterdam for **investing in practically oriented research** on responsible digitalization with its Creating010 Research Centre (Kenniscentrum) [e.g., 49, 165].

Even so, on the organizational side, Rotterdam W&I failed to satisfy multiple **ethical standards** as concluded by the Rekenkamer, including the lack of clearly assigned responsibility for the complete system, unsatisfactory transparency towards affected benefit recipients, and inadequate motivation for decisions related to ethics [196].

This chapter heavily relies on the tools of Systems-Theoretic Accident Model and Processes (STAMP) framework proposed by [138] and operationalized for AI use cases by [61]. STAMP aims to mitigate accidents by enforcing (behavioral) constraints on the system.

[138]: Leveson (2016), *Engineering a safer world: Systems thinking applied to safety*

[61]: Dobbe (2022), *System Safety and Artificial Intelligence*

8.1 Operating process

Leveson explains that in systems theory, *"systems are viewed as hierarchical structures, where each level imposes constraints on the activity of the level beneath it"* [138, p. 80]; these constraints dictate the behavior of the system and their inadequate (or missing) application leads to accidents. A complete *sociotechnical hierarchical safety control structure* would consider all agents that affect the final system including, e.g., legislatures and regulatory bodies (as they create legal frameworks for the system) or corporations involved in its development.

Our work focuses on the specific topic of algorithmic recourse. Hence, a complete safety control structure is not our goal. We draw the boundaries of analysis around the operating process, i.e., the set of components and activities immediately relevant to the (daily) operations of the system. We distill information from publicly available documents of [35, 173, 196] to develop a functional control diagram of the main operating process as presented in Figure 8.1.

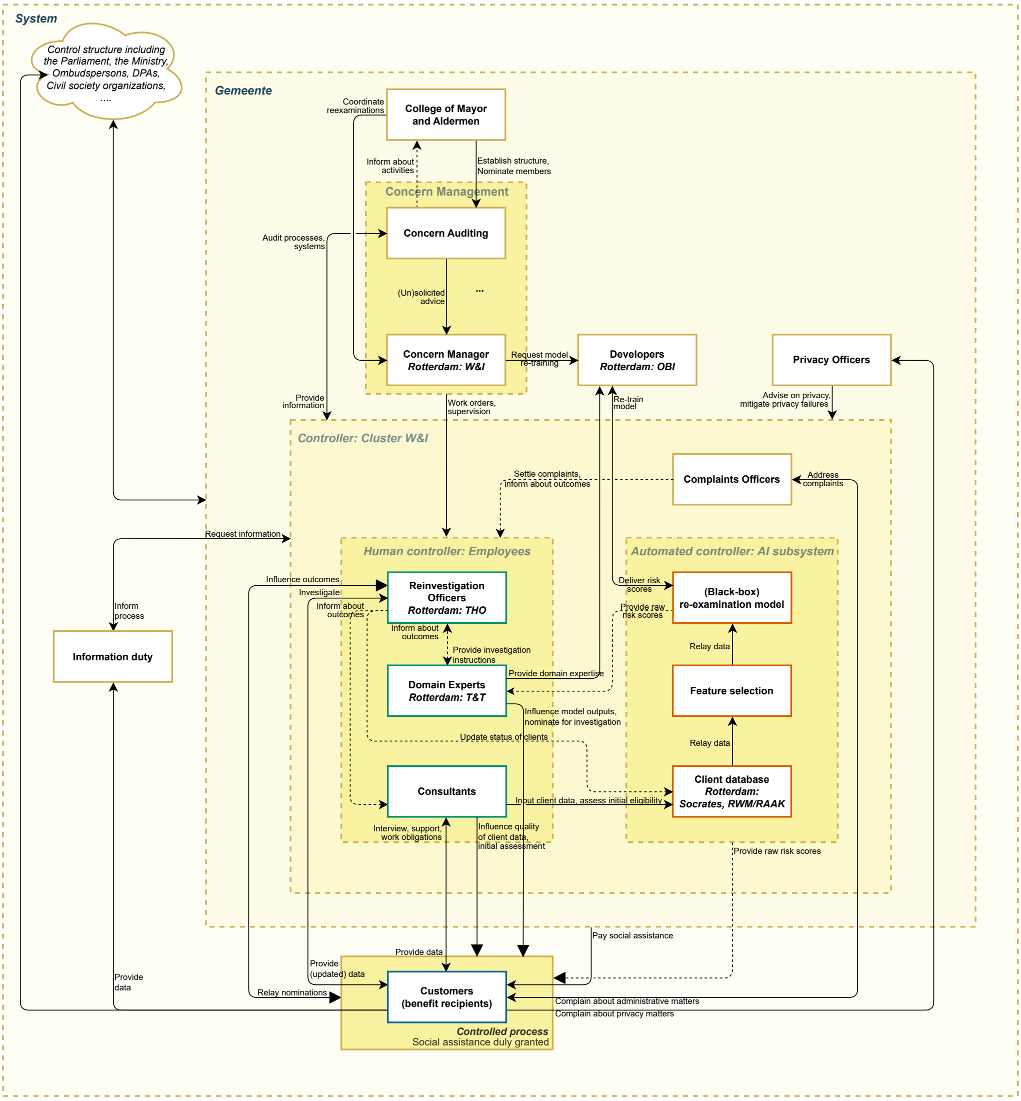
System
Control structure including the Parliament, the Ministry, Ombudspersons, DPAs, Civil society organizations, ...
Gemeente
Coordinate reexaminations
College of Mayor and Aldermen
Inform about activities
Establish structure, Nominate members
Concern Management
Audit processes, systems
Concern Auditing
(Un)solicited advice
...
Concern Manager Rotterdam: W&I
Request model re-training
Developers Rotterdam: OBI
Privacy Officers
Provide information
Work orders, supervision
Re-train model
Advise on privacy, mitigate privacy failures
Controller: Cluster W&I
Settle complaints, inform about outcomes
Complaints Officers
Address complaints
Request information
Human controller: Employees
Automated controller: AI subsystem
Influence outcomes
Investigate
Inform about outcomes
Reinvestigation Officers Rotterdam: THO
Deliver risk scores
Provide raw risk scores
(Black-box) re-examination model
Inform process
Information duty
Inform about outcomes
Provide investigation instructions
Relay data
Provide domain expertise
Domain Experts Rotterdam: T&T
Feature selection
Influence model outputs, nominate for investigation
Update status of clients
Relay data
Consultants
Input client data, assess initial eligibility
Client database Rotterdam: Socrates, RWM/RAAK
Interview, support, work obligations
Influence quality of client data, initial assessment
Provide raw risk scores
Pay social assistance
Provide data
Provide (updated) data
Provide data
Relay nominations
Customers (benefit recipients)
Complain about administrative matters
Complain about privacy matters
Controlled process
Social assistance duly granted

Figure 8.1: Functional control diagram of the main operating process in the Rotterdam case.
Relationships marked by solid lines are established based on [35, 173, 196]; relationships marked by dashed lines are assumed. We focus on the system and omit many parts of the hierarchical control structure; they are highlighted in the thought bubble. To increase readability, we collapse bi-directional arrows; the responsibility of an actor is stated on their side of a relationship.

Our evaluation focuses on the actors identified in Section 6.2 and corresponds to the state of the operating process when the model was still in use. Wherever relationships can be established with certainty, we mark them with a solid line. **A small number of constraints is unspecified** and we use our best judgment to infer the missing elements. In any case, we believe that such assumptions do not affect the following analysis in any meaningful manner.

For example, OBI is asked to re-run the model but it is not clear *who* submits this request [196], and thus we assume it is a responsibility of the concern management, as they are also involved in the initial planning of re-examinations.

Based on our analysis and insights of the experts in Chapter 7, we identify ten functional requirements relevant to the case study:

(a) maximize the number of **properly targeted re-examinations**;
(b) minimize the number of unnecessary re-examinations;
(c) minimize the impacts of re-examinations on benefit recipients;
(d) build a model that relies on factors relevant to the aim pursued;
(e) build a model that can improve over time;
(f) allow W&I officers to understand model nominations;
(g) allow W&I officers to review and revise model nominations;
(h) ensure sufficient transparency toward benefit recipients;
(i) ensure high reliability of the model and its nominations;
(j) attend to contextual factors in the selection process.

Simply put, true positives. Municipalities want to conduct re-examinations that actually lead to the **identification of benefits unduly granted**.

8.2 Process models

Having discussed the operating process, we turn our attention to the process models of the human and AI controllers. Leveson distinguishes four components of a process model [138, p. 87]:

(a) *goal*: safety constraints to be enforced by the controller;
(b) *action condition*: tools to enforce these safety constraints;
(c) *observability condition*: feedback from the controlled process;
(d) *model condition*: controllers' understanding (the estimated model) of the behavior of the controlled process.

8.2.1 Controlled process

We consider "social assistance duly granted" to be the controlled process. This means that *bijstand* should be provided to people (1) who need it, (2) who are eligible to receive it, (3) for the periods when they are eligible, (4) and at the amounts for which they are eligible. With this, the goals are two-fold. On the one hand, people who are not (or no longer) eligible for *bijstand* should be re-examined without delay. On the other hand, people who are eligible for *bijstand* should not have their privileges unnecessarily suspended during the re-examination, *cf.* case of the Rotterdam resident in Section 6.1. Of course, the latter depends on the administrative procedures of re-examinations themselves, rather than the procedures of nomination, but we note that minimizing the number of false positive nominations would equivalently address this issue.

8.2.2 Human Controller: Employees

Three teams in Rotterdam W&I jointly fulfill the role of the human controller. First, *consultants* are responsible for the quality of the data as their evaluation of the clients is stored in the database and used to train the model. Moreover, they carry out the **initial approval** of applicants. Second, *domain experts* informed the development of the model, including data selection. Based on the ranking generated by the model and other selection tools, they create a list of benefit recipients that must undergo re-examination. Third, *reinvestigation officers* carry out the actual assessment of benefit recipients to decide if they remain eligible for *bijstand*. We do not consider the OBI team to control the process as they act as an extension of the W&I teams in this setting, but their activities still have an impact on the system. For example, they may introduce bugs into the model.

In practice, the **initial approval** is also supported by an algorithm. Its entry in the Algorithm Register informs us that the algorithm proposes a decision to the consultant along with an explanation thereof. See also Section 5.4.

Goal. Ensure that residents have sufficient income to participate in society without falling into arrears, and protect social assistance processes from abuse by people who are not eligible for support.

Action condition. Affect the quality of nominations by modifying selection criteria for the model, safeguarding the quality of the training data, and revising the predictions so that, for instance, residents are not re-examined more than legally permissible. Additionally, they may (attempt to) identify implicit biases in the model by actively monitoring the groups of benefit recipients that are recommended for re-examination and relaying feedback to developers.

Observability condition. Cannot reliably know the number of false negatives but, after re-examination, can guarantee the number of true positives and false positives, which can be a measure of success. Reliable identification of the implicit biases of the model depends on the explainability standards but, in principle, it is possible to evaluate if (groups of) citizens are nominated for re-examination more often than expected based on features and outcomes.

Model condition. Human controller is formed by three groups of professionals who are knowledgeable about social welfare processes, so their models likely correspond to reality. Nonetheless, coordination issues between them may negatively impact the model condition; while the teams have non-overlapping responsibilities, they must still effectively share knowledge. For example, if a reinvestigation officer does not know which method was used to nominate a benefit recipient, they will not be able to decide if the model is reliable. As the automated controller relies on an opaque model, the W&I officers may have a general idea of its logic, but reviewing specific decisions requires additional (interpretability) mechanisms.

8.2.3 Automated Controller: AI subsystem

In principle, the automated controller impacts the controlled process only indirectly because its predictions are supposed to be filtered by humans. Nonetheless, the analysis of Lighthouse Reports unveiled that the model predictions exhibited biases [35]. Yet, the city carried out **a large number of re-examinations acting on the risk scores** assigned by the model. This suggests that the AI subsystem had a non-negligible *direct* impact on the controlled process.

In 2019, the model was used to carry out the targeted **nomination of 1,376 individuals**. Altogether, Rotterdam re-examined 6,232 recipients, meaning that the model was responsible for 22% of all re-examinations [196].

Rotterdam employed a gradient boosting machine model [161] trained on 315 features. Thus, the model predictions could not be readily interpreted, both due to the opaque algorithm that generated them as well as the complexity of the underlying data. As many feature names are uninformative for someone without domain expertise (e.g., *contacten_onderwerp_arbeidsdiagnose_dariuz*), we make use of the documents published by Lighthouse [82] to translate the features to English and attempt to make sense of their meaning. In broad terms, the features belong to 14 categories, including types of appointments with the municipality, barriers to reintegration into the workplace, personal characteristics, or relationships. Rotterdam's data scientists calculated the relative importance of all features compared to the baseline of "age at investigation". The relative importance scores of the top 10 features are stated in Table 8.1.

Table 8.1: Top features by importance.

Feature	Score
Age at investigation	100.0
Number of existing cost-sharer relationships	35.5
Number of no-show appointments	27.9
Number of applying expertise competencies	25.1
Number of contacts about income	24.1
Number of days living at current address	23.3
Presence of existing cost-sharer relationships	20.7
Presence of development action plans in the past	19.1
Number of instruments used in activation ladder	17.5
Number of contacts last year about outgoing documents	15.4

Goal. Support the W&I officers in the targeted re-examination of benefit recipients most likely to engage in fraudulent activities.

Action condition. Discover patterns in historical reinvestigation data to associate the profiles of current *bijstand* recipients with risk scores, which are converted into a ranking and provided to domain experts. The AI subsystem relies on a gradient boosting machine model – an ensemble of weak tree learners – to make predictions.

Observability condition. Observes the state of the system with delay as it is re-trained and re-run only once a year [196, p. 40]. It receives a form of feedback on the generated matches when W&I officers update the training data with outcomes of re-examinations.

Model condition. There is an inherent but necessary mismatch between the automated controller's understanding of the system and its state – the model relies on a regression/ranking algorithm, but the feedback it receives can only take the form of a binary label: eligible or non-eligible. Further, it may be impossible to decide with certainty whether the black-box model has learned valid associations. While surface-level biases can be **readily discovered**, even then it can take several reinvestigation cycles to observe the patterns. Finally, the behavior of the automated controller, and thus its model of the system, can be modified by human experts through means such as feature selection or explicit debiasing strategies [29].

Refer to the **method in the Lighthouse Reports investigation**. The journalists looked at statistical measures of outcome parity and then at the influence of individual attributes on these outcomes. This procedure allowed them to evaluate the model through more complex archetypical personas such as a *"Financially Struggling Single Mother"* or a *"Migrant Worker"* [35].

8.3 Inadequate control

As already explained, we do not intend to analyze a specific incident, but rather assess the hazards – potential causes of future safety incidents – that could emerge in the system. Thus, we apply a (simplified) form of Safety-Theoretic Process Analysis (STPA). As explained by Leveson, STPA is used to collect *"information about how the behavioral safety constraints, which are derived from the system hazards, can be violated"* [138, p. 212]. There are four types of inadequate control actions that could put the system in an unsafe state:

(a) Control actions never provided or never followed.
(b) Control actions provided at the incorrect time.
(c) Control actions executed for an incorrect amount of time.
(d) Unsafe control actions provided.

We do not intend to describe inadequate control in the complete system; we only look at the control actions related to the use of a machine learning model. For this, we revisit the operating process in Figure 8.1 and consider which controls may fail in the system.

We identify four scenarios of inadequate control related to data quality, interpretation of outputs, transparency, and handling of complaints. These are mapped in Figure 8.2 and explained in Sections 8.3.1 to 8.3.4. For each scenario, we describe what may go wrong (i.e., what types of hazards it could produce), discuss why it ought to be addressed (i.e., how likely it is to produce said hazards), and theorize whether algorithmic recourse could be applied as a risk/harm mitigation strategy. For the last part, we emphasize that safety interventions should not be considered mutually exclusive. We consider algorithmic recourse as *one possible mechanism*.

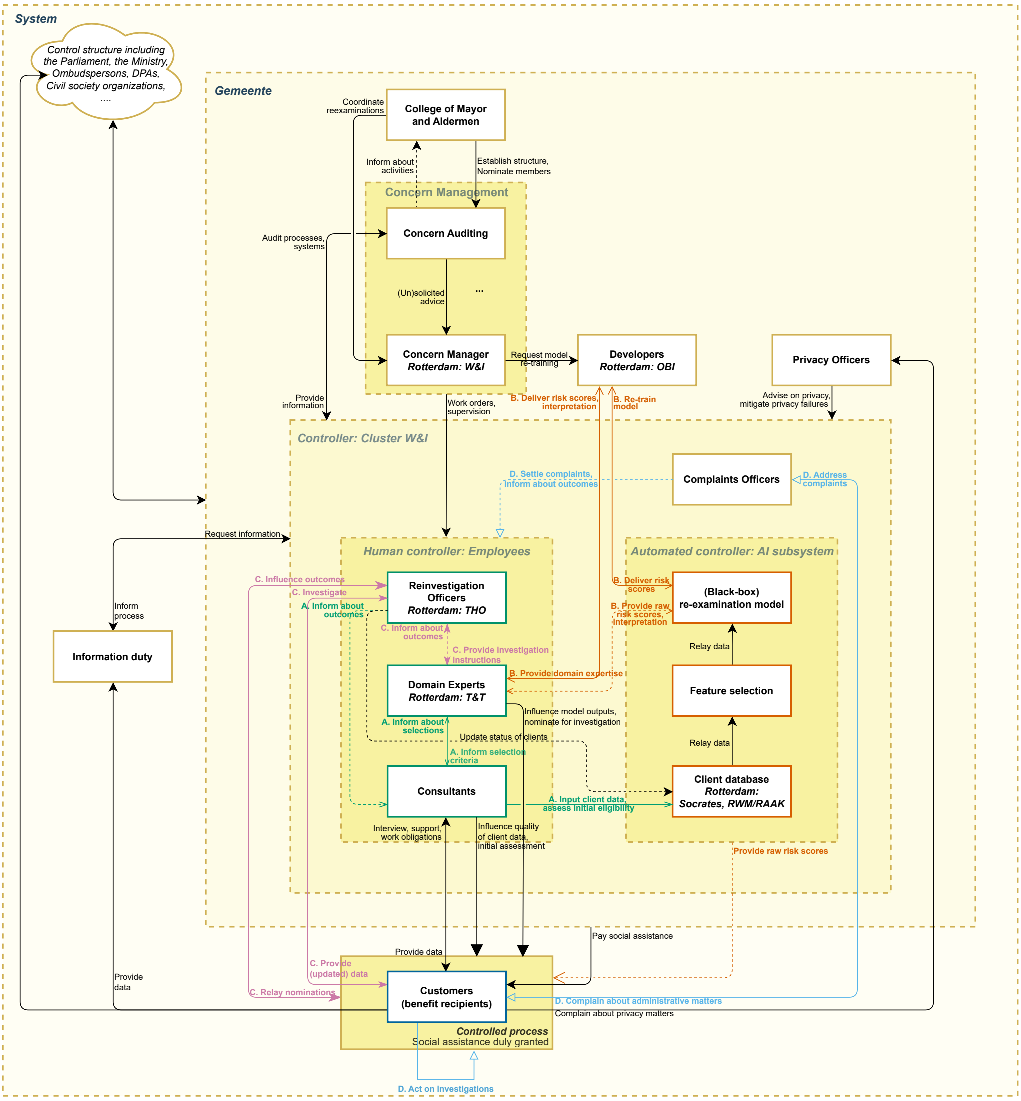
System
Control structure including the Parliament, the Ministry, Ombudspersons, DPAs, Civil society organizations, ...
Gemeente
Coordinate reexaminations
College of Mayor and Aldermen
Inform about activities
Establish structure, Nominate members
Concern Management
Concern Auditing
Audit processes, systems
(Un)solicited advice
...
Concern Manager Rotterdam: W&I
Request model re-training
Developers Rotterdam: OBI
Privacy Officers
Provide information
Work orders, supervision
B. Deliver risk scores, interpretation
B. Re-train model
Advise on privacy, mitigate privacy failures
Controller: Cluster W&I
D. Settle complaints, inform about outcomes
Complaints Officers
D. Address complaints
Request information
Inform process
Information duty
Human controller: Employees
Automated controller: AI subsystem
C. Influence outcomes
C. Investigate
A. Inform about outcomes
Reinvestigation Officers Rotterdam: THO
B. Deliver risk scores
(Black-box) re-examination model
B. Provide raw risk scores, interpretation
C. Inform about outcomes
C. Provide investigation instructions
Relay data
B. Provide domain expertise
Domain Experts Rotterdam: T&T
Feature selection
Influence model outputs, nominate for investigation
A. Inform about selections
Update status of clients
Relay data
A. Inform selection criteria
Consultants
Client database Rotterdam: Socrates, RWM/RAAK
A. Input client data, assess initial eligibility
Interview, support, work obligations
Influence quality of client data, initial assessment
Provide raw risk scores
Pay social assistance
Provide data
C. Provide (updated) data
Provide data
C. Relay nominations
Customers (benefit recipients)
D. Complain about administrative matters
Complain about privacy matters
Controlled process
Social assistance duly granted
D. Act on investigations

Figure 8.2: Interventions on the main operating process in the Rotterdam case.
We identify four scenarios (letters, colors, and arrow types) of inadequate control related to the use of a black-box model.

8.3.1 Scenario A: insufficient data quality

Unsafe control actions are provided

Description: Two types of unsafe control actions can be identified with respect to this scenario: (1) the use of uninformative or harmful features and (2) the collection of **biased data**. The "garbage-in, garbage-out" principle is at play here: non-meaningful data is likely to produce non-meaningful outcomes because the model learns spurious correlations rather than worthwhile patterns in the data.

Relatedly, the Rotterdam model was trained on historical re-examination outcomes, where the **selection mechanisms often targeted specific groups** of benefit recipients, and thus negative outcomes were more likely to be associated with these groups [196].

Discussion: In its analysis of the case, Algorithm Audit established a normative commission that proposed a set of eligible and non-eligible profiling criteria for this setting [10]. Their framework includes eight criteria for the analysis of profiling variables, such as clear *"linkage with aim pursued"*, *"subjective"* evaluation, or propensity to produce *"proxy discrimination"*. Many variables in the Rotterdam W&I model would be automatically disqualified under this framework, e.g., the evaluation of personal competencies and personal characteristics, or the number of children by age group.

Could algorithmic recourse help? **No.** At the same time, we observe that many examples of criteria provided by Algorithm Audit as **valid grounds for profiling**, e.g., *"no show at appointment with municipality"*, *"reminders for providing information"*, *"participation in trajectory to work"* [10], are highly-actionable for benefit recipients. From the first principles, this seems reasonable: an individual's attitude towards the parts of the process they can affect may be indicative of their attitude towards the entire process. Thus, deciding on the meaning of actionability in a specific context (as discussed in Section 4.3) could be informative for feature selection.

Other **eligible criteria** suggested by Algorithm Audit include age, type of living situation, or cost sharing.

8.3.2 Scenario B: erroneous interpretation of outcomes

Control actions are never provided or never followed

Description: There are two types of control actions that could be missing in this scenario. First, domain experts are expected to intervene on the outcomes of the profiling model, but it may happen that a required intervention is not carried out. Second, the model does not provide any explanation of the logic involved in its decision-making, which makes many interventions impossible.

Discussion: While T&T officers may be able to apply simple interventions (e.g., exempt certain people from the re-examination, or assess if certain groups are overrepresented among the top risk scores), any more complex interventions – such as identifying and acting on model biases – will be difficult without a reliable understanding of the logic behind specific decisions. Moreover, a model that outputs interpretable signals may contribute to the generation of organizational knowledge. For example, it may suggest fraud patterns for subsequent analysis by domain experts.

Could algorithmic recourse help? **Yes.** Algorithmic recourse may be a valuable mechanism to improve the understanding of patterns discovered by the model. Concretely, this setting yields itself to the *actionable knowledge discovery* alternative formulation of AR where automated techniques are used to distill knowledge that may be useful for decision-makers from the parameters of the black-box model [e.g., 3, 38], e.g., informing them which groups of benefit recipients tend to be associated with higher risk of fraud.

8.3.3 Scenario C: weak transparency standards

Control actions are never provided or never followed

Description: As highlighted by [196, p. 42], the benefit recipients were not informed when they were investigated due to algorithm-driven selection procedures, and it is unclear if the reinvestigation officers were aware of the selection tools applied in each case.

Discussion: Transparency with regard to the selection methods is necessary for the reinvestigation officers to identify wrongful nominations and evaluate the machine learning model, and for the benefit recipients to exercise their rights. Moreover, sufficient transparency standards towards data subjects are required by GDPR Articles 13(2)f, 14(2)g, and 15(1)h as discussed in Section 5.3.

Could algorithmic recourse help? **No.** This scenario of inadequate control relates to an organizational process. Open communication about the selection methods is a prerequisite to algorithmic recourse.

8.3.4 Scenario D: unreliable handling of complaints

Control actions are provided at the incorrect time

Description: Again, we observe two forms of inadequate control. First, benefit recipients may be reluctant to submit complaints. Second, a "stable" model is likely to nominate the same individuals for re-examination multiple times if their features do not change.

Discussion: As established in Section 7.2.9, it is difficult for people to complain about unfair treatment when the decision process is **necessary for the fulfillment of their basic needs**. Thus complaints may be delayed with respect to the moment when harms begin to occur. Moreover, Section 7.2.10 highlights that (consequential) algorithmic systems require a variety of strong oversight mechanisms.

The **objective of W&I customers** is *not* receiving benefits, but, e.g., having enough money to avoid food insecurity.

Could algorithmic recourse help? **Yes.** Its potential values for these scenarios have been emphasized in Section 7.3. Algorithmic recourse allows benefit recipients to self-control the process because certain behaviors that affect risk scores can be **reasonably expected from them**. Moreover, it could be an explicit encouragement to file complaints because it shows benefit recipients that the W&I department prefers that its clients receive beneficial outcomes. Finally, it may serve as an additional auditing mechanism that lies with end-users, rather than the parties operating the model.

For example, **the municipality may inform a benefit recipient**: *"You have been consistently missing appointments. If you would like to reduce the likelihood of a re-examination, please attend the appointments on time from now on."*

8.4 Algorithmic recourse as a safety mechanism

Across the four scenarios of inadequate control, we identified four ways in which algorithmic recourse could help address the potential violations of behavioral constraints in the system.

(a) In its *actionable knowledge discovery* formulation, it can contribute to the development of organizational knowledge and reduce the necessity of algorithm-driven selection over time.
(b) It increases the agency of end-users (benefit recipients), proposing actions that could reduce their risk scores, and thus help them decrease the likelihood of subsequent re-examinations.
(c) If provided on an opt-out basis, it may encourage end-users to engage with complaints processes when necessary.
(d) If sufficiently faithful to the model, it may serve as an additional oversight tool, allowing the end-users to identify certain harmful outputs of the model.

Of course, our analysis in this chapter is purely theoretical, and the implementation of algorithmic recourse is bound to bring about challenges and unexpected dynamics in the system. We consider the task of evaluating algorithmic recourse in a controlled manner in Chapter 9 and propose a proof-of-concept method to address its challenges in Chapter 10. For now, we simply observe that algorithmic recourse *may* be a useful safety mechanism, and thus it merits our further consideration in this application.

Finally, we address two important criticisms of our analysis. First, as recognized by [204], the cycle of applying a black-box model and then attempting to interpret tends to be counterproductive, and in high-stakes settings simpler models should be preferred. While we agree with this point, we consider the model that was put in place by the municipality, and there may have been valid reasons to consider a tree-based gradient boosting machine, such as its dominant performance on tabular data [92]. Moreover, we note that even highly-interpretable linear models become difficult to analyze when, e.g., operating on a large number of features, as also recognized by the experts in Chapter 7. Second, there exists a disconnect between the social requirements for technical solutions and the requirements that these solutions can support in practice [2]. In other words, even if algorithmic recourse solutions discussed in Chapter 4 could be adapted to this setting, they are not guaranteed to fully address the challenges of inadequate control. Again, we concede this point but note that algorithmic recourse does not need to fully bridge the socio-technical gap to remain a helpful mechanism. As this document focuses on the value of algorithmic recourse, we have narrowed down our analysis to this solution, but it is likely other interventions could address the inadequate control scenarios that we have identified equally well.

[204]: Rudin (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'

[92]: Grinsztajn, Oyallon, and Varoquaux (2022), 'Why do tree-based models still outperform deep learning on typical tabular data?'

[2]: Ackerman (2000), 'The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility'

Evaluation challenges 9

In this brief chapter, we introduce a conceptual framework for the evaluation of algorithmic recourse interventions in real-world contexts. Our model is a product of contextualizing the findings of the case study in Chapters 5- 8 against the backdrop of research reviewed in Chapter 4. We depict this model in Figure 9.1 and discuss it in Section 9.1. Then, we consider the problem of operationalizing the criteria for algorithmic recourse in Section 9.2.

9.1 Conceptual framework for algorithmic recourse

We identify three components that have influence over the successful implementation of algorithmic recourse mechanisms and, hence, the quality of recommendations that will be issued to the end-users.

Most importantly, algorithmic recourse necessitates a rich understanding of the constraints and the requirements of the application of interest. This includes actors that are involved in the system (Section 6.2), the existing organizational processes (Section 8.1), the potential abuse of the mechanism (e.g., whether opening up the model could negatively affect the control of the process), and the multi-agent dynamics (e.g., whether the end-users provided with recommendations would need to compete for limited resources). Answering these questions is a condition sine qua non for algorithmic recourse in real-world contexts because the other two aspects of its successful applications are ***necessarily* domain-specific**.

Recall our discussion in Section 4.3.1 where we considered the **differences in algorithmic recourse interventions** in education and medicine.

Next, the actionability of recommendations depends on multiple factors, such as the preferences of the recipient, the legal and ethical standards (i.e., some changes cannot be prescribed in good faith), the ability to effectively communicate these recommendations with the end-users, and a variety of contextual requirements that follow from the application. In the literature, these tend to be modeled as constraints over features and their values (Section 4.2.3), but other requirements may relate, for example, to the limited time for the implementation of a recommendation or the amount of support that the organization can provide their customers in the process.

Finally, the application will influence the technical challenges that need to be overcome for the successful implementation of the mechanism. For example, the quality of the data and/or model may be insufficient to generate reasonable recommendations (Section 8.3.1), and thus, an organization needs to focus its initial efforts elsewhere. Moreover, factors such as the regularity of model retraining will influence recommendation invalidation rates (Section 4.2.7).

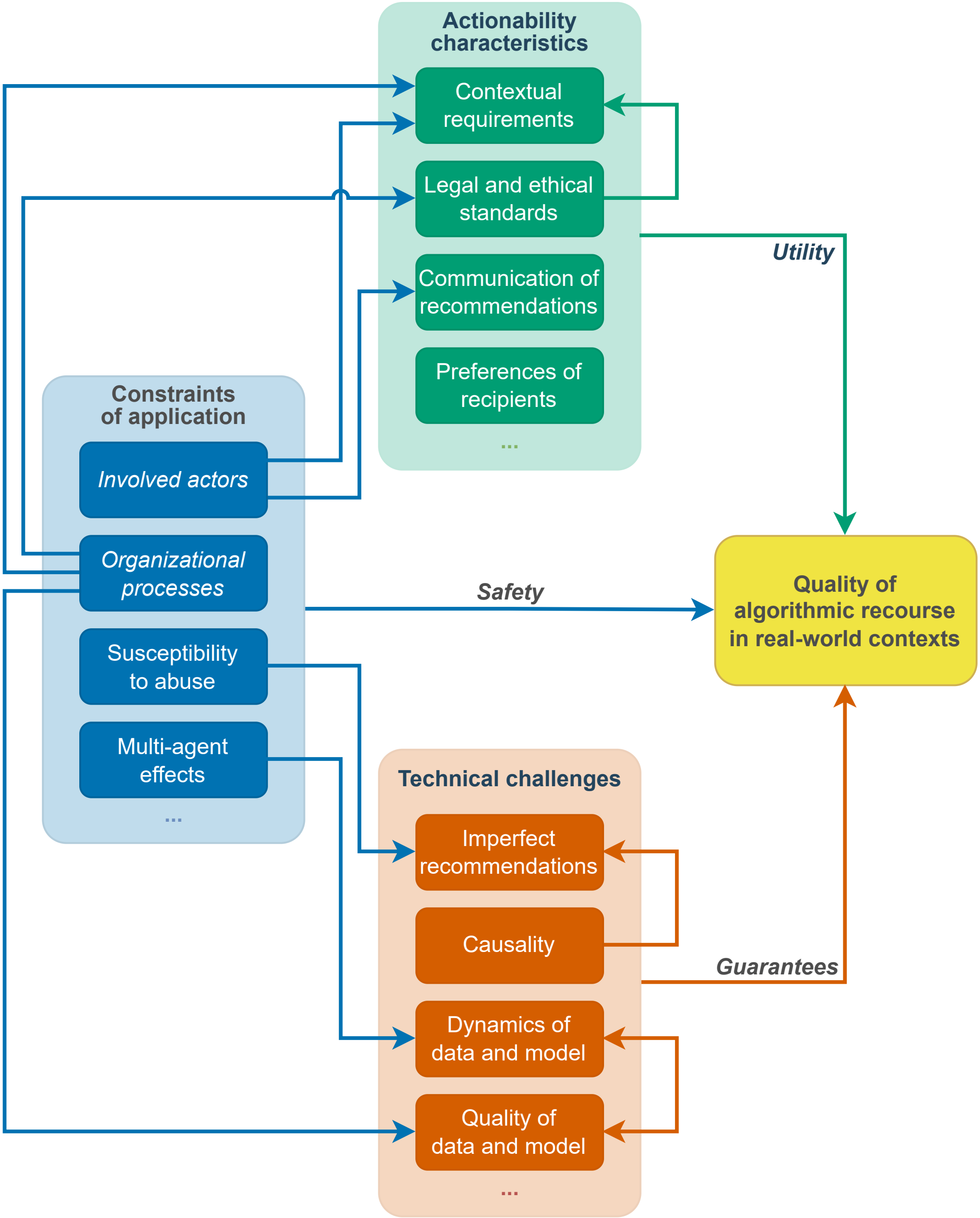
Actionability characteristics
Contextual requirements
Legal and ethical standards
Communication of recommendations
Preferences of recipients
...
Utility
Constraints of application
Involved actors
Organizational processes
Susceptibility to abuse
Multi-agent effects
...
Safety
Quality of algorithmic recourse in real-world contexts
Technical challenges
Imperfect recommendations
Causality
Dynamics of data and model
Quality of data and model
...
Guarantees

Figure 9.1: Conceptual framework for the evaluation of algorithmic recourse interventions in real-world contexts. Arrows represent "depends-on" relationships between components. For example, if the application is highly susceptible to abuse by end-users (meaning that the generated recommendations could be misused), they must be held to particularly high standards. Understanding the constraints imposed by the application is a pre-requisite for the deliberation on the characteristics of actionability or the technical challenges to be overcome, as the latter two components are necessarily domain-specific.

9.2 Measuring the quality of algorithmic recourse

The task of algorithmic recourse evades robust operational metrics precisely because it is domain-specific. At a minimum, three factors will influence the quality of an algorithmic recourse mechanism.

First, the actionability component entails the consideration of the utility of recommendations -— it should be possible for an agent to put the received recommendation into practice and achieve their goals with respect to the system. In practical terms, this means that the system must be able to generate **valid** recommendations across a range of potential constraints on the optimization process. We note that all other common desiderata for algorithmic recourse recommendations should be considered secondary to this requirement.

An AR recommendation is valid if the model assigns the proposed counterfactual instance to the target class.

Second, the capacity of system owners to address the technical challenges of algorithmic recourse relates to the guarantees given to end-users. A recommendation should remain valid long enough that its recipient has a genuine opportunity to implement it. Also, if they (bona fide) exert the expected effort, they should be rewarded with a better outcome. Hence, the system must be able to generate recommendations whose validity extends through a range of potential environmental scenarios (e.g., a set period of time). While this directly relates to the desideratum of "robustness" for CEs (see Section 4.2.6), we note that the validity of recommendations may generally be enforced through non-technical interventions such as "recourse contracts" [64, 74], but these are again context-specific.

Third, the ability of an algorithmic recourse mechanism to respect the constraints of an application requires a safety-oriented evaluation. We emphasize that algorithmic recourse could impact system processes and stakeholders both negatively (e.g., Section 4.2.7) and positively (e.g., Section 8.4), so it is necessary to evaluate it against baselines without such mechanisms. A variety of **performance indicators** may then be applied to measure the "safety component", including the trends in the number of incidents over time, or the time between the first occurrence of a hazard and its detection.

In the Rotterdam case, **these indicators** could correspond to a change in the number of false positive investigations, or the time required to identify that the model exhibits harmful biases.

Given this complexity, it may seem that the study of algorithmic recourse in real-world contexts requires pilot deployments, similar to the `Hired.com` evaluation [164] highlighted in Section 4.2.9. However, as noted for the Rotterdam case in Chapter 6, even when human-in-the-loop studies are actively monitored, they may still produce harm that would otherwise be avoidable. Is all hope lost? Not necessarily. We argue that digital twin solutions [216, 236] may be a promising way to improve the evaluation of algorithmic recourse mechanisms but – to the best of our knowledge – there are no frameworks that would allow decision-makers to explore the dynamics of algorithmic recourse in the context of a particular decision-making process. We address this gap in the next chapter.

Digital twins for public administration 10

In this chapter, we discuss the simulation framework that we have developed as a proof-of-concept instantiation of our evaluation model and applied to the Rotterdam case. First, in Section 10.1, we differentiate our work from the existing benchmarking tools for algorithmic recourse. Next, in Section 10.2, we describe and motivate the design of our framework. Then, in Section 10.3, we demonstrate how our tool could be used to evaluate an algorithmic recourse mechanism in the decision-making setting of our case study. Finally, in Section 10.4, we address the existing limitations.

10.1 Overview of existing solutions

We are aware of three computational frameworks for the generation and benchmarking of CEs and AR recommendations:

- `Counterfactual And Recourse Library` (`CARLA`) is a Python library published in 2021 [178] that supports 13 generators for algorithmic recourse; the recommendations may be evaluated on 9 integrated metrics, including Minkowski distances, success rates, constraint violations, or the generation time. At the time of writing in late October 2024, the `CARLA` GitHub repository has not seen activity since February 2023. Thus, it seems likely that the project has been discontinued.

- `CounterfactualExplanations.jl` (`CE.jl`) is a package that forms part of the *Trustworthy AI in Julia* (*TAIJA*) ecosystem [13]. `CE.jl` has a similar focus to `CARLA` in that it is geared towards the generation and evaluation of (actionable) counterfactual explanations [15]. The package is under active development; as of October 2024, it implements 14 generators and allows users to arbitrarily compose gradient-based generators to test new solutions. `CE.jl` can fit and explain models trained using `Flux`, and further integrates with `PyTorch` and `Torch for R`.

- `AlgorithmicRecourseDynamics.jl` is another package from *TAIJA* that allows for the measurement of dynamics that may arise when algorithmic recourse recommendations are implemented at scale [14]. While the package allows for the evaluation of some important multi-agent characteristics of real-world systems, the simulations remain relatively abstract in that the recommendations are isolated from outside events (decision-making processes), and the agents instantaneously materialize required changes.

Existing frameworks focus on algorithmic recourse *recommendations*. However, as we argued in Chapter 9, exploring the *mechanisms* of algorithmic recourse is necessary to understand its impacts on decision-making systems. We introduce `SimulatedRecourse.jl` as a proof-of-concept solution to facilitate safety-oriented evaluations. Our tool is **available** under the MIT License in the *TAIJA* organization at https://github.com/JuliaTrustworthyAI/SimulatedRecourse.jl.

All experimental studies discussed in this chapter were carried out on the **commit be939a9** version of the code.

10.2 Architecture and design choices

This section focuses on the design and development of our framework. In Section 10.2.1, we motivate the choice of agent-based modeling as the simulation paradigm. Next, in Section 10.2.2, we explain the structure of our codebase. Finally, in Section 10.2.3, we discuss how to configure experiments in the tool.

10.2.1 Motivation for agent-based modeling

Our problem potentially yields itself to two modeling formalisms for dynamic systems: Discrete-Event System Specification (DEVS) or Discrete-Time System Specification (DTSS) [24]. The former places emphasis on the decision-making process because DEVS simulations are driven by events (e.g., "an agent submits an application", "an investigation is triggered by the municipality"); the latter focuses on the behavior of individual agents within this process because at each time step, the model allows all agents to take actions.

If the simulation model triggers fewer than one event per agent per time step, DEVS simulations are guaranteed to be more efficient. Thus, the choice of a modeling formalism may have non-negligible impacts on the quality of the final simulation framework. We decide to follow the DTSS paradigm – specifically by applying agent-based modeling (ABM) methods – for three reasons:

1. We envision the time steps of the models to represent relatively large periods of time, at the order of one calendar month. A higher level of granularity is simply not needed. For example, implementing an algorithmic recourse recommendation is unlikely to happen overnight. Thus, we expect a DTSS model to be competitive with a DEVS model. In any case, we allow for the reconfiguration of the duration of a time step.
2. We aim to develop a simulation framework that improves on the state-of-the-art in terms of grounding in real-world contexts. Hence, it should be possible to meaningfully enforce heterogeneity between individual agents. For example, they may differ in feature values, cost functions, or their "flow" through the stages of the decision-making process. We argue that this goal aligns more closely with the characteristics of DTSS rather than DEVS simulation models.

3. Following our recommendation in Section 4.3.2, we aspire to build a practitioner-friendly artifact. ABMs are seen as the de facto standard method for simulations in policy analysis [79, 142], and they have been adopted in a variety of related fields including computational social sciences [e.g., 85, 148], economics [e.g., 22, 106], public health [e.g., 150, 239], or urbanism [e.g., 44, 207]. Thus, an ABM tool should be more intuitive for practitioners who may want to use it in the future.

Having established the advantages of an agent-based model, we note that our use case does not require (extensive) interactions between the agents. Rather, we expect to observe the emergence of certain dynamics through the interactions of agents with their environment. Implementing the decision-making process as interactions of customers and employees of an organization could have had some advantages – for example, it would have allowed us to better diagnose **the bottlenecks** that an algorithmic recourse mechanism could introduce – this is not necessarily the main goal of our simulations. Instead, the customers interact directly with the organization .

We can still observe the changes in **the flows of agents** throughout the process, but at a higher level of abstraction.

This requires us to introduce some features characteristic of discrete-event models into the agent-based model. In particular, we split the decision-making process into **`SystemStages`** and **`AgentStages`**; the agents passively flow through the former and actively engage in the latter. In turn, a decision-making process becomes a chain of action-reaction events, separating the concerns of the agents and their environment. We elaborate on this topic in Section 10.2.2.

10.2.2 Structure and integration of other libraries

An important goal for our tool is its generalizability to different decision-making processes to allow for the evaluation of algorithmic recourse mechanisms in a variety of scenarios. While we develop the tool with the Rotterdam case in mind, wherever possible we decouple the overarching logic of the framework from the specific logic of the case study. When discussing the architecture of the tool, we highlight the design choices taken to facilitate this goal.

`SimulatedRecourse.jl` is built using the `Agents.jl` framework for agent-based modeling in Julia [53]. Simulations in `Agents.jl` can be as simple as defining four components: (1) a struct representing an agent that populates the simulation, (2) a type of space where the agents operate, (3) the stepping functions for the agents and the model, and (4) the types of data to be collected from the simulation.

Agent

We make use of a single type of agent – `Customer` (Listing 10.1) – that represents an end-user (e.g., resident) interacting with the decision-making process (e.g., municipal social assistance). Each `Customer` is defined by a set of `features`, their `status` in the simulation (e.g., "accepted" or "rejected"), and the `outcome` of the decision-making process (e.g., the exact amount of received assistance). Notably, the `Customers` may be defined by multiple sets of `features` **in different stages of the process**. We decide to store the "main" features of the agent – ones that are most relevant for the decision-making process – at the agent level, while additional features can be easily stored on the model level in the form of a `DataFrame`.

```
@agent struct Customer(GraphAgent)
    features::DataFrameRow
    status::Symbol
    outcome::Int
    options::Vector{Symbol}
    properties::Dict{Symbol, Any}
end
```

Listing 10.1: Definition of `Customer`.

In the Rotterdam case, the decision about social assistance depends on 10 features as described in Section 5.2, but the nomination for investigations was carried out by a model trained on (mostly non-overlapping) 315 features.

Additionally, each agent stores a `Vector` of `options` and a `Dict` of `properties`. The former is used to pass parameters that influence the behavior of the simulation between stages (e.g., if an agent was randomly selected for investigation, they will not be able to ask for an algorithmic recourse recommendation); the latter is used to store additional types of information that are collected during the execution of the simulation (to be configured by the user).

Space

`Agents.jl` allows the agents to operate in a discrete space ($\mathbb{I}^n$), continuous space ($\mathbb{R}^n$), on a map, or on a graph. The last of these options is a natural fit for our task as we can represent an arbitrary decision-making process as a set of stages (nodes) connected by directed edges that describe the processing of `Customer` agents. For example, (`Idle`) ⟶ (`Application`) ⟶ (`Decision`), corresponds to the path of an agent that exists outside of the boundary of the simulation in (`Idle`), becomes active and applies for assistance in (`Application`), and their status is evaluated by the organization in (`Decision`). Although Julia does not support traditional inheritance, we decide to organize all stages in a logical hierarchy, presented as a UML diagram in Figure 10.1. As already described, we divide the stages into those where the agents are active (`AgentStage`) and those where the agents are passive (`SystemStage`). Additionally, we define `Idle` as an instantiation of an `EnvironmentStage` that stores agents outside of the simulation boundary, i.e., ones that are not actively participating in the decision-making process. For each individual experiment, the user may reconfigure the process graph stored as a `.txt` file. For example, they may remove a stage or change the default transition probabilities between stages.

We further separate the existence of an agent in a stage from the logic of that stage, as the latter may depend on a simulation experiment. Thus, the user is expected to implement or extend the logic, following a common API as defined in Listing 10.2. The `process` functions are defined for every stage `X` allowing us to make use of Julia's multiple dispatch, and hence the state of the simulation can be updated in a simple loop that we will describe in the next section.

```
function process(
    current_stage::X,
    agent::Customer,
    sim::ABM,
    stage_logic::Function)

    stage_logic(current_stage,
                agent, sim)

    return
end
```

Listing 10.2: Definition of `process`.

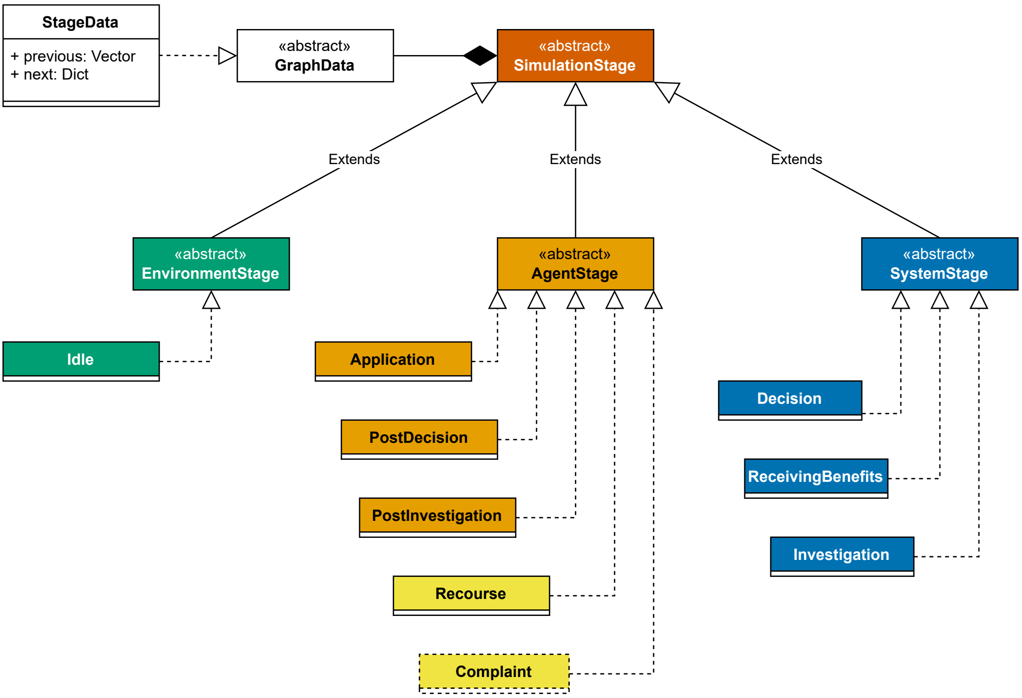

Figure 10.1: Logical organization of the decision-making process stage.
We implement the logic for eight stages marked in solid lines, but further stages (e.g., `Complaint`) can be readily added. Each concrete stage stores a `StageData` object that probabilistically describes transitions in and out of that stage.

Stepping functions

Our proof-of-concept tool implements only a stepping function for the agents, presented in Listing 10.3. It is called for every agent at every time step to update this agent and process them in the current stage. This Listing also shows how our framework is split into two submodules: `Case` and `Process` to facilitate generalizability. In particular, the `update_agent!` function can be used to modify the features of an agent so that, e.g., their eligibility for assistance may change over time – in our instantiation in Section 10.2.3 we use this function to model some basic macroeconomic phenomena.

```
function agent_step!(
    agent::Customer,
    sim::ABM)

    Case.update_agent!(agent,sim)

    stage = sim.stages[agent.pos]
    Process.process(stage,agent,sim)
end
```

Listing 10.3: Agent stepping function.

We have not implemented the stepping function for the model in the first iteration of the tool due to its lower priority. Nonetheless, our simulations could definitely benefit from model-wide updates. For instance, `model_step!` could be employed to periodically retrain the model and explore the dynamics of algorithmic recourse, akin to the goals of `AlgorithmicRecourseDynamics.jl` of [14].

Data collection

Integration with `Agents.jl` allows the users to freely define agent-level and simulation-level properties that should be collected at runtime. In many cases, this will only require declaring a property in the initialization method of the simulation – `initialize(·)` – and adding a few lines of code in the logic of relevant stages. As one example, we may be interested in the number of times that each agent is nominated for investigation; we can define an agent-level counter as `agent.properties[:experienced_investigations] = 0` and then update this value whenever the agent enters the `Investigation` stage. Then, `Agents.jl` returns a `DataFrame` with the value of the counter per agent per time step, to be aggregated by the user.

Algorithmic recourse recommendations

As a final point in this brief overview of `SimulatedRecourse.jl`, we discuss the integration with `CounterfactualExplanations.jl` to generate algorithmic recourse recommendations. Depending on the configuration of the decision-making process, the users may simulate a system where the nominations are (1) completely random, (2) additionally rely on a model, and (3) further allow for algorithmic recourse. In cases (2) and (3), we rely on the functionalities of `CounterfactualExplanations.jl` – we use it to train the investigation models and to generate counterfactuals for agents. If a `Customer` moves to the `Recourse` stage, a recommendation will be generated and stored. Then, the agent will probabilistically implement the recommendation; in every cycle, they have a $p_i \in [0,1]$ chance of modifying the value of feature i by one unit. Thus, we implement a very simple measure of difficulty, but propose a more robust way to simulate the implementation of recommendations in Section 10.4.

10.2.3 Configuration of the experiments

`SimulatedRecourse.jl` is principally intended to allow for the exploration of algorithmic recourse mechanisms, so it was developed to be highly configurable. We provide an overview of the main configuration capabilities in the following paragraphs.

Configuring process graphs
The user may arbitrarily configure the decision-making process graph by adding and removing stages, adding and removing edges between stages, and changing the transition probabilities. We present an example configuration for the process graph that corresponds to our case study in Listing 10.4: on the first line we declare the total number of stages, the following fifteen lines describe edges in the format `from to probability`, and the last eight lines assign human-readable names to the stages. The transition probabilities may change at runtime. For example, if an agent is rejected in the **`Investigation`** stage, they will not be allowed to directly move back to the **`ReceivingBenefits`** stage, so the transition `7 5 0.00` would be in any case excluded, with other probabilities re-normalized.

```
8
1 1 0.05
1 2 0.95
2 3 1.00
3 4 1.00
4 1 0.05
4 5 0.95
5 1 0.05
5 5 0.90
5 6 0.05
6 7 1.00
7 1 0.00
7 5 0.00
7 8 1.00
8 1 0.05
8 5 0.95
1 Idle
2 Application
3 Decision
4 PostDecision
5 ReceivingBenefits
6 Investigation
7 PostInvestigation
8 Recourse
```

Listing 10.4: Process configuration.

Configuring agent data
The user may also modify the data used in the process by specifying paths to the "decision data", the "test data", and potentially the "train data". We first explain the decision data since it is slightly more involved – it refers to the dataset used to make decisions about the eligibility for assistance (Section 5.2). In principle, our framework can support any dataset although modifying the decision data may require changing the logic of the **`Decision`** stage. The latter two types of data are used by the investigation model; modifying them is as simple as specifying the paths to two datasets described by the same features. For example, a decision-maker may decide to evaluate how a particular set of features influences the process or the ability of **`Customers`** to implement recommendations. If provided with a training dataset, our tool automatically trains a model and runs a simulation on the dataset. To enable AR, the user also needs to provide a configuration file in the `.json` format that includes the indices of continuous and categorical features, the mutability and the domain constraints, and the suggested actionability of features.

Configuring algorithmic recourse recommendations
Finally, the user may specify in the configuration file any machine learning model and any algorithmic recourse generator supported by `CounterfactualExplanations.jl`. It may be that some generators perform better in a specific domain, e.g., the `ClaPROAR` algorithm was introduced to mitigate endogenous domain and model shifts [14], so if a simulation involves model retraining, we would expect that it produces different results from simpler generators. As long as `CE.jl` supports a specific combination of the model and the generator, `SimulatedRecourse.jl` will also support it.

10.3 End-to-end evaluation

We present our tool by simulating the **(simplified)** decision-making process of the case study using the data provided by Lighthouse Reports. In this section, we describe all steps required to prepare an experiment in `SimulatedRecourse.jl` but we note that many of them need to be carried out only once, which makes follow-up experiments much simpler. After configuring a baseline simulation study in Section 10.3.1, we evaluate the impacts of an algorithmic recourse mechanism on the system in Section 10.3.2.

Given that we make a number of strong assumptions, we can best describe our experiments as ***heavily inspired* by the Rotterdam case**.

10.3.1 Instantiation of the case study

1. Configuring the process graph
We configure the process graph to reflect the operating process from Chapter 8 as closely as possible; it is presented in Figure 10.2 with the transition probabilities omitted. All agents start in the **`Idle`** stage, where they also return whenever removed from the social assistance process. Of particular note is the **`PostInvestigation`** stage where the agents may decide to react to a negative outcome of an investigation by pursuing algorithmic recourse. A similar intervention could be employed in the **`PostDecision`** stage; however, the assistance standard for a **`Customer`** is calculated by a rule-based system, so standard algorithmic recourse solutions do not apply. This is exactly the problem addressed in Chapter 11.

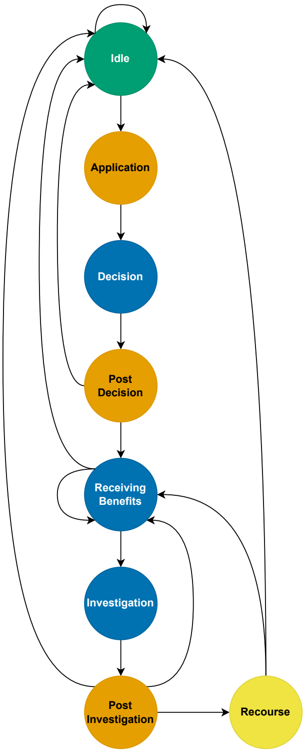

Figure 10.2: Process graph.

2. Defining the case-specific logic
This is the most complex step of the case study setup, as we need to define the logic for each of the stages from the previous paragraph:

(**`Idle`**) Apply for social assistance with certain probability. If a recommendation is available (i.e., the agent was previously removed from the process), attempt to implement it.

(**`Application`**) Move to the next stage.

(**`Decision`**) Apply logic from the Participatiewet to calculate the amount of assistance to be awarded to the agent.

(**`PostDecision`**) Move to the next stage depending on the calculated amount of assistance. If it is larger than 0, move to **`ReceivingBenefits`** and otherwise move back to **`Idle`**.

(**`ReceivingBenefits`**) With certain probability nominate the agent for investigation. If a model is available and outputs a high risk score, also nominate the agent for investigation.

(**`Investigation`**) Apply logic from the Participatiewet to re-calculate the assistance standard for the agent.

(`PostInvestigation`) Move back to `Idle` or request `Recourse` if the outcome of an investigation is negative, otherwise move back to `ReceivingBenefits`. If the status of an agent did not change but an investigation was started because of a model, the agent may still request `Recourse` to reduce their risk score.

(`Recourse`) Generate an algorithmic recourse recommendation for the agent to help them reduce their future risk score. It will be implemented outside of the `Recourse` stage.

3. Preparing the decision data

Unfortunately, the data provided by Lighthouse Reports does not allow us to calculate the amount of assistance awarded to agents as it does not contain the features relevant to the decision. Therefore, we need to generate our own dataset for the agents. We make use of the CBS statistics [42] to estimate the distributions of 18 features that are combined to produce the 10 features needed for the assessment of eligibility. We make our best effort to provide a realistic initialization for the agents. Nonetheless, this initialization is far from perfect: we need to treat all features as independent, and we are unable to fit the exact distributions for these features.

As one example, we explain the process of estimating the total income of an agent. First, we find that 44.2% of inhabitants aged 15+ are married and that there are 1.12 million unmarried couples in the Netherlands. We also find that there are roughly 15 million people in the Netherlands aged 15 or older. This allows us to approximate the likelihood that a person has a partner. Second, we find that 75% of inhabitants are in the labor force and that for working people the median income is €39100 per year, which is ≈€3200 per month. We draw the incomes from a Gamma(4, 800) distribution – it starts at 0 with a mean of exactly 3200 and it is skewed towards lower numbers. Finally, we can sample the income of an agent by combining the two distributions. If they have a partner, we also sample their income and combine them under a feature `total_income`. The pseudocode is given in Listing 10.5. We additionally decide to scale down some features – multiply them by a fraction – until we achieve a reasonable number of agents eligible for social assistance (≈20% when the simulation begins).

```
married = sample(
    [true,false],[0.442,0.558]
    )

cohabitation = married ?
    false : sample(
    [true,false],[0.264,0.736]
    )

has_partner = married || cohabitation

own_income = (
        sample(
        [true,false],[0.75,0.25]
        ) ?
        rand(Gamma(4, 800)) : 0
    ) * 0.25

partner_income = (
        has_partner ? (
            sample(
            [true,false],[0.75,0.25]
            ) ?
            rand(Gamma(4, 800)) : 0
        ) : 0
    ) * 0.25

total_income = own_income +
      partner_income
```

Listing 10.5: Estimation of income.

Decision-makers likely have access to the features of the actual customers, so they should be able to skip this step by (re)sampling real-world data. This may also improve the quality of the model and align it more closely with the goals of digital twin systems.

4. Preparing the investigation data and the model
Although Lighthouse Reports published the parameters of the gradient boosting machine used by Rotterdam W&I department, the model was trained in R and it cannot be readily explained by `CE.jl` because its outcomes are continuous in range [0, 1] (where 1 represents the highest possible risk score) and regression models are not currently supported by `CE.jl`. Instead, we decide to train a surrogate model for a classification task. First, we generate further 130000 synthetic samples using the Gaussian copula model trained by Lighthouse Reports, and then binarize the risk scores on a threshold used by Rotterdam (roughly 0.7). Then we use the new samples to train a `NeuroTreeModel` – a form of a differentiable decision tree [113] – on the resulting classification task. If all 315 features are used, the surrogate model achieves accuracy of ≈95% (i.e., the predictions of the two models agree in 95% of the cases) whereas guessing the majority class would lead to accuracy of ≈85% in expectation. We use the synthetic data initially generated by Lighthouse Reports as test data in the simulations.

5. Defining the `update_agent!` function
If the features of agents were static, then the amount of assistance calculated in the **`Investigation`** stage would always agree with the amount calculated in the **`Decision`** stage, which would negatively impact our ability to analyze **certain aspects of algorithmic recourse**. Thus, we model some external phenomena in the **`update_agent!`** function, again making use of the CBS data [42]. We allow for eight features to change every "year": `children`, `current_age`, `has_partner`, `other_adults_in_household`, `first_home_equity`, `own_income`, `total_assets`, `total_income`. For example, the agent may lose a partner or move in with a new partner, both of which will further impact their total assets and total income. Of course, all of these changes fall outside the simulation boundary, so we do not attempt (and do not need) to be particularly exhaustive.

For instance, recommendations would never be issued to agents rejected after a reinvestigation in this setting.

6. Implementing an algorithmic recourse mechanism
To support the "realistic" implementation of recommendations, we need to define three types of constraints for each feature. First, we infer the domain constraints from the data by looking at the minimum and maximum values of features found in the test set. Second, to decide on the mutability constraints we reason about changes that can be expected from the agents. Third, we need to estimate the difficulty of changing each feature and we want to at least remain consistent in the ordering of features by difficulty. To that end, we decide to follow a **pairwise outranking procedure**. If all 315 features were considered, this would lead to almost 50,000 comparisons. Thus, we select **26 features from the original dataset**, 16 of which may be considered mutable. We carry out 120 pairwise comparisons to establish a ranking of features ordered by "inverse actionability" and convert it to probabilistic difficulty scores.

Pairwise outranking is a multi-criteria decision-making method to establish a global ranking by comparing all possible pairs of options. Then, the ranking is constructed by looking at the number of times each option is selected.

We apply the insights of Algorithm Audit [10] to select the features (mostly) acceptable for risk profiling. Among others, we remove all features that are subjective or seem irrelevant to the task. We also remove all features with a relevance score of less than 5. Unfortunately, this degrades the performance of our surrogate model to ≈91%.

10.3.2 Experiments and results

Having instantiated the case study, we run three experiments using `SimulatedRecourse.jl`. First, a system where **`Customers`** can *only* be randomly nominated for investigations (no model is in use) is the baseline. Then, we add a model as an additional selection method. Finally, we enable the agents to request and implement algorithmic recourse recommendations over time.

Each study is executed with a random sample of 2500 agents from the synthetic dataset produced by Lighthouse Reports (20% of the complete dataset), simulating their actions **over 22 "years"** (264 time steps). All simulations in `SimulatedRecourse.jl` are fully reproducible, so the agents active in run *i* of all three studies are exactly the same, but their actions diverge when the pseudorandom number generators get out of sync due to different processing of agents in **`ReceivingBenefits`** and **`Recourse`** stages, depending on the study. As our experiments constitute a "baseline measurement", we also decide to make use of the simple Wachter *et al.* generator [261]. For the study with algorithmic recourse mechanism enabled, we configure the process graph as in Listing 10.4, i.e., **all agents with a high risk score** request a recommendation; in the other two studies we simply remove the **`Recourse`** stage and set **`7 1 0.85`** and `7 5 0.15` as default transition probabilities. All other settings that can be enforced in the tool follow from the discussions in Chapters 5 and 6, e.g., **`Customers`** can be nominated for investigation every 24 cycles, and the **probability of random nomination** is 0.175.

As all **`Customers`** start in the **`Idle`** state, **it takes slightly over two "years"** (specifically, 28 cycles) for the first investigation to be triggered. We use the first two years as a burn-in and focus on the following 20 years.

This reflects a setting where **recommendations are given on an opt-out basis**.

This decodes as: move from stage 7 (**`PostInvestigation`**) to stage 1 (**`Idle`**) with probability p=0.85 and to stage 5 (**`ReceivingBenefits`**) with p=0.15, but this may be adjusted at runtime.

The logic of the **`ReceivingBenefits`** is configured to first evaluate an agent using the model, so its use does not have a significant impact on the **overall probability of nomination**.

We evaluate the experimental studies on six criteria:

- *Total investigations*: the number of investigations due to any selection method carried out in the simulation period.
- *Model investigations*: the number of investigations due to model selection carried out in the simulation period.
- *Total investigations per agent*: the mean number of investigations due to any selection method experienced by agents who have undergone at least one investigation.
- *Model investigations per agent*: the mean number of investigations due to model selection experienced by agents who have undergone at least one model investigation.
- *Distribution of agents over stages*: the proportion of agents in each stage at the end of the simulation period.
- *Wall time*: the elapsed real runtime of a single-threaded simulation on an Intel Core i7-8750H CPU @ 2.2 GHz.

Table 10.1: Selected results for three simulation studies, averaged over 100 runs. Standard deviation is provided in parentheses. *Emphasized* numbers indicate that model investigations can repeatedly target the same individuals when AR is not available.

Study	Setting		Total investig. (occurrences)	Model investig. (occurrences)	Total investig. per agent (occurrences) ↓	Model investig. per agent (occurrences) ↓	Wall time (ms) ↓
	Model	Recourse					
1. Baseline	✗	✗	2392.41 (64.39)	–	**2.10 (0.04)**	–	**2713.93 (91.33)**
2. Model	✓	✗	2513.95 (65.57)	467.75 (26.63)	*2.18 (0.03)*	*2.61 (0.11)*	7859.41 (312.92)
3. Recourse	✓	✓	2469.43 (61.17)	211.67 (12.23)	**2.14 (0.03)**	**1.19 (0.03)**	10384.22 (444.27)

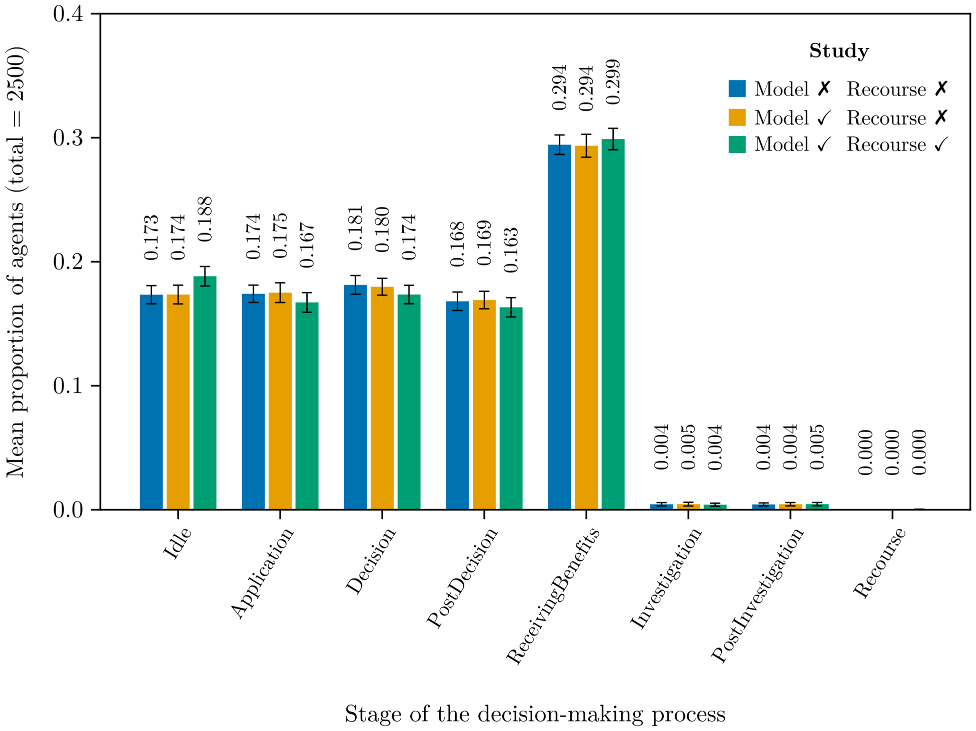

Figure 10.3: Distribution of agents over the stages of the decision-making process at the end of three simulation studies ($t = 264$). All results are averaged over 100 runs; we do not observe statistically significant shifts between stages when AR is introduced.

Two findings stand out in the results. First, as shown in Table 10.1, the use of model-driven nominations introduces a harmful dynamic in the system where **Customers** who have been nominated at least once by the model are more likely to be nominated again than in the baseline. This is exactly the behavior observed in Rotterdam as reported in Section 6.1. As we need to treat the predictions of the original model as the ground truth labels for the surrogate model, we effectively work with a completely different and likely more "accurate" model than that of Rotterdam. This is important because we would expect **the magnitude of this dynamic** to depend on the quality of the model. For example, if a certain "quota" for model-driven nomination is to be filled and the model is biased toward certain agents, then the agents that *experience at least one* model-driven nomination will likely *experience many more* nominations, on average, than benefit recipients who are classified as low-risk.

We also explored how **the maximum allowable frequency of investigations influences the effect** – if we allow for investigations once every 12 months, the average number of investigations per agent grows 2.2 times, but the average number of model investigations per agent increases as much as 2.6 times.

Our model is static, but if it was periodically re-trained such that it can learn about the outcomes of the previous investigations, it may still repeatedly target some individuals due to its **bounded accuracy**. Naturally, introducing an AR mechanism mitigates this effect because individuals who have been nominated by the model may be able to reduce their risk scores for subsequent runs of the model, and so the number of model investigations per agent becomes smaller than the number of total investigations per agent.

Even if the municipality was able to develop a model with **100% accuracy**, the features of customers are not static.

Second, in Figure 10.3, we observe no significant differences in the number of agents at each stage of the decision-making process. That the proportion of agents in **ReceivingBenefits** stage is similar, suggests that agents who have received a recommendation are not "favored" by the system. This makes sense because the decision logic and the investigation model rely on different sets of features. We also note that in the "Recourse study", slightly more agents are in **Idle** and their proportion is lower in the following three stages – this happens because agents spend several cycles implementing the recommendations before trying to rejoin the process.

10.4 Present shortcomings and future development

In this section, we address the main shortcoming of our case study – the limited verification and validation of the experimental models (Section 10.4.1) – and outline several improvements for our tool that could improve its utility for the evaluation of algorithmic recourse across different real-world applications (Section 10.4.2).

10.4.1 Verification and validation

Verification and validation are essential processes for establishing the credibility of a simulation model [211]. All simulation models should agree with the underlying conceptual model (verification;

[211]: Schlesinger (1979), 'Terminology for model credibility'

e.g., the simulation should be tested for bugs) and its actualization in the real-world (validation; e.g., the parameters of the simulation should correspond to the parameters of the target system).

The conceptual model for the decision-making process of *bijstand* has been thoroughly analyzed in our case study, though we implement it with some simplifications. Evaluating whether our implementation is bug-free is a major challenge because we cannot readily test it automatically. Instead, we decide to add a logging system, run a small-scale model, and manually verify the logs to check whether (1) the agents move between the stages as expected, and (2) the main parts of the logic are implemented correctly. For instance, we check the nomination for investigations and the fulfillment of algorithmic recourse recommendations by agents. We believe this approach is sufficient to verify the simulation model with respect to its main requirements, but it certainly does not eliminate the risk of minor bugs influencing the outcomes of the simulations.

Regarding validation, we note that even though our end-to-end evaluation of the tool is grounded in the setting of W&I Rotterdam, the simulations are bound to differ from the real-world system in significant ways. Some of these relate to the limitations of our tool, but most of the differences stem from the lack of suitable data. We highlighted the assumptions we had to concede in Section 10.3.1: our instantiation of the process graph is simplified, our model is only a surrogate of the original model (and trained for a different task), or our data is fully-synthetic and likely an insufficient reflection of the real-world distributions. Wherever possible, we still make use of real-world parameters to simulate the process as accurately as possible, but this does not offset many assumptions we need to make. Nonetheless, we believe that our experiments remain valuable even if they remain *exploratory* rather than *explanatory*.

While thorough validation of the framework would be beneficial, it may not be strictly necessary in our use case. By and large, our tool is a proof-of-concept solution to demonstrate the evaluation of algorithmic recourse mechanisms in real-world contexts, and it can be treated as a step towards the development of more robust solutions that can actually rely on real-world data. This has been recognized as a scenario where the burden of validation is greatly reduced [88]. Moreover, through our experiments, we show that algorithmic recourse *can* reduce risks related to model-driven selection and *does not need to* impact the dynamics of the decision-making system. In other words, our end-to-end evaluation retains value as a computational "thought experiment", and the limited validation does not detract from this purpose [153]. Finally, because we allow for highly configurable experiments, users of `SimulatedRecourse.jl` may test the decision-making processes in a variety of scenarios – this exploratory nature of agent-based modeling may be insightful even if it does not reflect a specific real-world system [217].

[88]: Graebner (2018), 'How to Relate Models to Reality? An Epistemological Framework for the Validation and Verification of Computational Models'

[153]: Mayo-Wilson and Zollman (2021), 'The computational philosophy: simulation as a core philosophical method'

[217]: Šešelja (2021), 'Exploring Scientific Inquiry via Agent-Based Modelling'

10.4.2 Improving `SimulatedRecourse.jl`

`SimulatedRecourse.jl` can be extended in a few important ways. First, most importantly, it would benefit from a higher degree of generalizability. As established in our conceptual framework in Chapter 9, the evaluation of algorithmic recourse must account for the characteristics of a specific decision-making process, logic (i.e., procedures) being one of them. Unfortunately, the current tool would require major re-development if it was to be applied in a different setting, which limits its utility for decision-makers. One solution could be to re-design the tool such that it supports reusable "logic blocks" (akin to automation platforms, e.g., [202] or [221]), reducing the amount of required code duplication. Second, a similar modeling approach may be used to explore other safety interventions and their interactions. We highlight an example in Figure 10.1 with the **`Complaint`** stage, which exists – in several forms – in Rotterdam but has been omitted from our simulations. Third, supporting model-level updates (e.g., re-training) could ground the simulations more strongly in reality and allow for the evaluation of new types of dynamics, potentially even subsuming `AlgorithmicRecourseDynamics.jl` [14]. Finally, as explained in Section 10.2, the implementation of recommendations remains naive in the current version of the tool because the difficulty of affecting features is independent of agents. In our opinion, defining agent-level cost functions would not meaningfully improve the quality of the simulations. Instead, we propose to borrow a simpler approach from the domain of automated negotiation and endow agents with "archetypical " behavioral strategies to affect the predefined actionability values (see, for instance, [141]).

[202]: Rockwell Automation (2024), *Arena Simulation Software*

[221]: Simio LLC (2024), *The Simio Discrete Event Simulation Platform*

[141]: Lin, Kraus, Baarslag, Tykhonov, Hindriks, and Jonker (2014), 'Genius: An integrated environment for supporting the design of generic automated negotiators'

Algorithmic recourse in expert systems 11

As we learned in Chapter 7, public administration tends to rely on hand-crafted "if-then" business rules rather than machine learning models. These rules follow directly from legal acts and produce a form of an expert system, sometimes also referred to as a rule-based system. Expert systems have a long tradition in the (Dutch) government and have been applied to, e.g., *"**grant benefits**, issue licenses or automatically collect traffic fines"* but, despite their widespread acceptance, they may still be *"very complex due to many variables and rules"* [191]. This was also recognized by one of the experts who noted that even algorithms relying on simple rules can **produce harm** or become prohibitively difficult to understand, giving the example of tax systems. As such, we believe there are important reasons to consider algorithmic recourse in expert systems, but in our survey in Chapter 4 we have not identified any previous attempts to generate actionable recommendations in this setting.

This **use case** directly relates to our case study. We present the rules used to decide on the eligibility and the amount of for *bijstand* in Figure 11.1.

An interesting and potent example is the Australian Robodebt, a rule-based system that issued over 430,000 incorrect welfare debt notices, leading to a court settlement of ≈€1.1 billion [68].

In this chapter, we propose a simple and provably-correct algorithm to generate algorithmic recourses in rule-based systems that rely on hand-crafted rules. We start with a small background on expert systems in Section 11.1. Next, in Section 11.2, we define and explain the algorithm, and in Section 11.3, we prove its correctness. Finally, Section 11.4 discusses some domain-specific practical considerations.

11.1 Background on expert systems

Expert systems are among the earliest forms of artificial intelligence; they were an important research direction on algorithmic decision-making in the 1970s and 1980s [55, 269], but have been largely superseded by "learning" algorithms since then. A typical expert system consists of a knowledge base – a set of domain facts – and an **inference engine** used to derive new knowledge from the existing facts in an automated manner [20]. Notably, expert systems have attracted much work on explainability because their decisions can be **readily traced** [e.g., 125, 235, 275]. As already emphasized, we have not identified any methods for algorithmic recourse in expert systems, although there exist several approaches to the generation of (counterfactual) explanations in related settings. These include [231] that focuses on explanations for *"pretrained decision trees and fuzzy rule-based classifiers"*, [107] that proposes a contrastive explanation engine for smart environment systems, [280] that defines a dialogue explainability framework for rule-based systems acting on proof trees, or [31] that produces counterfactual explanations for surrogate decision-tree models. While the above solutions target related challenges, they remain distinct from the problem of generating actionable recommendations for **hand-crafted "if-then" trees**.

In the setting of interest, this **inference engine** is as simple as an algorithm that scans through the tree of if-then rules until it finds a leaf node (outcome).

For the same reason, **"if-then" rules** are sometimes employed as tools for classification and explainability, e.g., as Anchors [198], Bayesian Rule Lists [137], Local Rule-based Explanations [94], or Actionable Recourse Summaries [193].

Arguably, **this setting** is less complex than any of its related problems, so it can benefit from a simpler approach that relies on basic tree operations.

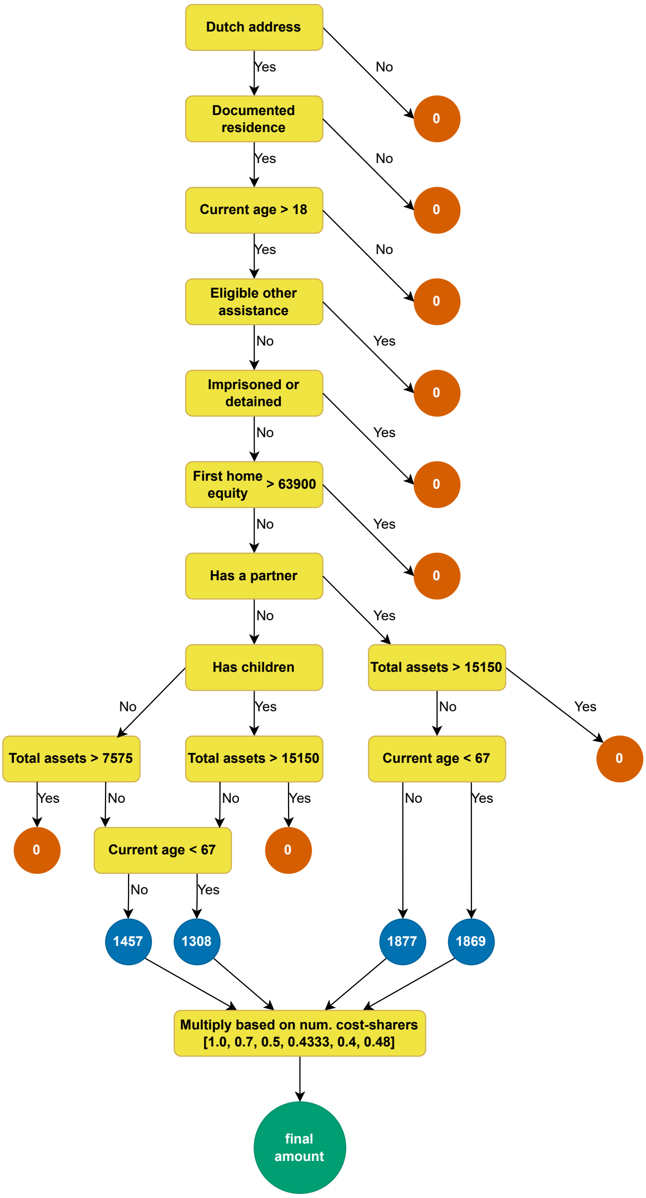

Figure 11.1: Decision tree for the assessment of eligibility for *bijstand*.
Rectangular nodes represent if-then rules and circular nodes describe the (intermediate) outcomes, i.e., the amount of assistance. We slightly simplify the tree for the sake of clarity. In reality, the “current age < 67” nodes depend on the birth year of the applicant, and the final decision node (“multiply based on num. cost-sharers”) should be represented as a set of nodes with atomic conditions.

11.2 Proposed algorithm

Our primary goal is to address a challenge that we have identified with respect to algorithmic recourse in the context of our case study. As such, we do not attempt to guarantee the optimality of the following procedure (in that there likely exists an equivalent algorithm with higher computational efficiency) because rule-based systems/decision trees derived from legislation are bound to be relatively small in size. Nonetheless, we can guarantee that our algorithm is totally correct (in that it can generate algorithmic recourse recommendations for any data instance). While the rule-based system used to assess the eligibility for *bijstand* (Figure 11.1) is *very* simple, meaning that most people will be able to decide what needs to change about an instance to overturn a negative decision, the same algorithm can be applied on much more complex "if-then trees". Anyhow, even the simplest rule-based systems may have underlying implementation errors, so methods to generate CEs or AR recommendations retain value in such settings.

Algorithm 11.1 Algorithmic recourse in expert systems

```
Require:
        $T$    the expert system given as a tree of if-then rules
        $s$    the rejected (negatively classified) sample
        $A$    the set of accepting (positive) leaves

procedure GenerateRecourse($T, s, A$)
    $T \leftarrow$ ConvertToAtomicTree($T$)
    $costs \leftarrow [0, \dots, 0]$
    $paths \leftarrow [[\,], \dots, [\,]]$

    for $accept \in A$ do
        $current \leftarrow$ FindClassifyingLeaf($T, s$)

        while $current \neq accept$ do
            $ancestor \leftarrow$ FindLCAncestor($accept, current$)
            $s, intervention\_cost \leftarrow$ Intervene($s, ancestor$)

            Append($paths[accept], ancestor$)
            Add($costs[accept], intervention\_cost$)

            $current \leftarrow$ FindClassifyingLeaf($T, s$)
        end while

    end for

    return $paths[$Min($costs$)$]$, Min($costs$)
end procedure
```

Algorithm 11.1 describes a way to generate AR recommendations in expert systems. We prove its correctness in Section 11.3, but first, we explain the pseudocode. We require three inputs:

- The expert system encoded as a tree of (hand-crafted) if-then rules, such as the one shown in Figure 11.1.
- A sample that is negatively classified by the expert system (i.e., a sample whose classification may be improved according to some measure of quality or preference).
- The set of all leaves corresponding to positive outcomes.

We assume that T may include non-atomic Boolean conditions. Thus, we convert it to an equivalent form where every internal node represents a single **atomic condition**. We also define variables to keep track of the recourse paths and their associated costs.

In other words, our goal is to convert a generic k-ary tree into a **binary tree** with the additional property that all logical connectives (other than possibly not) are expressed over the edges.

Then, for every accepting state $accept$, we find the lowest common ancestor of $accept$ and the leaf that currently classifies sample s. This $ancestor$ is the deepest node that has both $accept$ and s as its descendants, and it can be determined in linear time if all nodes in the data structure of T keep track of their parents [60]. We need to intervene on the $ancestor$ (i.e., change the value of the attribute of its if-then rule) to modify the path taken by s one node closer to $accept$. This intervention is associated with some cost that informs the total cost of recourse with respect to a particular $accept$ leaf (we propose some ways to estimate this cost in Section 11.4). Then, we store the node that required an intervention and the cost, and verify whether s is now in the accepting state. Finally, we output the recommendation, e.g., as the path that entails minimal cost.

This conversion **leaves the tree in a state** that also allows for the application of other algorithms, such as [31]; further (computational) research is required to decide when each of the approaches becomes advantageous.

11.3 Proof of total correctness

We start by acknowledging an important assumption: our algorithm is defined with hand-crafted rule-based systems in mind. In effect, the number of if-then rules in T is **finite and computationally tractable**. The same holds for the set of accepting states A.

Of course, **a tree where the if-then rules are inferred from data will also have a finite size** (it must be finite to classify all samples), but our algorithm may be inefficient for such models.

Our proof has two parts. First, in Section 11.3.1, we prove by construction that a generic decision tree can be converted into a binary tree with only atomic conditions in its internal nodes. Then, in Section 11.3.2, we use the same technique to prove that the algorithm can find recourse for any instance and that it terminates.

11.3.1 Part 1: Constructing binary atomic trees

Consider a generic tree T_G that may have nodes with more than two children (e.g., if the outcome depends on a combination of variables) and whose internal nodes describe arbitrary propositional formulas. We can **construct an equivalent binary atomic tree** in two steps.

We observe that there are **several ways to approach this construction**, e.g., in the first step, we could consider a *left-child, right-sibling* approach [143].

We start by converting T_G into a binary tree T_B; only the nodes with more than two children require attention. In a (multi-way) decision tree, the outcomes encoded by the children of a node are disjoint and their union covers the complete domain [77]. We visit all nodes in T_G (pre-order traversal) starting at its root. For each sub-tree rooted at T_S with more than two children, we apply the following operation: (1) create a new sub-tree rooted at T_S, (2) assign the left child of the original sub-tree as the left child in the new sub-tree, and (3) assign the negation of the left child as the right child. Then, all other children of the original sub-tree become the children of the right child of the new sub-tree. If this right child has more than two children, we apply the same operation again. As T_G was finite, this procedure terminates, yielding a finite binary decision tree T_B. We present this procedure in Figure 11.2.

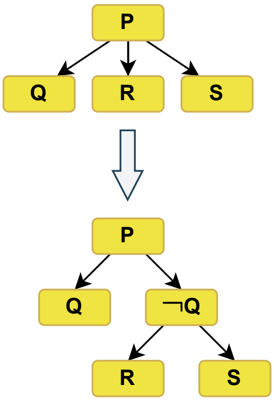

Figure 11.2: Converting T_G into T_B.

Next, we must address the arbitrary Boolean functions represented by the internal nodes of T_B. We note that any Boolean function can be transformed into the disjunctive normal form, i.e., a disjunction of conjunctions. Then, we again pre-order traverse the tree. For every internal node representing a non-atomic condition, we work inwards by splitting the disjunctions. This is presented in Figure 11.3 where Q itself can stand for another disjunction, in which case we iteratively split Q as well. Finally, we modify the conjunctions as presented in Figure 11.4, where similarly Q may require iterative treatment. As T_B was finite and each step in the construction requires a finite number of steps, the procedure terminates and yields T_A, a finite binary tree with atomic conditions in all internal nodes.

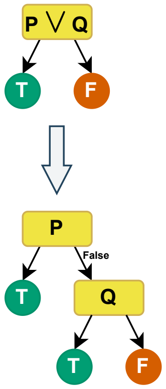

Figure 11.3: Converting Boolean `OR`.

11.3.2 Part 2: Generating recourse

Now, we can focus on the generation of algorithmic recourse. We start by observing that the classification of an arbitrary factual sample f can be represented as a path $T_1, \dots, T_m$ of if-then decision nodes visited by that sample, while its counterfactual c with respect to an accepting state a would trace the path $T_1, \dots, T_n$. In other words, the factual and the counterfactual begin at the same root and there exists some node T_i where the samples first took a different path. We again use the proof by construction to show that a factual can be turned into its counterfactual with respect to some accepting state a in a finite number of steps. First, we analyze the case of a decision stump, i.e., a situation where the factual and counterfactual differ by one decision. Then, we look at an arbitrary tree.

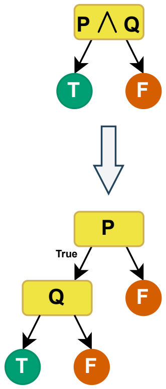

Figure 11.4: Converting Boolean `AND`.

Decision stump

Consider a decision stump T_A whose node classifies samples based on **the value of attribute** A into its two leaves: $accept$ and $reject$. We can move a sample f from the $reject$ state to the $accept$ state by modifying its value of the attribute A to a (minimal) value required by $accept$. Thus, we can trivially generate recourse for a decision stump. This step is represented on Line 9 of Algorithm 11.1.

We ensured that each individual decision relies on **one attribute** in Part 1.

Arbitrary tree

Now consider an arbitrary tree. While the exact order of if-then rules that it implements is unimportant, there exists (sequentially) the first node T_i where the value of attribute i of a factual sample f moves it into child T_j of T_i (i.e., one level closer to the reject state) while its counterfactual c with respect to state accept is moved into child T_k. In other words, before the intervention on T_i the paths taken by c and f agree up to and including node T_i: $T_1, T_2, \ldots, T_i$.

Node T_i is the lowest common ancestor of f and c and we can identify it in linear time. If each node stores the reference to its parent, we can simply traverse "backward" from c and from f, collecting both of their paths and then finding the first node where the two paths separate (Line 8 in Algorithm 11.1). Now, if we intervene on attribute i of c so that it moves into child T_k, the path shared by c and f in the tree becomes at least one step longer $T_1, T_2, \ldots T_i, T_k, \ldots$.

Note that the path of c and f from T_k onward may or may not agree. If it agrees, then we have effectively reached a case equivalent to the decision stump already at T_i. If it does not agree, then we repeat the process on the subtree rooted at T_k, which is one level smaller than the subtree rooted at T_i and thus move one step closer to the base case decision stump. Every time we intervene on some node, the height of the subtree that we still need to consider decreases by at least one, and so it is guaranteed to reach 0 at some point.

As we established, the **while** loop at Line 7 terminates. Furthermore, as the **for** loop at Line 5 of Algorithm 11.1 iterates over a finite set of accepting states in A, it must also terminate. Thus, we have proven the total correctness of the proposed algorithm.

11.4 Practical considerations

Finally, we need to address two important practicalities: the types of allowed interventions and the estimation of their costs. Regarding the former, we note that by construction the if-then rules may only take the form of **Boolean conditions or comparisons**. In this first (categorical) case, the intervention entails flipping the value of the attribute, while in the second (continuous) case, it entails modifying the value until the outcome of the comparison changes. Of course, the fact that a factual can be turned into a counterfactual does not necessarily mean that algorithmic recourse with respect to a particular state is viable. The task of ascertaining which attributes in the decision tree can be affected should be jointly carried out by domain experts and the end-users. To decide which path of recourse would entail the minimum cost (or otherwise require optimal effort) for the affected individual, we propose to make use of methods long-established in the field of multi-criteria decision-making, such as the Multi-Attribute Value Theory [127], the Analytic Hierarchy Process [205], or the Best-Worst Method [197].

For example, *"is true"*, *"is false"*, *"is (not) equal"*, *"is less/greater than"*, etc.

Conclusions 12

This chapter concludes the document. We start with Section 12.1, where we revisit the research questions and provide our answers. Then, in Section 12.2, we highlight the limitations of current work, which ties to the future challenges and the recommendations for researchers and practitioners that are discussed in Section 12.3. We wrap up in Section 12.4 with some final remarks about the challenges of algorithmic recourse and beyond.

12.1 Answers to research questions

In Chapter 3, we indicated that the research objective of this thesis is to address the socio-technical gap of algorithmic recourse by connecting its technical affordances with its social requirements. We explained that the problem space of algorithmic recourse is underexplored, but its in-depth understanding is required to encourage, shape, and support further development in the solution space. To that end, we formulated eight research questions that have been investigated throughout this thesis. Now, we recapitulate our answers and place them in the context of the complete document.

(RQ 1) How are the goals and tasks of algorithmic recourse defined and understood by researchers in the field?

We approached this question in Sections 4.2.1-4.2.3, where we found that the algorithmic recourse literature tends to focus on the technical aspects of the problem. While the overarching goal of AR seems clear – helping affected individuals overturn undesirable algorithmic decisions – important operational aspects of the task remain undefined. For instance, roughly one-fourth of all papers surveyed equate AR with "*actionable*" counterfactual explanations, but the understanding of this concept is very limited. Typically, it is explained as the ability of algorithmic recourse recommendations to respect certain constraints on the features, but questions such as *who* defines what is actionable tend to be sidestepped. Thus, we proposed to operationalize AR as *the provision of recommendations aligned with the preferences of non-expert users in an attempt to help them improve outcomes in an ADM setting* (Section 4.3.2).

RQ 1 impact on the research objective: Developing a shared understanding of the problem space is necessary to identify what is the *envisioned* role of algorithmic recourse mechanisms, i.e., what types of shortcomings of ADM systems the researchers aim to address.

At this point, we ought to give nuance to our definition. Our case study in Chapters 6-8 highlighted that three different problems seem to be confounded under the term "algorithmic recourse". It refers to (1) actionable **recommendations** (e.g., *"Alice was provided with algorithmic recourse."*), (2) the **process** of improving outcomes (e.g., *"Bob achieved algorithmic recourse with respect to the system."*), and (3) the task of implementing **mechanisms** to support end-users in this process (e.g., *"Charlie was unhappy with a decision but their bank offers algorithmic recourse."*). None of these problems are purely technical; barring a few exceptions, the literature tends to focus on this first, narrowest meaning of algorithmic recourse.

(RQ 2) What types of practical considerations are recognized and neglected in the literature on algorithmic recourse?

We found a large number of practical considerations for algorithmic recourse in the existing literature and discussed them in-depth in Sections 4.2.4-4.2.7. We observe that some topics are fashionable and attract major interest – for example, generating algorithmic recourse in causal settings – but many problems crucial for the real-world deployment of algorithmic recourse mechanisms have been observed and yet remain neglected: these include legal and ethical frameworks, supporting multi-agent systems, or mitigating the risks of abuse. Some problems may still belong to the "unknown unknowns" category, especially as the form of algorithmic recourse would depend on a system (see Section 9.1).

RQ 2 impact on the research objective: Identifying the practical considerations that have guided existing research and/or are observed as challenges for the future allows us to learn about the (promised) technical affordances of AR.

(RQ 3) How can algorithmic recourse complement the existing safety mechanisms for ADM tools in public administration?

We looked at this problem in Section 7.3. Our interviewees explained that algorithmic tools are commonly employed in public administration, but most of its needs are served by rule-based systems with no machine learning components because the Dutch General Administrative Law Act establishes strong transparency requirements for administrative decisions. This is not to say that if-then algorithms are without problems. For instance, they may rely on faulty logic in which case counterfactual explanations and/or algorithmic recourse recommendations could be a diagnostic tool. While the experts did not perceive algorithmic recourse as a solution that should be mandated in governmental systems, they recognized its value in the broader toolkit of possible safety interventions.

RQ 3 impact on the research objective: Learning about the social constraints and needs of a domain where AR could be implemented allows us to develop a better understanding of the socio-technical gap, including the *true* role that AR mechanisms could fulfill.

Of course, as we learned in Chapter 6, public administration entities may still attempt to use machine learning models. Risk profiling is not an administrative decision in the meaning of GALA. Hence, model-driven selection of benefits recipients was admissible in Rotterdam. As we argued in Section 2.3, algorithms in public administration contexts ought to be kept to particularly high standards. Algorithmic recourse, a set of solutions aiming to promote the agency of people, may be an important way to promote these goals.

(RQ 4) What needs of public administration could be addressed by algorithmic recourse but have not been explored yet?

We have co-discovered with our interviewees that algorithmic recourse may support public administration in three ways beyond the standard task of strengthening end-user agency (Section 7.3 and Section 8.4). First, algorithmic recourse may still be useful in expert systems, but all existing literature focuses on generating algorithmic recourse for machine learning models. We made a step towards addressing this gap in Chapter 11. Second, encouraging affected individuals to contest (algorithmic) decisions is a widely recognized problem. Providing recommendations on how to improve (algorithmic) outcomes shifts some burden of contesting a decision from the individual to the organization and may serve as "evidence" that the decision-maker prefers positive outcomes for the decision subjects. Third, counterfactual explanations and/or algorithmic recourse could be a form of an auditing tool available to the end-users because they may unveil some biases in the model. While we believe that the latter two values warrant further research, we cannot support their practicality with any empirical results.

RQ 4 impact on the research objective: Establishing whether AR mechanisms hold promise for additional positive impacts on ADM systems enables us to propose new directions for research in the solution space.

(RQ 5) How can the authorities explore the potential value of algorithmic recourse before implementing it in a system?

We proposed to make use of the techniques from the field of system safety to explore the place for algorithmic recourse mechanisms in a particular system. In Chapter 8, we applied System-Theoretic Process Analysis of [138] on the risk profiling system of Rotterdam, which allowed us to discover several potential use cases for algorithmic recourse, validating the utility of the technique for this purpose.

RQ 5 impact on the research objective: We aimed to appraise AR mechanisms from the perspective of a realistic socio-technical system. By establishing the applicability of STPA to reason about the value of algorithmic recourse interventions, we can support practitioners in the responsible adoption of AR.

We purposefully narrowed the scope of our analysis in Chapter 8 to algorithmic recourse, but of course, similar analyses will allow decision-makers to reason about other safety interventions [see 200]. Moreover, when applied iteratively, they allow to reason about the "safety of safety interventions". We decided to pursue a different approach to the exploration of hazards introduced by algorithmic recourse (RQ 7), but STPA is a solid alternative option.

(RQ 6) What are the ways to evaluate the quality of algorithmic recourse recommendations in practical settings?

In Chapter 9, we contributed a conceptual framework where three factors influence the quality of an algorithmic recourse mechanism. First, the utility of the recommendations requires that they accurately capture the characteristics of actionability. Second, the guarantees that implementing a recommendation will lead to the expected outcome, which depends on the ability to accommodate for the imperfect realization of the mechanism. Third, the capacity of an organization to implement the mechanism in a safe manner,

RQ 6 impact on the research objective: While benchmarking solutions for AR recommendations are readily available, we have not identified any tools to reason about the quality of AR mechanisms. Thus, our conceptual framework is another way to improve the understanding of the problem space.

i.e., in a way that respects the (business) constraints and requirements of the application. Notably, these correspond to the three meanings of algorithmic recourse that we have defined in RQ 1, and so the former two components – the actionable recommendations and the process of improving outcomes – are highly dependent on the latter component. Thus, we do not attempt to define generic metrics and instead provide guidance on what such metrics should capture. As the evaluation of AR in practical settings is a complex challenge, we propose digital twin solutions as a way to address it.

(RQ 7) To what extent do "digital twin" solutions allow for the reliable exploration of potential dynamics of algorithmic recourse before implementing it in a system?

We developed a simulation framework to explore the dynamics of algorithmic recourse in the Rotterdam case in Chapter 10. We place "digital twin" in quotation marks in this research question because our framework cannot be considered a proper digital twin due to our limited ability to verify and (especially) validate it. Still, we find satisfactory results in that the dynamics we observe correspond to the problems described by Lighthouse Reports in the Rotterdam case: agents that are nominated by a model for investigation are more likely to be re-nominated in the future. After introducing an algorithmic recourse mechanism, we find that it is able to counteract the phenomenon. Regarding dynamics, we only focus on the shifts in the decision-making process, but we observe no clear-cut unwanted effects. `SimulatedRecourse.jl` remains a proof-of-concept solution, so it is ultimately beyond our capability to decide whether digital twins are *the* way forward for algorithmic recourse, but they are definitely a promising solution. Even our simple models are capable of producing results that should be insightful for experts.

RQ 7 impact on the research objective: Exploring the value of simulation models for (more) realistic evaluation of AR mechanisms is a way to strengthen the connection between the technical affordances of algorithmic recourse, and the social and organizational constraints of systems that it could serve.

(RQ 8) What are the ways to align research on algorithmic recourse with the requirements of realistic domains?

Already in Section 4.3.2, we formulated five suggestions on how to (effectively) move from algorithmic recourse recommendations to algorithmic recourse mechanisms based on the literature review. These include: (1) broadening the scope of research, (2) engaging end-users, affected individuals, and communities, (3) accepting a socio-technical perspective, (4) accounting for emergent effects, and (5) attending to other operational aspects. We made our best effort to follow these suggestions in this draft. At the risk of sounding presumptuous, we believe that several of our contributions could not have materialized if we did not look at algorithmic recourse from the point of view of a real-world system. Indeed, the best way to bring algorithmic recourse out of theoretical computer science literature and into practical contexts may simply be to ground future research in the requirements of selected applications.

RQ 8 impact on the research objective: With this document, we showcased the process of reasoning about an AR mechanism in a real-world setting. We propose several forms of analyses that will be helpful (but not "required", viable alternatives probably exist) in connecting highly theoretical research on AR with the practical requirements of decision-making systems.

We return to the example from Section 4.3.2 where we explained that research on causality in algorithmic recourse, perceived to be one of the major challenges (highlighted in 29.7% of all publications that propose directions for future research), relies on the assumption that machine learning models in practical applications are causal. They are predominantly not. Causality does not appear to be seen by practitioners as a prerequisite to developing and deploying artificial intelligence systems in real-world contexts, and neither should it be seen as a requirement for algorithmic recourse tasks. That is to say, some "popular" research directions for AR are not necessarily substantive for real-world systems; they look far into the future **without contributing solutions** that could be applied to extant problems of (black-box) algorithmic decision-making.

Some desiderata may even **negatively affect the understanding** of a model we are trying to explain. For example, they may instill a false sense of the quality of the black-box model predictions [16].

There exist important practical challenges to introducing algorithmic recourse into decision-making systems. Many of them cannot be tackled by computer science literature alone. As was recognized by one of the experts in Chapter 7, *"developing an algorithm is actually a non-technical question"*. Similarly, aligning AR with the requirements of realistic domains will require the integration of various perspectives and multi-dimensional analyses. We believe that the process presented in this document may be of guidance. To the best of our knowledge, this has been the first attempt to design algorithmic recourse into an existing real-world decision-making system that goes beyond the generation of actionable recommendations, and so building upon our approach may help address the challenge.

12.2 Limitations of the current work

Where relevant, we have already highlighted the shortcomings of the various parts of our research in the corresponding sections of the document (Sections 4.4, 8.4, and 10.4). However, we still need to address two important limitations, both related to the overarching Design Science Research approach.

First, certain aspects of our work are motivated by inductive rather than deductive reasoning. As one example, while the evaluation framework in Chapter 9 is informed by the literature review, it mainly generalizes the observations that we made in the case study, meaning that it will not necessarily apply one-to-one in all settings. We strived to be completely transparent when our arguments follow from a (subjective) design process rather than empirical science. Notwithstanding, most of our work remains scientific in nature, including the systematized literature review in Chapter 4, the expert interviews in Chapter 7, the simulation framework and quantitative experiments in Chapter 10, and the method to generate actionable recommendations in expert systems in Chapter 11. Moreover, we note that the design artifacts are still testable and falsifiable.

Second, relatedly, we worked on an "external" case without the involvement of its owners. As our request for an interview with the experts from the Rotterdam W&I department was denied, our analyses of the case were left without confirmation from the decision-makers. This should not impact the validity of our arguments in a significant way because we supported them with evidence from high-quality publicly available sources. Still, this makes the decision-making system depicted in Chapters 6, 8, and 10 a simplification of reality. Further, our design artifacts have not been evaluated in the real-world systems for which they have been developed. We come close to a complete application of the Design Science Research process but fall short in our ability to close the "relevance cycle". Unfortunately, this also means that our work does not directly address the problem of limited applications of algorithmic recourse that we have identified in Section 4.2.1. We lay necessary groundwork for this task by developing the tools to apply algorithmic recourse, but the gap in the literature remains unsolved.

12.3 Challenges and future work

We direct this section to three groups of readers. We have already provided computer science researchers with the most important recommendations in Chapter 4. Here we summarize them in one point: we advocate foregoing the attempts to narrow down the task of algorithmic recourse to specific solutions (e.g., causal CEs) and instead focus on the *spirit* of the problem. Also, we have argued that many challenges ahead of practical AR may not be solvable with technical interventions, but the development of better tooling remains essential and certain desiderata for actionable recommendations may need to be fulfilled regardless of the system. Next, we invite researchers from other fields to consider the problems of algorithmic recourse; perspectives from business, law, economics, media studies, organization studies, psychology, sociology, systems theory, etc. can meaningfully contribute to the developing body of work. As one example, these fields are better equipped – both in terms of the established theoretical frameworks and the available research methods – to define the social aspects of actionability. Finally, even if they are not ready to attempt algorithmic recourse in real-world systems, we encourage practitioners to pursue it in the form of "thought experiments", similar to our analyses in this draft. A broad outlook on AR mechanisms informed by experts across a variety of potential domains would be immensely helpful to define what is *actually* important in AR research. In retrospect, the lack of solutions for expert systems is a glaring oversight but we have not identified earlier approaches to address this gap, potentially because the set of problems experienced by practitioners and the set of challenges tackled by researchers do not always overlap.

12.4 Final remarks

As the title of this document suggests, our goal was to find recourse for algorithmic recourse. On account of much theoretical interest but little to no practical applications of AR, we aimed to evaluate if this field of research holds value for real-world algorithmic decision-making systems. While researchers in the machine learning community have recognized Cynthia Rudin's 2019 call to stop explaining black-box models [204] – as evidenced by *Web Of Science* placing it among the top 1% of articles in computer science – the fact of the matter remains that black-box models are widely used in "consequential settings". This was also the case in Rotterdam W&I.

Will practitioners who (arguably) err by employing black-box models for high-stakes decisions be conscientious enough to look towards algorithmic recourse? Rudin recognized that policy-making efforts, in particular GDPR, attempt to establish the "right to explanation" rather than the "right to interpretable decisions" [204]. This is still true under the AI Act. As it focuses purely on AI systems and does not outlaw black-box decision-making, instead mandating sufficient explanation standards for automated decisions, we can reasonably expect post-hoc explainability techniques to become commonplace.

Thus, research on methods to generate counterfactual explanations and algorithmic recourse recommendations remains important, especially methods that guarantee faithfulness to the model [16]. Similarly, more research is needed on the ways to benefit from explainability techniques in practice, to tailor them to specific domains, and – through that – to improve responsible uptake of black-box models. This is a grand challenge that goes beyond algorithmic recourse to other explainability tasks and beyond social assistance to other decision-making contexts. Additionally, it is not only a challenge for science and engineering but also for policy-making. Even more broadly, algorithmic recourse could be seen as a *just one* process of contestable artificial intelligence [9, 116] in the late stages of model lifecycle or as *just one* mechanism to support meaningful human (end-user) control of algorithmic systems [41, 224].

Have we found recourse for algorithmic recourse? We believe so, but it requires a much broader interpretation of what it means to provide algorithmic recourse than currently observed in the literature. Ultimately, it is about the development of mechanisms to help people react to (unfavorable) algorithmic decisions. Actionable recommendations, the focal point of ongoing research on AR, may be *a necessary but, by no means, sufficient* part of such mechanisms.

Bibliography

[1] Abubakar Abid, Mert Yuksekgonul, and James Zou. 'Meaningfully Debugging Model Mistakes using Conceptual Counterfactual Explanations'. In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 66–88 (cited on page 44).

[2] Mark S. Ackerman. 'The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility'. In: *Human–Computer Interaction* 15.2-3 (2000), pp. 179–203. DOI: 10.1207/S15327051HCI1523_5 (cited on page 54).

[3] Gediminas Adomavicius and Alexander Tuzhilin. 'Discovery of Actionable Patterns in Databases: The Action Hierarchy Approach'. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. KDD'97. Newport Beach, CA: AAAI Press, 1997, 111–114 (cited on pages 6, 53).

[4] Farzana Afrin, Margaret Hamilton, and Charles Thevathyan. 'Exploring Counterfactual Explanations for Predicting Student Success'. In: *Computational Science – ICCS 2023*. Vol. 14074 LNCS. Springer Nature Switzerland, 2023, pp. 413–420. DOI: 10.1007/978-3-031-36021-3_44 (cited on pages 14, 24, 112).

[5] Muhammad Afzaal, Jalal Nouri, Aayesha Zia, Panagiotis Papapetrou, Uno Fors, Xiu Wu Yongchaoand Li, and Rebecka Weegar. 'Automatic and Intelligent Recommendations to Support Students' Self-Regulation'. In: *2021 International Conference on Advanced Learning Technologies (ICALT)*. July 2021, pp. 336–338. DOI: 10.1109/ICALT52272.2021.00107 (cited on pages 14, 20, 24, 110).

[6] Charu C. Aggarwal, Chen Chen, and Jiawei Han. 'The Inverse Classification Problem'. In: *Journal of Computer Science and Technology* 25 (2010), pp. 458–468 (cited on page 6).

[7] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. 'Counterfactual Shapley Additive Explanations'. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, 1054–1070. DOI: 10.1145/3531146.3533168 (cited on pages 14, 20, 111).

[8] Madison Alder. *DOJ seeks public input on AI use in criminal justice system*. Accessed 10.10.2024. Apr. 2024. URL: https://fedscoop.com/doj-seeks-input-on-criminal-justice-ai/ (cited on page 1).

[9] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 'Contestable AI by design: Towards a framework'. In: *Minds and Machines* (2022), pp. 1–27 (cited on pages 6, 85).

[10] Algorithm Audit. *Risk Profiling for Social Welfare Re-examination. Advice document*. Tech. rep. AA:2023:02:A. Algorithm Audit, Nov. 2023 (cited on pages 52, 67).

[11] Algorithm Audit. *Risk Profiling for Social Welfare Re-examination. Problem statement*. Tech. rep. AA:2023:02:P. Algorithm Audit, May 2023 (cited on pages 30, 31, 33).

[12] Hissah Alotaibi and Ronal Singh. 'Metrics for Evaluating Actionability in Explainable AI'. In: *PRICAI 2023: Trends in Artificial Intelligence*. Springer Nature Singapore, 2023, pp. 481–487 (cited on pages 14, 112).

[13] Patrick Altmeyer. *Trustworthy Artificial Intelligence in Julia*. Accessed: 2024-10-30. 2024. URL: https://www.taija.org/ (cited on page 58).

[14] Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia C. S. Liem. 'Endogenous Macrodynamics in Algorithmic Recourse'. In: *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 2023, pp. 418–431. DOI: 10.1109/SaTML54575.2023.00036 (cited on pages 7, 14, 21, 23, 24, 58, 63, 64, 72, 112).

[15] Patrick Altmeyer, Arie van Deursen, and Cynthia C. S. Liem. 'Explaining Black-Box Models through Counterfactuals'. In: *Proceedings of the JuliaCon Conferences* 1.1 (2023), p. 130. DOI: 10.21105/jcon.00130 (cited on page 58).

[16] Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. 'Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024, pp. 10829–10837 (cited on pages 26, 44, 83, 85).

[17] David Alvarez Melis and Tommi Jaakkola. 'Towards Robust Interpretability with Self-Explaining Neural Networks'. In: *Advances in Neural Information Processing Systems* 31 (2018) (cited on page 5).

[18] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 'Power to the People: The Role of Humans in Interactive Machine Learning'. In: *AI Magazine* 35.4 (2014), pp. 105–120 (cited on page 26).

[19] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. 'Evaluating Robustness of Counterfactual Explanations'. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. Dec. 2021, pp. 01–09. DOI: 10.1109/SSCI50451.2021.9660058 (cited on pages 14, 21, 110).

[20] Asefeh Asemi, Andrea Ko, and Mohsen Nowkarizi. 'Intelligent libraries: a review on expert systems, artificial intelligence, and robot'. In: *Library Hi Tech* 39.2 (2020), pp. 412–434 (cited on page 73).

[21] Association for Computing Machinery. *Words Matter. Alternatives for Charged Terminology in the Computing Profession.* Accessed: 2024-11-10. 2021. URL: https://www.acm.org/diversity-inclusion/words-matter (cited on page 2).

[22] Robert L. Axtell and J. Doyne Farmer. 'Agent-based modeling in economics and finance: Past, present, and future'. In: *Journal of Economic Literature* (2022), pp. 1–101 (cited on page 60).

[23] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 'The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2020, 80–89. DOI: 10.1145/3351095.3372830 (cited on pages 9, 14, 26, 110).

[24] Fernando J. Barros. 'Modeling formalisms for dynamic structure systems'. In: *ACM Trans. Model. Comput. Simul.* 7.4 (Oct. 1997), 501–515. DOI: 10.1145/268403.268423 (cited on page 59).

[25] Hosein Barzekar and Susan McRoy. 'Achievable Minimally-Contrastive Counterfactual Explanations'. In: *Machine Learning and Knowledge Extraction* 5.3 (2023), pp. 922–936. DOI: 10.3390/make5030048 (cited on pages 14, 22, 112).

[26] Barry Becker and Ronny Kohavi. *Adult*. UCI Machine Learning Repository. 1996. DOI: 10.24432/C5XW20 (cited on page 23).

[27] Sander Beckers. 'Causal Explanations and XAI'. In: *Proceedings of the First Conference on Causal Learning and Reasoning*. Ed. by Bernhard Schölkopf, Caroline Uhler, and Kun Zhang. Vol. 177. Proceedings of Machine Learning Research. PMLR, 2022, pp. 90–109 (cited on pages 14, 111).

[28] Andrew Bell, Joao Fonseca, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. *Fairness in Algorithmic Recourse Through the Lens of Substantive Equality of Opportunity*. 2024. URL: https://arxiv.org/abs/2401.16088 (cited on page 27).

[29] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018. URL: https://arxiv.org/abs/1810.01943 (cited on page 49).

[30] Alexander Berman, Ellen Breitholtz, Christine Howes, and Jean-Philippe Bernardy. 'Explaining Predictions with Enthymematic Counterfactuals'. In: *CEUR Workshop Proceedings*. Vol. 3319. CEUR-WS, 2022, pp. 95–100 (cited on pages 14, 20, 111).

[31] Tom Bewley, Salim I. Amoukou, Saumitra Mishra, Daniele Magazzeni, and Manuela Veloso. 'Counterfactual Metarules for Local and Global Recourse'. In: *arXiv preprint arXiv:2405.18875* (2024) (cited on pages 73, 76).

[32] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 'Explainable Machine Learning in Deployment'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, 648–657. DOI: 10.1145/3351095.3375624 (cited on page 7).

[33] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 'Power to the People? Opportunities and Challenges for Participatory AI'. In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2022, pp. 1–8 (cited on page 44).

[34] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 'The Values Encoded in Machine Learning Research'. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, 173–184. DOI: 10.1145/3531146.3533083 (cited on page 1).

[35] Justin-Casimir Braun, Eva Constantaras, Htet Aung, Gabriel Geiger, Dhruv Mehrotra, and Daniel Howden. *Suspicion Machine Methodology*. 2023. URL: https://www.lighthousereports.com/methodology/suspicion-machine/ (cited on pages 29, 32, 33, 35, 45, 46, 49).

[36] Matt Burgess, Evaline Schot, and Gabriel Geiger. *This Algorithm Could Ruin Your Life*. Mar. 2023. URL: https://www.wired.com/story/welfare-algorithms-discrimination/ (cited on pages 32, 33).

[37] Sarah Buss and Andrea Westlund. 'Personal Autonomy'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2018. Metaphysics Research Lab, Stanford University, 2018 (cited on page 6).

[38] Longbing Cao. 'Actionable knowledge discovery and delivery'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.2 (2012), pp. 149–163 (cited on page 53).

[39] Miguel Á. Carreira-Perpiñán and Suryabhan Singh Hada. 'Counterfactual Explanations for Oblique Decision Trees: Exact, Efficient Algorithms'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (May 2021), pp. 6903–6911. DOI: 10.1609/aaai.v35i8.16851 (cited on pages 14, 110).

[40] Claude Castelluccia and Daniel Le Métayer. *Understanding algorithmic decision-making: Opportunities and challenges*. Tech. rep. European Parliament, Mar. 2019 (cited on page 8).

[41] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker, Jeroen van den Hoven, Deborah Forster, and Reginald L. Lagendijk. 'Meaningful human control: actionable properties for AI system development'. In: *AI and Ethics* 3.1 (2023), pp. 241–255 (cited on page 85).

[42] Centraal Bureau voor de Statistiek. *Centraal Bureau voor de Statistiek*. Subpages therein. URL: https://www.cbs.nl/ (cited on pages 29, 33, 66, 67).

[43] Brian Chapman, Edward C. Page, and Frederick C. Mosher. *Public administration*. In: *Encyclopedia Britannica*. 2024 (cited on page 7).

[44] Liang Chen. 'Agent-based modeling in urban and architectural research: A brief literature review'. In: *Frontiers of Architectural Research* 1.2 (2012), pp. 166–177. DOI: https://doi.org/10.1016/j.foar.2012.03.003 (cited on page 60).

[45] Yatong Chen, Jialu Wang, and Yang Liu. 'Strategic Recourse in Linear Classification'. In: *Workshop on Consequential Decision Making in Dynamic Environments*. 2020 (cited on pages 14, 19, 110).

[46] Ziheng Chen, Fabrizio Silvestri, Gabriele Tolomei, Jia Wang, He Zhu, and Hongshik Ahn. 'Explain the Explainer: Interpreting Model-Agnostic Counterfactual Explanations of a Deep Reinforcement Learning Agent'. In: *IEEE Transactions on Artificial Intelligence* 5.4 (2024), pp. 1443–1457. DOI: 10.1109/TAI.2022.3223892 (cited on pages 14, 19, 22, 112).

[47] Furui Cheng, Yao Ming, and Huamin Qu. 'DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models'. In: *IEEE Transactions on Visualization & Computer Graphics* 27.02 (Feb. 2021), pp. 1438–1447. DOI: 10.1109/TVCG.2020.3030342 (cited on pages 14, 19, 21, 22, 110).

[48] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 'Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders'. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, 1–12. DOI: 10.1145/3290605.3300789 (cited on page 26).

[49] Sunil Choenni, Niels Netten, Mortaza Shoae-Bargh, and Rochelle Choenni. 'On the Usability of Big (Social) Data'. In: *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*. 2018, pp. 1167–1174. DOI: 10.1109/BDCloud.2018.00172 (cited on page 45).

[50] Lea Cohausz. 'Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science'. In: *Proceedings of the 15th International Conference on Educational Data Mining*. Durham, United Kingdom: International Educational Data Mining Society, July 2022, pp. 361–367. DOI: 10.5281/zenodo.6853069 (cited on pages 14, 24, 111).

[51] Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, and Beatriz San Miguel Gonzalez. 'Counterfactual Explanations as Interventions in Latent Space'. In: *Data Mining and Knowledge Discovery* (2022). DOI: 10.1007/s10618-022-00889-2 (cited on pages 14, 22, 111).

[52] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 'Multi-Objective Counterfactual Explanations'. In: *Parallel Problem Solving from Nature – PPSN XVI*. Cham: Springer International Publishing, 2020, pp. 448–469. DOI: 10.1007/978-3-030-58112-1_3 (cited on pages 14, 110).

[53] George Datseris, Ali R. Vahdati, and Timothy C. DuBois. 'Agents.jl: a performant and feature-full agent-based modeling software of minimal code complexity'. In: *Simulation* 100.10 (2024), pp. 1019–1031 (cited on page 60).

[54] Debanjan Datta, Feng Chen, and Naren Ramakrishnan. 'Framing Algorithmic Recourse for Anomaly Detection'. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, 283–293. doi: 10.1145/3534678.3539344 (cited on page 16).

[55] Randall Davis. 'Expert Systems: Where Are We? And Where Do We Go from Here?' In: *AI Magazine* 3.2 (June 1982), p. 3. doi: 10.1609/aimag.v3i2.367 (cited on page 73).

[56] Randall Davis, Andrew W. Lo, Sudhanshu Mishra, Arash Nourian, Manish Singh, Nicholas Wu, and Ruixun Zhang. 'Explainable Machine Learning Models of Consumer Credit Risk'. In: *Journal of Financial Data Science* 5.4 (2022), pp. 9–39. doi: 10.3905/jfds.2023.1.141 (cited on pages 14, 19, 111).

[57] Marcelo de Sousa Balbino, Luis Enrique Zárate Gálvez, and Cristiane Neri Nobre. 'CSSE - An agnostic method of counterfactual, selected, and social explanations for classification models'. In: *Expert Systems with Applications* 228 (2023), p. 120373. doi: https://doi.org/10.1016/j.eswa.2023.120373 (cited on pages 14, 112).

[58] Giovanni De Toni, Bruno Lepri, and Andrea Passerini. 'Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis'. In: *Machine Learning* 112.4 (2023), pp. 1389–1409. doi: 10.1007/s10994-022-06293-7 (cited on pages 14, 22, 112).

[59] Sarah Dean, Sarah Rich, and Benjamin Recht. 'Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, 436–445. doi: 10.1145/3351095.3372866 (cited on page 16).

[60] Hristo N. Djidjev, Grammati E. Pantziou, and Christos D Zaroliagis. 'Computing shortest paths and distances in planar graphs'. In: *Automata, Languages and Programming: 18th International Colloquium Madrid, Spain, July 8–12, 1991 Proceedings 18*. Springer. 1991, pp. 327–338 (cited on page 76).

[61] Roel Dobbe. *System Safety and Artificial Intelligence*. 2022. url: https://arxiv.org/abs/2202.09292 (cited on page 45).

[62] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 'Hard choices in artificial intelligence'. In: *Artificial Intelligence* 300 (2021), p. 103555. doi: https://doi.org/10.1016/j.artint.2021.103555 (cited on page 24).

[63] Roel Dobbe and Anouk Wolters. 'Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context'. In: *Minds and Machines* 34.2 (2024), pp. 1–51 (cited on page 24).

[64] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, and Bernhard Schölkopf. 'On the Adversarial Robustness of Causal Algorithmic Recourse'. In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 5324–5342 (cited on pages 14, 19, 26, 57, 111).

[65] Michael Downs, Jonathan L. Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. 'CRUDS: Counterfactual Recourse Using Disentangled Subspaces'. In: *ICML Workshop on Human Interpretability in Machine Learning* (2020), pp. 1–23 (cited on pages 14, 22, 26, 110).

[66] Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schoelkopf, and Setareh Maghsudi. 'Robustness Implies Fairness in Causal Algorithmic Recourse'. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, 984–1001. DOI: 10.1145/3593013.3594057 (cited on pages 14, 20, 21, 112).

[67] Julia El Zini and Mariette Awad. 'Beyond Model Interpretability: On the Faithfulness and Adversarial Robustness of Contrastive Textual Explanations'. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, 2022, pp. 1391–1402. DOI: 10.18653/v1/2022.findings-emnlp.100 (cited on page 16).

[68] Lena Enqvist. 'Rule-based versus AI-driven benefits allocation: GDPR and AIA legal implications and challenges for automation in public social security administration'. In: *Information & Communications Technology Law* 33.2 (2024), pp. 222–246. DOI: 10.1080/13600834.2024.2349835 (cited on pages 8, 73).

[69] Epoch AI. *Key Trends and Figures in Machine Learning*. Accessed: 2024-10-14. 2023. URL: https://epochai.org/trends (cited on page 5).

[70] Seyedehdelaram Esfahani, Giovanni De Toni, Bruno Lepri, Andrea Passerini, Katya Tentori, and Massimo Zancanaro. 'Preference Elicitation in Interactive and User-centered Algorithmic Recourse: an Initial Exploration'. In: *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '24. New York, NY, USA: Association for Computing Machinery, 2024, 249–254. DOI: 10.1145/3627043.3659556 (cited on page 27).

[71] Andrew Estornell, Yatong Chen, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. 'Incentivizing Recourse through Auditing in Strategic Classification'. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence, Aug. 2023, pp. 400–408. DOI: 10.24963/ijcai.2023/45 (cited on pages 14, 19, 21, 112).

[72] Virginia Eubanks. *Digital dead end: Fighting for social justice in the information age*. MIT Press, 2012 (cited on page 8).

[73] Virginia Eubanks. 'Want to Predict the Future of Surveillance? Ask Poor Communities'. In: *The American Prospect* 15 (2014) (cited on page 8).

[74] Andrea Ferrario and Michele Loi. 'The Robustness of Counterfactual Explanations Over Time'. In: *IEEE Access* 10 (2022), pp. 82736–82750. DOI: 10.1109/ACCESS.2022.3196917 (cited on pages 14, 26, 57, 111).

[75] FICO. *Explainable Machine Learning Challenge*. 2018. URL: https://community.fico.com/s/explainable-machine-learning-challenge (cited on page 23).

[76] João Fonseca, Andrew Bell, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. 'Setting the Right Expectations: Algorithmic Recourse Over Time'. In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '23. New York, NY, USA: Association for Computing Machinery, 2023. DOI: 10.1145/3617694.3623251 (cited on page 27).

[77] Eibe Frank and Ian H. Witten. *Selecting multiway splits in decision trees*. Working Paper 96/31. Hamilton, New Zealand: University of Waikato, Department of Computer Science, 1996 (cited on page 77).

[78] Susanne Friese, Jacks Soratto, and Denise Pires de Pires. 'Carrying out a computer-aided thematic content analysis with ATLAS.ti'. In: *IWMI Working Papers* 18 (Apr. 2018) (cited on page 17).

[79] Bernardo Alves Furtado. 'Simulation Modeling as a Policy Tool'. In: *The Routledge Handbook of Policy Tools*. Ed. by Michael Howlett. Taylor & Francis, 2023. DOI: 10.4324/9781003163954 (cited on page 60).

[80] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 'Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals'. In: *Proceedings of the 2021 International Conference on Management of Data*. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 577–590. DOI: 10.1145/3448016.3458455 (cited on pages 14, 20, 21, 110).

[81] Ruijiang Gao and Himabindu Lakkaraju. 'On the Impact of Algorithmic Recourse on Social Segregation'. In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. JMLR.org, 2023 (cited on pages 14, 20–22, 24, 26, 112).

[82] Gabriel Geiger. *How We Did It: Unlocking Europe's Welfare Fraud Algorithms*. July 2023. URL: https://pulitzercenter.org/how-we-did-it-unlocking-europes-welfare-fraud-algorithms (cited on page 49).

[83] Gabriel Geiger, Eva Constantaras, Justin-Casimir Braun, Htet Aung, Evaline Schot, Saskia Klassen, Romy van Dijk, David Davidson, Dhruv Mehrota, Morgan Meaker, Matthew Burgess, Kyle Thomas, Alyssa Walker, Katherine Lam, Sam Lavigne, Amy Qu, Raagul Nagendran, Hari Moorthy, Ishita Tiwari, Danielle Carrick, Lily Boyce, Andrew Couts, James Temperton, Daniel Howden, Soizic Penicaud, Pablo Jiménez Arandia, Reinier Tromp, Tom Claessens, Antonella Napolitano, Ariadne Papagapitos, Marina Walker, Boyoung Lim, Eeva Liukku, Melissa van Amerongen, Willemijn Sneep, Sascha Meijer, Roelif van der Meer, Tom Simonite, and Fanis Kollias. *Suspicion Machines*. Mar. 2023. URL: https://www.lighthousereports.com/investigation/suspicion-machines/ (cited on page 32).

[84] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 'PRINCE: Provider-Side Interpretability with Counterfactual Explanations in Recommender Systems'. In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. WSDM '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 196–204. DOI: 10.1145/3336191.3371824 (cited on page 16).

[85] Nigel Gilbert and Klaus G Troitzsch. *Simulation for the Social Scientist*. USA: Open University Press, 2005 (cited on page 60).

[86] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 'ViCE: Visual Counterfactual Explanations for Machine Learning Models'. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 531–535. DOI: 10.1145/3377325.3377536 (cited on pages 14, 21, 22, 110).

[87] Bryce Goodman and Seth Flaxman. 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"'. In: *AI Magazine* 38.3 (2017), pp. 50–57 (cited on page 30).

[88] Claudius Graebner. 'How to Relate Models to Reality? An Epistemological Framework for the Validation and Verification of Computational Models'. In: *Journal of Artificial Societies and Social Simulation* 21.3 (2018), p. 8. DOI: 10.18564/jasss.3772 (cited on page 71).

[89] Crystal Grant. *Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism*. Accessed 22.05.2024. Oct. 2022. URL: https://www.aclu.org/news/privacy-technology/algorithms-in-health-care-may-worsen-medical-racism (cited on page 1).

[90] Maria J. Grant and Andrew Booth. 'A typology of reviews: an analysis of 14 review types and associated methodologies'. In: *Health Information & Libraries Journal* 26.2 (2009), pp. 91–108. DOI: https://doi.org/10.1111/j.1471-1842.2009.00848.x (cited on pages 14, 15).

[91] Stephan Grimmelikhuijsen and Albert Meijer. 'Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response'. In: *Perspectives on Public Management and Governance* 5.3 (Mar. 2022), pp. 232–242. DOI: 10.1093/ppmgov/gvac008 (cited on page 1).

[92] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 'Why do tree-based models still outperform deep learning on typical tabular data?' In: Red Hook, NY, USA: Curran Associates Inc., 2022 (cited on page 54).

[93] Riccardo Guidotti. 'Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking'. In: *Data Mining and Knowledge Discovery* (2022). DOI: 10.1007/s10618-022-00831-6 (cited on pages 6, 14, 21, 111).

[94] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Francesca Naretto, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 'Stable and Actionable Explanations of Black-Box Models through Factual and Counterfactual Rules'. In: *Data Mining and Knowledge Discovery* (2022). DOI: 10.1007/s10618-022-00878-5 (cited on pages 14, 73, 111).

[95] Riccardo Guidotti and Salvatore Ruggieri. 'Ensemble of Counterfactual Explainers'. In: *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2021, pp. 358–368. DOI: 10.1007/978-3-030-88942-5_28 (cited on pages 14, 110).

[96] Hangzhi Guo, Feiran Jia, Jinghui Chen, Anna Squicciarini, and Amulya Yadav. 'RoCourseNet: Robust Training of a Prediction Aware Recourse Model'. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 619–628. DOI: 10.1145/3583780.3615040 (cited on pages 14, 20, 112).

[97] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 'Equalizing Recourse Across Groups'. In: *arXiv* (2019) (cited on pages 14, 16, 21, 110).

[98] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. 'Generating Robust Counterfactual Explanations'. In: *Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2023*. Berlin, Heidelberg: Springer-Verlag, 2023, pp. 394–409. DOI: 10.1007/978-3-031-43418-1_24 (cited on pages 14, 112).

[99] Victor Guyomard, Françoise Fessant, Tassadit Bouadi, and Thomas Guyet. 'Post-Hoc Counterfactual Generation with Supervised Autoencoder'. In: *Communications in Computer and Information Science*. Vol. 1524 CCIS. Springer Science and Business Media Deutschland GmbH, 2021, pp. 105–114. DOI: 10.1007/978-3-030-93736-2_10 (cited on pages 14, 110).

[100] Suryabhan Singh Hada and Miguel Á. Carreira-Perpiñán. 'Exploring Counterfactual Explanations for Classification and Regression Trees'. In: *Communications in Computer and Information Science*. Vol. 1524 CCIS. Springer Science and Business Media Deutschland GmbH, 2021, pp. 489–504. DOI: 10.1007/978-3-030-93736-2_37 (cited on pages 14, 110).

[101] 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. New York, NY, USA: Association for Computing Machinery, 2022, pp. 666–675. DOI: 10.1145/3511808.3557424 (cited on pages 14, 111).

[102] Ian Hardy, Jayanth Yetukuri, and Yang Liu. 'Adaptive Adversarial Training Does Not Increase Recourse Costs'. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 432–442. DOI: 10.1145/3600211.3604704 (cited on pages 14, 112).

[103] Sadik Hasan. 'Governance and Public Administration'. In: *Global Encyclopedia of Public Administration, Public Policy, and Governance*. Ed. by Ali Farazmand. Cham: Springer International Publishing, 2018, pp. 1–6. DOI: 10.1007/978-3-319-31816-5_1820-1 (cited on page 7).

[104] Trevor Hastie and Robert Tibshirani. 'Generalized Additive Models'. In: *Statistical Science* 1.3 (1986), pp. 297 –310. DOI: 10.1214/ss/1177013604 (cited on page 5).

[105] Zhian He and Eric Lo. 'Answering Why-not Questions on Top-k Queries'. In: *2012 IEEE 28th International Conference on Data Engineering* (2012), pp. 750–761. DOI: 10.1109/ICDE.2012.8 (cited on page 6).

[106] Scott Heckbert, Tim Baynes, and Andrew Reeson. 'Agent-based modeling in ecological economics'. In: *Annals of the New York Academy of Sciences* 1185.1 (2010), pp. 39–53. DOI: https://doi.org/10.1111/j.1749-6632.2009.05286.x (cited on page 60).

[107] Lars Herbold, Mersedeh Sadeghi, and Andreas Vogelsang. 'Generating Context-Aware Contrastive Explanations in Rule-based Systems'. In: *Proceedings of the 2024 Workshop on Explainability Engineering*. New York, NY, USA: Association for Computing Machinery, 2024, 8–14. DOI: 10.1145/3648505.3648507 (cited on page 73).

[108] Alan R. Hevner. 'A Three Cycle View of Design Science Research'. In: *Scandinavian Journal of Information Systems* (2007), pp. 87–92 (cited on page 12).

[109] Alan R. Hevner and Samir Chatterjee. 'Design Science Research in Information Systems'. In: *Design Research in Information Systems: Theory and Practice*. Boston, MA: Springer US, 2010, pp. 9–22. DOI: 10.1007/978-1-4419-5653-8_2 (cited on page 12).

[110] Alan R. Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. 'Design Science in Information Systems Research'. In: *MIS Quarterly* (2004), pp. 75–105 (cited on page 12).

[111] Hans Hofmann. *Statlog (German Credit Data)*. UCI Machine Learning Repository. 1994. DOI: 10.24432/C5NC77 (cited on pages 18, 23).

[112] Jacqueline Höllig, Aniek F. Markus, JJef de Slegte, and Prachi Bagave. 'Semantic Meaningfulness: Evaluating Counterfactual Approaches for Real-World Plausibility and Feasibility'. In: *Communications in Computer and Information Science*. Vol. 1902 CCIS. Springer Science and Business Media Deutschland GmbH, 2023, pp. 636–659. DOI: 10.1007/978-3-031-44067-0_32 (cited on pages 14, 20, 112).

[113] Jeremie Desgagne-Bouchard. *NeuroTreeModels.jl*. 2024. URL: https://github.com/Evovest/NeuroTreeModels.jl (cited on page 67).

[114] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 'Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems'. In: *arXiv* (2019) (cited on pages 14, 16, 26, 110).

[115] Sarathi K, Shania Mitra, Deepak P, and Sutanu Chakraborti. 'Counterfactuals as Explanations for Monotonic Classifiers'. In: *CEUR Workshop Proceedings*. Vol. 3389. CEUR-WS, 2022, pp. 177–188 (cited on pages 14, 20, 111).

[116] Margot E. Kaminski and Jennifer M. Urban. 'The right to contest AI'. In: *Columbia Law Review* 121.7 (2021), pp. 1957–2048 (cited on page 85).

[117] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. 'Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization'. In: *Transactions of the Japanese Society for Artificial Intelligence* 36.6 (2021). DOI: 10.1527/TJSAI.36-6_C-L44 (cited on pages 14, 110).

[118] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. 'Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees'. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Vol. 151. PMLR, 2022, pp. 1846–1870 (cited on pages 14, 19, 21, 111).

[119] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike, Kento Uemura, and Hiroki Arimura. 'Ordered Counterfactual Explanation by Mixed-Integer Linear Optimization'. In: *35th AAAI Conference on Artificial Intelligence, AAAI 2021*. Vol. 13A. Association for the Advancement of Artificial Intelligence, 2021, pp. 11564–11574 (cited on pages 14, 20, 110).

[120] 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*. Vol. 108. PMLR, 2020, pp. 895–905 (cited on pages 14, 110).

[121] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 'A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations'. In: *ACM Computing Surveys* 55.5 (Dec. 2022). DOI: 10.1145/3527848 (cited on pages 14, 18, 21, 111).

[122] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 'Algorithmic Recourse: From Counterfactual Explanations to Interventions'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 353–362. DOI: 10.1145/3442188.3445899 (cited on pages 6, 7, 14, 110).

[123] 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* 33 (2020), pp. 265–277 (cited on pages 14, 22, 26, 110).

[124] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 'Towards Causal Algorithmic Recourse'. In: *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Cham: Springer International Publishing, 2020, pp. 139–166. DOI: 10.1007/978-3-031-04083-2_8 (cited on page 21).

[125] Robert Kass and Tim Finin. 'The Need for User Models in Generating Expert System Explanations'. In: *International Journal of Expert Systems* 1.4 (Oct. 1988) (cited on page 73).

[126] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 'If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques'. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Survey Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4466–4474. DOI: 10.24963/ijcai.2021/609 (cited on page 22).

[127] Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, 1993 (cited on page 78).

[128] Nwaike Kelechi and Licheng Jiao. 'Quantifying Actionability: Evaluating Human-Recipient Models'. In: *IEEE Access* 11 (2023), pp. 119811–119823. DOI: 10.1109/ACCESS.2023.3324906 (cited on pages 14, 19, 24, 112).

[129] Seunghun Koh, Byung Hyung Kim, and Sungho Jo. 'Understanding the User Perception and Experience of Interactive Algorithmic Recourse Customization'. In: *ACM Trans. Comput.-Hum. Interact.* 31.3 (Aug. 2024). DOI: 10.1145/3674503 (cited on page 27).

[130] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. 'Causal Perspective on Meaningful and Robust Algorithmic Recourse'. In: *ICML Workshop on Algorithmic Recourse* (2021) (cited on pages 14, 21, 110).

[131] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. 'Improvement-Focused Causal Recourse (ICR)'. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. doi: 10.1609/aaai.v37i10.26398 (cited on pages 14, 112).

[132] Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. 'Towards Bridging the Gaps between the Right to Explanation and the Right to Be Forgotten'. In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. JMLR.org, 2023 (cited on pages 14, 20, 21, 112).

[133] Julius von Kügelgen, Nikita Agarwal, Jakob Zeitler, Afsaneh Mastouri, and Bernhard Schölkopf. 'Algorithmic Recourse in Partially and Fully Confounded Settings Through Bounding Counterfactual Effects'. In: *arXiv* (2021) (cited on pages 14, 16, 110).

[134] Anisio Lacerda, Claudio Almeida, Leonardo Ferreira, Adriano Pereira, Gisele L. Pappa, Wagner Meira, Debora Miranda, Marco A. Romano-Silva, and Leandro Malloy Diniz. 'Algorithmic Recourse in Mental Healthcare'. In: *2023 International Joint Conference on Neural Networks (IJCNN)*. June 2023, pp. 1–8. doi: 10.1109/IJCNN54540.2023.10191158 (cited on pages 14, 19, 24, 25, 112).

[135] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. *How We Analyzed the COMPAS Recidivism Algorithm*. May 2016. url: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (cited on page 23).

[136] Derek Leben. 'Explainable AI as Evidence of Fair Decisions'. In: *Frontiers in Psychology* 14 (2023). doi: 10.3389/fpsyg.2023.1069426 (cited on pages 9, 14, 20, 24, 43, 112).

[137] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 'Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model'. In: *The Annals of Applied Statistics* 9.3 (2015), pp. 1350 –1371. doi: 10.1214/15-AOAS848 (cited on page 73).

[138] Nancy G. Leveson. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016 (cited on pages 45, 47, 50, 81).

[139] Karen Levy, Kyla E. Chasalow, and Sarah Riley. 'Algorithms and decision-making in the public sector'. In: *Annual Review of Law and Social Science* 17.1 (2021), pp. 309–334 (cited on page 7).

[140] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. 'GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations'. In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023 (cited on pages 14, 21, 112).

[141] Raz Lin, Sarit Kraus, Tim Baarslag, Dmytro Tykhonov, Koen Hindriks, and Catholijn M. Jonker. 'Genius: An integrated environment for supporting the design of generic automated negotiators'. In: *Computational Intelligence* 30.1 (2014), pp. 48–70 (cited on page 72).

[142] Fabian Lorig, Loïs Vanhée, and Frank Dignum. 'Agent-Based Social Simulation for Policy Making'. In: *Human-Centered Artificial Intelligence: Advanced Lectures*. Ed. by Mohamed Chetouani, Virginia Dignum, Paul Lukowicz, and Carles Sierra. Cham: Springer International Publishing, 2023, pp. 391–414. doi: 10.1007/978-3-031-24349-3_20 (cited on page 60).

[143] Alen Lovrencic and Paul E. Black. 'binary tree representation of trees'. In: *Dictionary of Algorithms and Data Structures*. Ed. by Paul E. Black. National Institute of Standards and Technology, Nov. 2008 (cited on page 76).

[144] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. 'FOCUS: Flexible Optimizable Counterfactual Explanations for Tree Ensembles'. In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*. Vol. 36. 2022, pp. 5313–5322 (cited on pages 14, 111).

[145] Scott M. Lundberg and Su-In Lee. 'A unified approach to interpreting model predictions'. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, 4768–4777 (cited on page 5).

[146] Shucen Ma, Jianqi Shi, Yanhong Huang, Shengchao Qin, and Zhe Hou. 'Minimal-unsatisfiable-core-driven Local Explainability Analysis for Random Forest'. In: *International Journal of Software and Informatics* 12.4 (2022), pp. 355–376. DOI: 10.21655/ijsi.1673-7288.00280 (cited on pages 14, 111).

[147] Jonne Maas. 'Machine learning and power relations'. In: *AI & SOCIETY* 38.4 (2023), pp. 1493–1500 (cited on page 8).

[148] Michael W. Macy and Robert Willer. 'From Factors to Actors: Computational Sociology and Agent-Based Modeling'. In: *Annual Review of Sociology* 28.1 (2002), pp. 143–166 (cited on page 60).

[149] Alexander Maedche, Shirley Gregor, Stefan Morana, and Jasper Feine. 'Conceptualization of the Problem Space in Design Science Research'. In: *Extending the Boundaries of Design Science Theory and Practice*. Ed. by Bengisu Tulu, Soussan Djamasbi, and Gondy Leroy. Cham: Springer International Publishing, 2019, pp. 18–31 (cited on page 9).

[150] Paul P. Maglio and Patricia L. Mabry. 'Agent-Based Models and Systems Science Approaches to Public Health'. In: *American Journal of Preventive Medicine* 40.3 (2011), pp. 392–394 (cited on page 60).

[151] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 'Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers'. In: *NeurIPS 2019 Workshop "Do the right thing": machine learning and causal inference for improved decision making*. 2019 (cited on pages 14, 22, 26, 110).

[152] Ričards Marcinkevičs and Julia E. Vogt. 'Interpretable and explainable machine learning: a methods-centric overview with concrete examples'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13.3 (2023), e1493 (cited on page 5).

[153] Conor Mayo-Wilson and Kevin J. S. Zollman. 'The computational philosophy: simulation as a core philosophical method'. In: *Synthese* 199.1 (2021), pp. 3647–3673 (cited on page 71).

[154] Raphael Mazzine, Sofie Goethals, Dieter Brughmans, and David Martens. 'Counterfactual Explanations for Employment Services'. In: *International workshop on AI for Human Resources and Public Employment Services* (2021) (cited on pages 14, 19, 110).

[155] Md Golam Moula Mehedi Hasan and Douglas A. Talbert. 'Mitigating the Rashomon Effect in Counterfactual Explanation: A Game-theoretic Approach'. In: *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS*. Vol. 35. Florida Online Journals, University of Florida, 2022. DOI: 10.32473/flairs.v35i.130711 (cited on pages 14, 111).

[156] Tim Miller. 'Explanation in artificial intelligence: Insights from the social sciences'. In: *Artificial Intelligence* 267 (2019), pp. 1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007 (cited on page 5).

[157] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. *Handreiking Algoritmeregister. Aan de slag met het Algoritmeregister.* Tech. rep. 1.0. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, Nov. 2023 (cited on page 31).

[158] Jonathan Moore, Nils Hammerla, and Chris Watkins. 'Explaining Deep Learning Models with Constrained Adversarial Examples'. In: *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2019, pp. 43–56. DOI: 10.1007/978-3-030-29908-8_4 (cited on pages 14, 110).

[159] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 'Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 607–617. DOI: 10.1145/3351095.3372850 (cited on pages 14, 23, 110).

[160] Madhumita Murgia. *Algorithms are deciding who gets organ transplants. Are their decisions fair?* Accessed 22.05.2024. Nov. 2023. URL: https://www.ft.com/content/5125c83a-b82b-40c5-8b35-99579e087951 (cited on page 1).

[161] Alexey Natekin and Alois Knoll. 'Gradient boosting machines, a tutorial'. In: *Frontiers in neurorobotics* 7 (2013), p. 21 (cited on page 49).

[162] Philip Naumann and Eirini Ntoutsi. 'Consequence-Aware Sequential Counterfactual Generation'. In: *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2021, pp. 682–698. DOI: 10.1007/978-3-030-86520-7_42 (cited on pages 14, 110).

[163] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. 'CounteRGAN: Generating Counterfactuals for Real-Time Recourse and Interpretability Using Residual GANs'. In: *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence, UAI 2022*. Association For Uncertainty in Artificial Intelligence (AUAI), 2022, pp. 1488–1497 (cited on pages 14, 22, 26, 111).

[164] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. 'Providing Actionable Feedback in Hiring Marketplaces Using Generative Adversarial Networks'. In: *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, 2021, pp. 1089–1092. DOI: 10.1145/3437963.3441705 (cited on pages 14, 19, 22, 24, 57, 110).

[165] Niels Netten, Mortaza S. Bargh, and Sunil Choenni. 'Exploiting Data Analytics for Social Services: On Searching for Profiles of Unlawful Use of Social Benefits'. In: *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*. ICEGOV '18. New York, NY, USA: Association for Computing Machinery, 2018, 550–559. DOI: 10.1145/3209415.3209481 (cited on page 45).

[166] Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. 'Distributionally Robust Recourse Action'. In: *arXiv* (2023) (cited on pages 14, 16, 20, 112).

[167] Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. 'Feasible Recourse Plan via Diverse Interpolation'. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. PMLR, 2023, pp. 4679–4698 (cited on pages 14, 112).

[168] 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*. Vol. 180. PMLR, 2022, pp. 1498–1508 (cited on pages 14, 20, 22, 111).

[169] Andrew O'Brien and Edward Kim. 'Toward Multi-Agent Algorithmic Recourse: Challenges From a Game-Theoretic Perspective'. In: *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS*. Vol. 35. Florida Online Journals, University of Florida, 2022. DOI: 10.32473/flairs.v35i.130614 (cited on pages 7, 14, 21, 111).

[170] Andrew O'Brien, Edward Kim, and Rosina Weber. 'Investigating Causally Augmented Sparse Learning as a Tool for Meaningful Classification'. In: *2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. Sept. 2023, pp. 33–37. DOI: 10.1109/AIKE59827.2023.00013 (cited on pages 14, 112).

[171] Ming Lun Ong, Anthony Li, and Mehul Motani. 'Explainable and Actionable Machine Learning Models for Electronic Health Record Data'. In: *IFMBE Proceedings*. Vol. 79. Cham: Springer International Publishing, 2021, pp. 91–99. DOI: 10.1007/978-3-030-62045-5_9 (cited on pages 14, 20, 24, 25, 110).

[172] Overheid.nl. *Het Algoritmeregister*. Subpages therein. URL: https://algoritmes.overheid.nl/ (cited on page 31).

[173] Overheid.nl. *Lokale wet- en regelgeving*. Subpages therein. URL: https://lokaleregelgeving.overheid.nl/ (cited on pages 31, 34, 45, 46).

[174] Overheid.nl. *Wettenbank*. Subpages therein. URL: https://wetten.overheid.nl/ (cited on page 29).

[175] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews'. In: *Bmj* 372 (2021) (cited on page 15).

[176] Axel Parmentier and Thibaut Vidal. 'Optimal Counterfactual Explanations in Tree Ensembles'. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. PMLR, 2021, pp. 8422–8431 (cited on pages 14, 20, 110).

[177] 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*. Vol. 151. PMLR, 2022, pp. 4574–4594 (cited on pages 14, 20, 111).

[178] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. 'CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms'. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)*. 2021 (cited on pages 14, 19, 23, 58, 110).

[179] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 'Learning Model-Agnostic Counterfactual Explanations for Tabular Data'. In: *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*. 2020, pp. 3126–3132. DOI: 10.1145/3366423.3380087 (cited on pages 14, 22, 26, 110).

[180] Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. 'On the Privacy Risks of Algorithmic Recourse'. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. PMLR, 2023, pp. 9680–9696 (cited on pages 14, 19, 112).

[181] Judea Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009 (cited on page 6).

[182] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 'FACE: Feasible and Actionable Counterfactual Explanations'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 344–350. DOI: 10.1145/3375627.3375850 (cited on pages 14, 22, 110).

[183] Programmaraad Samenvoordeklant.nl. *Werkwijzer Tegenprestatie*. Tech. rep. Programmaraad Samenvoordeklant.nl, Sept. 2018 (cited on page 29).

[184] Wenting Qi and Charalampos Chelmis. 'Improving Algorithmic Decision–Making in the Presence of Untrustworthy Training Data'. In: *2021 IEEE International Conference on Big Data (Big Data)*. 2021, pp. 1102–1108. DOI: 10.1109/BigData52589.2021.9671677 (cited on pages 14, 110).

[185] Marcos M. Raimundo, Luis Gustavo Nonato, and Jorge Poco. 'Mining Pareto-optimal Counterfactual Antecedents with a Branch-and-Bound Model-Agnostic Algorithm'. In: *Data Mining and Knowledge Discovery* (2022). DOI: 10.1007/s10618-022-00906-4 (cited on pages 14, 111).

[186] Goutham Ramakrishnan, Yun Chan Lee, and Aws Albarghouthi. 'Synthesizing Action Sequences for Modifying Model Decisions'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 5462–5469 (cited on pages 14, 22, 110).

[187] Natraj Raman, Daniele Magazzeni, and Sameena. Shah. 'Bayesian Hierarchical Models for Counterfactual Estimation'. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. PMLR, 2023, pp. 1115–1128 (cited on pages 14, 21, 112).

[188] Gomathy Ramaswami, Teo Susnjak, and Anuradha Mathrani. 'Supporting Students' Academic Performance Using Explainable Machine Learning with Automated Prescriptive Analytics'. In: *Big Data and Cognitive Computing* 6.4 (2022). DOI: 10.3390/bdcc6040105 (cited on pages 14, 24, 111).

[189] Zbigniew W. Ras and Alicja Wieczorkowska. 'Action-Rules: How to Increase Profit of a Company'. In: *Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg, 2000, pp. 587–592 (cited on page 6).

[190] Peyman Rasouli and Ingrid Chieh Yu. 'CARE: Coherent Actionable Recourse Based on Sound Counterfactual Explanations'. In: *International Journal of Data Science and Analytics* 17 (2022). DOI: 10.1007/s41060-022-00365-6 (cited on pages 14, 111).

[191] Rathenau Instituut. *Governing algorithmic decision-making in government. The role of the Senate.* Tech. rep. Rathenau Instituut, Nov. 2021 (cited on pages 7, 73).

[192] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. 'Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts'. In: *arXiv* (2021) (cited on pages 7, 14, 16, 21, 23, 110).

[193] Kaivalya Rawal and Himabindu Lakkaraju. 'Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses'. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020 (cited on pages 14, 21, 23, 73, 110).

[194] Rechtspraak.nl. *SyRI legislation in breach of European Convention on Human Rights*. Feb. 2020. URL: https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-legislation-in-breach-of-European-Convention-on-Human-Rights.aspx (cited on page 30).

[195] Annabelle Redelmeier, Martin Jullum, Kjersti Aas, and Anders Løland. 'MCCE: Monte Carlo sampling of realistic counterfactual explanations'. In: *Data Mining and Knowledge Discovery*. Springer Nature, 2024, pp. 421–437. DOI: 10.1007/s10618-024-01017-y (cited on pages 14, 16, 22, 112).

[196] Rekenkamer Rotterdam. *gekleurde technologie. verkenning ethisch gebruik algoritmes.* Tech. rep. Rekenkamer Rotterdam, Apr. 2021 (cited on pages 32–35, 45–47, 49, 52, 53).

[197] Jafar Rezaei. 'Best-worst multi-criteria decision-making method'. In: *Omega* 53 (2015), pp. 49–57. DOI: https://doi.org/10.1016/j.omega.2014.11.009 (cited on page 78).

[198] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 'Anchors: high-precision model-agnostic explanations'. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018 (cited on page 73).

[199] Rijksoverheid. *Rijksoverheid*. Subpages therein. URL: https://www.rijksoverheid.nl/ (cited on page 28).

[200] Shalaleh Rismani, Roel Dobbe, and AJung Moon. 'From Silos to Systems: Process-Oriented Hazard Analysis for AI Systems'. In: *arXiv preprint arXiv:2410.22526* (2024) (cited on page 81).

[201] Horst W. J. Rittel and Melvin M. Webber. 'Dilemmas in a general theory of planning'. In: *Policy Sciences* (1973), pp. 155–169 (cited on page 12).

[202] Rockwell Automation. *Arena Simulation Software*. Accessed: 2024-11-04. 2024. URL: https://www.rockwellautomation.com/en-us/products/software/arena-simulation.html (cited on page 72).

[203] Alexis Ross, Himabindu Lakkaraju, and Osbert Bastani. 'Learning models for actionable recourse'. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 18734–18746 (cited on pages 14, 20, 110).

[204] Cynthia Rudin. 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215 (cited on pages 5, 54, 85).

[205] Thomas L. Saaty. 'Decision making with the analytic hierarchy process'. In: *International Journal of Services Sciences* 1.1 (2008), pp. 83–98 (cited on page 78).

[206] Pedram Salimi, Nirmalie Wiratunga, David Corsar, and Anjana Wijekoon. 'Towards Feasible Counterfactual Explanations: A Taxonomy Guided Template-Based NLG Method'. In: *Frontiers in Artificial Intelligence and Applications*. Vol. 372. IOS Press BV, 2023, pp. 2057–2064. DOI: 10.3233/FAIA230499 (cited on pages 14, 20, 22, 112).

[207] Lena Sanders, Denise Pumain, Hélene Mathian, France Guérin-Pace, and Stephane Bura. 'SIMPOP: A Multiagent System for the Study of Urbanism'. In: *Environment and Planning B: Planning and design* 24.2 (1997), pp. 287–305 (cited on page 60).

[208] Mihaela van der Schaar and Andrew Rashbass. *The case for Reality-centric AI*. Accessed 21.05.2024. Feb. 2023. URL: https://www.vanderschaar-lab.com/the-case-for-reality-centric-ai/ (cited on page 24).

[209] Eckard Schindler. *Judicial systems are turning to AI to help manage vast quantities of data and expedite case resolution*. Accessed 10.10.2024. Jan. 2024. URL: https://www.ibm.com/blog/judicial-systems-are-turning-to-ai-to-help-manage-its-vast-quantities-of-data-and-expedite-case-resolution/ (cited on page 1).

[210] Maximilian Schleich, Zixuan Geng, Yihong Zhang, and Dan Suciu. 'GeCo: Quality Counterfactual Explanations in Real Time'. In: *Proc. VLDB Endow.* 14.9 (May 2021), 1681–1693. DOI: 10.14778/3461535.3461555 (cited on pages 14, 20, 110).

[211] Stewart Schlesinger. 'Terminology for model credibility'. In: *Simulation* 32.3 (1979), pp. 103–104 (cited on page 70).

[212] Markus Schlosser. 'Agency'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University, 2019 (cited on page 6).

[213] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. '"There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making'. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1616–1628. DOI: 10.1145/3531146.3533218 (cited on pages 14, 111).

[214] Andrew Selbst and Julia Powles. '"Meaningful Information" and the Right to Explanation'. In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 48–48 (cited on page 30).

[215] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 'Fairness and Abstraction in Sociotechnical Systems'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, 59–68. DOI: 10.1145/3287560.3287598 (cited on page 24).

[216] Concetta Semeraro, Mario Lezoche, Hervé Panetto, and Michele Dassisti. 'Digital twin paradigm: A systematic literature review'. In: *Computers in Industry* 130 (2021), p. 103469. DOI: https://doi.org/10.1016/j.compind.2021.103469 (cited on page 57).

[217] Dunja Šešelja. 'Exploring Scientific Inquiry via Agent-Based Modelling'. In: *Perspectives on Science* 29.4 (Aug. 2021), pp. 537–557. DOI: 10.1162/posc_a_00382 (cited on page 71).

[218] Shubham Sharma, Alan H. Gee, David Paydarfar, and Joydeep Ghosh. 'FaiR-N: Fair and Robust Neural Networks for Structured Data'. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 946–955. DOI: 10.1145/3461702.3462559 (cited on pages 14, 21, 110).

[219] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 'CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 166–172. DOI: 10.1145/3375627.3375812 (cited on pages 14, 19, 21, 110).

[220] Sunny Shree, Jaganmohan Chandrasekaran, Yu Lei, Raghu N. Kacker, and D. Richard Kuhn. 'DeltaExplainer: A Software Debugging Approach to Generating Counterfactual Explanations'. In: *2022 IEEE International Conference On Artificial Intelligence Testing (AITest)*. 2022, pp. 103–110. DOI: 10.1109/AITest55621.2022.00023 (cited on page 44).

[221] Simio LLC. *The Simio Discrete Event Simulation Platform*. Accessed: 2024-11-04. 2024. URL: https://www.simio.com (cited on page 72).

[222] 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. New York, NY, USA: Association for Computing Machinery, 2023, pp. 293–294. DOI: 10.1145/3570991.3571039 (cited on pages 14, 20, 112).

[223] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. 'Directive Explanations for Actionable Explainability in Machine Learning Applications'. In: *ACM Trans. Interact. Intell. Syst.* 13.4 (Dec. 2023). DOI: 10.1145/3579363 (cited on pages 14, 112).

[224] Filippo Santoni de Sio and Jeroen Van den Hoven. 'Meaningful human control over autonomous systems: A philosophical account'. In: *Frontiers in Robotics and AI* 5 (2018), p. 15 (cited on page 85).

[225] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. 'Counterfactual Explanations Can Be Manipulated'. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 62–75 (cited on pages 14, 110).

[226] Bevan I. Smith, Charles Chimedza, and Jacoba H. Bührmann. 'Individualized Help for At-Risk Students Using Model-Agnostic and Counterfactual Explanations'. In: *Education and Information Technologies* 27.2 (Mar. 2022), pp. 1539–1558. doi: 10.1007/s10639-021-10661-6 (cited on pages 14, 24, 111).

[227] Jan-Tobias Sohns, Christoph Garth, and Heike Leitte. 'Decision Boundary Visualization for Counterfactual Reasoning'. In: *Computer Graphics Forum* 42.1 (2023), pp. 7–20. doi: 10.1111/cgf.14650 (cited on pages 14, 21, 112).

[228] Nina Spreitzer, Hinda Haned, and Ilse van der Linden. 'Evaluating the Practicality of Counterfactual Explanations'. In: *CEUR Workshop Proceedings*. Vol. 3277. CEUR-WS, 2022, pp. 31–50 (cited on pages 14, 20, 111).

[229] Laura State, Salvatore Ruggieri, and Franco Turini. 'Reason to Explain: Interactive Contrastive Explanations (REASONX)'. In: *Explainable Artificial Intelligence*. Vol. 1901 CCIS. Cham: Springer Nature Switzerland, 2023, pp. 421–437 (cited on pages 14, 112).

[230] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 'A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence'. In: *IEEE Access* 9 (2021), pp. 11974–12001 (cited on page 6).

[231] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martin Pereira-Fariña. 'Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers'. In: *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2020, pp. 1–8. doi: 10.1109/FUZZ48607.2020.9177629 (cited on page 73).

[232] Muhammad Suffian and Alessandro Bogliolo. 'Investigation and Mitigation of Bias in Explainable AI'. In: *CEUR Workshop Proceedings*. Vol. 3319. CEUR-WS, 2022, pp. 89–94 (cited on pages 14, 19, 21, 111).

[233] Muhammad Suffian, Pierluigi Graziani, Jose M. Alonso, and Alessandro Bogliolo. 'FCE: Feedback Based Counterfactual Explanations for Explainable AI'. In: *IEEE Access* 10 (2022), pp. 72363–72372. doi: 10.1109/ACCESS.2022.3189432 (cited on pages 14, 22, 111).

[234] Emily Sullivan and Philippe Verreault-Julien. 'From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse'. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 712–722. doi: 10.1145/3514094.3534185 (cited on pages 14, 20, 111).

[235] William R. Swartout and Johanna D. Moore. 'Explanation in Second Generation Expert Systems'. In: *Second Generation Expert Systems*. Ed. by Jean-Marc David, Jean-Paul Krivine, and Reid Simmons. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 543–585 (cited on page 73).

[236] Fei Tao, He Zhang, Ang Liu, and A. Y. C. Nee. 'Digital Twin in Industry: State-of-the-Art'. In: *IEEE Transactions on Industrial Informatics* 15.4 (2019), pp. 2405–2415. doi: 10.1109/TII.2018.2873186 (cited on page 57).

[237] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 'Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking'. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 465–474. DOI: 10.1145/3097983.3098039 (cited on pages 6, 24).

[238] Giovanni De Toni, Paolo Viappiani, Stefano Teso, Bruno Lepri, and Andrea Passerini. *Personalized Algorithmic Recourse with Preference Elicitation*. 2024. URL: https://arxiv.org/abs/2205.13743 (cited on page 27).

[239] Melissa Tracy, Magdalena Cerdá, and Katherine M. Keyes. 'Agent-based modeling in public health: current applications and future directions'. In: *Annual review of public health* 39.1 (2018), pp. 77–94 (cited on page 60).

[240] Maria Tsiakmaki and Omiros Ragos. 'A Case Study of Interpretable Counterfactual Explanations for the Task of Predicting Student Academic Performance'. In: *2021 25th International Conference on Circuits, Systems, Communications and Computers (CSCC)*. July 2021, pp. 120–125. DOI: 10.1109/CSCC53858.2021.00029 (cited on pages 14, 24, 110).

[241] Stratis Tsirtsis and Manuel Gomez-Rodriguez. 'Decisions, Counterfactual Explanations and Strategic Behavior'. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020 (cited on pages 14, 21, 110).

[242] UCI Machine Learning Repository. *South German Credit*. UCI Machine Learning Repository. 2019. DOI: 10.24432/C5X89F (cited on page 23).

[243] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. 'Towards Robust and Reliable Algorithmic Recourse'. In: *Advances in Neural Information Processing Systems*. Vol. 20. 2021, pp. 16926–19937 (cited on pages 7, 14, 20, 110).

[244] Berk Ustun, Alexander Spangher, and Yang Liu. 'Actionable Recourse in Linear Classification'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 10–19. DOI: 10.1145/3287560.3287566 (cited on pages 6, 14, 19, 110).

[245] UWV. *Hoelang WW-uitkering*. Apr. 2024. URL: https://www.uwv.nl/nl/ww/hoelang-ww (cited on page 29).

[246] Daniel Vale, Ali El-Sharif, and Muhammed Ali. 'Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law'. In: *AI and Ethics* 2.4 (2022), pp. 815–826 (cited on page 43).

[247] Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. 'An open source machine learning framework for efficient and transparent systematic reviews'. In: *Nature Machine Intelligence* 3.2 (2021), pp. 125–133. DOI: 10.1038/s42256-020-00287-7 (cited on page 15).

[248] Peter M. VanNostrand, Huayi Zhang, Dennis M. Hofmann, and Elke A. Rundensteiner. 'FACET: Robust Counterfactual Explanation Analytics'. In: *Proc. ACM Manag. Data* 1.4 (Dec. 2023). DOI: 10.1145/3626729 (cited on pages 14, 112).

[249] Michael Veale and Reuben Binns. 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data'. In: *Big Data & Society* 4.2 (2017), p. 2053951717743530. DOI: 10.1177/2053951717743530 (cited on page 8).

[250] Suresh Venkatasubramanian and Mark Alfano. 'The Philosophical Basis of Algorithmic Recourse'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 284–293. DOI: 10.1145/3351095.3372876 (cited on pages 6, 7, 9, 14, 18, 19, 25, 110).

[251] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. 'Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review'. In: *arXiv* (2022) (cited on pages 6, 14, 16, 21, 111).

[252] Sahil Verma, Keegan Hines, and John P. Dickerson. 'Amortized Generation of Sequential Algorithmic Recourses for Black-Box Models'. In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*. Vol. 36. Association for the Advancement of Artificial Intelligence, 2022, pp. 8512–8519 (cited on pages 14, 19, 111).

[253] Sahil Verma, Ashudeep Singh, Varich Boonsanong, John P. Dickerson, and Chirag Shah. 'RecRec: Algorithmic Recourse for Recommender Systems'. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 4325–4329. DOI: 10.1145/3583780.3615181 (cited on page 16).

[254] Tom Viering and Marco Loog. 'The Shape of Learning Curves: A Review'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2023), pp. 7799–7819. DOI: 10.1109/TPAMI.2022.3220744 (cited on page 4).

[255] Kilian Vieth-Ditlmann. *The algorithmic administration: automated decision-making in the public sector*. Accessed 22.05.2024. May 2024. URL: https://algorithmwatch.org/en/algorithmic-administration-explained/ (cited on page 1).

[256] Marco Virgolin and Saverio Fracaros. 'On the Robustness of Sparse Counterfactual Explanations to Adverse Perturbations'. In: *Artificial Intelligence* 316.C (Mar. 2023). DOI: 10.1016/j.artint.2022.103840 (cited on pages 14, 19, 112).

[257] 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. New York, NY, USA: Association for Computing Machinery, 2023, pp. 2211–2222. DOI: 10.1145/3580305.3599343 (cited on pages 14, 112).

[258] Julius Von Kugelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Scholkopf. 'On the Fairness of Causal Algorithmic Recourse'. In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*. Vol. 36. Association for the Advancement of Artificial Intelligence, 2022, pp. 9584–9594 (cited on pages 14, 21, 111).

[259] Yvette de Vries. *Gemeente Nissewaard wist al jaren dat fraudeopsporing bijstand niet deugde*. July 2021. URL: https://www.fnv.nl/nieuwsbericht/sectornieuws/uitkeringsgerechtigden/2021/07/gemeente-nissewaard-wist-al-jaren-dat-fraudeopspor (cited on page 31).

[260] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'. In: *International Data Privacy Law* 7.2 (2017), pp. 76–99 (cited on page 30).

[261] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR'. In: *Harvard Journal of Law & Technology* 31 (2017), p. 841 (cited on pages 5–7, 14–16, 68, 110).

[262] Paul Y. Wang, Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 'Demonstration of Generating Explanations for Black-Box Algorithms Using Lewis'. In: *Proc. VLDB Endow.* 14.12 (July 2021), pp. 2787–2790. DOI: 10.14778/3476311.3476345 (cited on pages 14, 21, 110).

[263] Yongjie Wang, Qinxu Ding, Ke Wang, Yue Liu, Xingyu Wu, Jinglong Wang, Yong Liu, and Chunyan Miao. 'The Skyline of Counterfactual Explanations for Machine Learning Decision Models'. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 2030–2039. DOI: 10.1145/3459637.3482397 (cited on pages 14, 110).

[264] 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. New York, NY, USA: Association for Computing Machinery, 2023, pp. 2596–2605. DOI: 10.1145/3583780.3614885 (cited on pages 14, 21, 112).

[265] Zhendong Wang, Isak Samsten, Vasiliki Kougia, and Panagiotis Papapetrou. 'Style-Transfer Counterfactual Explanations: An Application to Mortality Prevention of ICU Patients'. In: *Artif. Intell. Med.* 135.C (Jan. 2023). DOI: 10.1016/j.artmed.2022.102457 (cited on pages 14, 24, 25, 112).

[266] Zijie J. Wang, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. 'GAM Coach: Towards Interactive and User-Centered Algorithmic Recourse'. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. DOI: 10.1145/3544548.3580816 (cited on pages 14, 19, 23, 112).

[267] Greta Warren, Mark T. Keane, and Ruth M. J. Byrne. 'Features of Explainability: How Users Understand Counterfactual and Causal Explanations for Categorical and Continuous Features in XAI'. In: *CEUR Workshop Proceedings*. Vol. 3251. CEUR-WS, 2022 (cited on pages 14, 20, 111).

[268] Greta Warren, Barry Smyth, and Mark T. Keane. '"Better" Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI)'. In: *Case-Based Reasoning Research and Development: 30th International Conference, ICCBR 2022, Nancy, France, September 12–15, 2022, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2022, pp. 63–78. DOI: 10.1007/978-3-031-14923-8_5 (cited on pages 14, 19, 111).

[269] Donald A. Waterman. *A guide to expert systems*. USA: Addison-Wesley Longman Publishing Co., Inc., 1985 (cited on page 73).

[270] Tony Waters and Dagmar Waters. *Weber's rationalism and modern society: New translations on politics, bureaucracy, and social stratification*. Springer, 2015 (cited on page 7).

[271] Hans Weigand, Paul Johannesson, and Birger Andersson. 'An artifact ontology for design science research'. In: *Data & Knowledge Engineering* 133 (2021), p. 101878. DOI: https://doi.org/10.1016/j.datak.2021.101878 (cited on page 12).

[272] Daniel S. Weld and Gagan Bansal. 'The Challenge of Crafting Intelligible Intelligence'. In: *Commun. ACM* 62.6 (May 2019), 70–79. DOI: 10.1145/3282486 (cited on page 5).

[273] Leif Wenar. 'Rights'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University, 2023 (cited on page 8).

[274] Werk.nl. *Dienstverlening bij een bijstandsuitkering*. Mar. 2024. URL: https://www.werk.nl/werkzoekenden/over-werk-nl/dienstverlening/bijstand/ (cited on page 29).

[275] Michael R. Wick. 'Second Generation Expert System Explanation'. In: *Second Generation Expert Systems*. Ed. by Jean-Marc David, Jean-Paul Krivine, and Reid Simmons. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 614–640 (cited on page 73).

[276] Anjana Wijekoon, Nirmalie Wiratunga, Ikechukwu Nkisi-Orji, Kyle Martin, Chamath Palihawadana, and David Corsar. 'Counterfactual Explanations for Student Outcome Prediction with Moodle Footprints'. In: *CEUR Workshop Proceedings*. Vol. 2894. CEUR-WS, 2021, pp. 1–8 (cited on pages 14, 24, 110).

[277] Nirmalie Wiratunga, Anjana Wijekoon, Ikechukwu Nkisi-Orji, Kyle Martin, Chamath Palihawadana, and David Corsar. 'DisCERN: Discovering Counterfactual Explanations Using Relevance Features from Neighbourhoods'. In: *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. Nov. 2021, pp. 1466–1473. DOI: 10.1109/ICTAI52525.2021.00233 (cited on pages 14, 110).

[278] Claes Wohlin. 'Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering'. In: *EASE '14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. EASE '14 (2014). DOI: 10.1145/2601248.2601268 (cited on page 15).

[279] Bang Wong. 'Points of view: Color blindness'. In: *Nature methods* 8 (June 2011), p. 441. DOI: 10.1038/nmeth.1618 (cited on page 12).

[280] Yifan Xu, Joe Collenette, Louise Dennis, and Clare Dixon. 'Dialogue Explanations for Rule-Based AI Systems'. In: *Explainable and Transparent AI and Multi-Agent Systems*. Ed. by Davide Calvaresi, Amro Najjar, Andrea Omicini, Reyhan Aydogan, Rachele Carli, Giovanni Ciatto, Yazan Mualla, and Kary Främling. Cham: Springer Nature Switzerland, 2023, pp. 59–77 (cited on page 73).

[281] Jingquan Yan and Hao Wang. 'Self-Interpretable Time Series Prediction with Counterfactual Explanations'. In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. JMLR.org, 2023 (cited on pages 14, 112).

[282] I-Cheng Yeh. *Default of Credit Card Clients*. UCI Machine Learning Repository. 2016. DOI: 10.24432/C55S3H (cited on page 23).

[283] I-Cheng Yeh and Che-hui Lien. 'The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients'. In: *Expert Systems with Applications* 36.2 (2009), pp. 2473–2480 (cited on page 22).

[284] 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. New York, NY, USA: Association for Computing Machinery, 2023, pp. 742–751. DOI: 10.1145/3600211.3604708 (cited on pages 14, 22, 112).

[285] Songming Zhang, Xiaofeng Chen, Shiping Wen, and Zhongshan Li. 'Density-Based Reliable and Robust Explainer for Counterfactual Explanation'. In: *Expert Syst. Appl.* 226.C (Sept. 2023). DOI: 10.1016/j.eswa.2023.120214 (cited on pages 14, 112).

Appendix

Adapted search engine queries (Ch. 4) A

For ACM Digital Library:

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

For IEEE Xplore:

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

For SCOPUS:

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

Forms of contributions (Ch. 4) B

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

Year	Reference	Propose methods	Theoretical frameworks	Analyses	Apply	Benchmark	Review
2017	[261]	✓	✓				
2019	[97]	✓					
	[114]	✓					
	[151]	✓					
	[158]	✓					
	[244]	✓					
2020	[65]	✓					
	[159]	✓					
	[241]	✓					
	[45]	✓					
	[52]	✓					
	[86]	✓					
	[123]	✓					
	[120]	✓					
	[179]	✓					
	[182]	✓					
	[186]	✓					
	[219]	✓					
	[193]	✓					
	[23]		✓				
	[250]		✓				
2021	[122]	✓	✓				
	[243]	✓		✓			
	[80]	✓					
	[95]	✓					
	[99]	✓					
	[130]	✓					
	[133]	✓					
	[184]	✓					
	[39]	✓					
	[47]	✓					
	[117]	✓					
	[119]	✓					
	[162]	✓					
	[176]	✓					
	[203]	✓					
	[210]	✓					
	[263]	✓					
	[277]	✓					
	[218]	✓					
	[100]		✓				
	[19]			✓			
	[192]			✓			
	[225]			✓			
	[5]				✓		
	[154]				✓		
	[164]				✓		
	[171]				✓		
	[240]				✓		
	[262]				✓		
	[276]				✓		
	[178]					✓	

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

Year	Reference	Propose methods	Theoretical frameworks	Analyses	Apply	Benchmark	Review
2022	[74]	✓		✓			
	[64]	✓		✓			
	[7]	✓					
	[51]	✓					
	[94]	✓					
	[115]	✓					
	[268]	✓					
	[155]	✓					
	[101]	✓					
	[144]	✓					
	[146]	✓					
	[163]	✓					
	[168]	✓					
	[185]	✓					
	[190]	✓					
	[233]	✓					
	[232]	✓					
	[252]	✓					
	[118]	✓					
	[177]		✓	✓			
	[50]		✓		✓		
	[121]		✓				✓
	[27]		✓				
	[30]		✓				
	[169]		✓				
	[213]		✓				
	[234]		✓				
	[267]		✓				
	[228]		✓				
	[258]			✓			
	[56]				✓		
	[188]				✓		
	[226]				✓		
	[93]					✓	✓
	[251]					✓	✓

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

Year	Reference	Propose methods	Theoretical frameworks	Analyses	Apply	Benchmark	Review
2023	[66]	✓	✓				
	[57]	✓	✓				
	[206]	✓	✓				
	[14]	✓		✓			
	[81]	✓		✓			
	[132]	✓		✓			
	[256]	✓		✓			
	[266]	✓			✓		
	[265]	✓			✓		
	[98]	✓					
	[222]	✓					
	[25]	✓					
	[128]	✓					
	[58]	✓					
	[96]	✓					
	[166]	✓					
	[167]	✓					
	[170]	✓					
	[187]	✓					
	[227]	✓					
	[229]	✓					
	[248]	✓					
	[264]	✓					
	[281]	✓					
	[284]	✓					
	[285]	✓					
	[140]	✓					
	[257]	✓					
	[131]	✓					
	[136]		✓				
	[223]		✓				
	[71]			✓			
	[102]			✓			
	[180]			✓			
	[4]				✓		
	[134]				✓		
	[12]					✓	
	[112]					✓	
2024	[46]	✓					
	[195]	✓					

Invitation letter for interviews (Ch. 7) C

Interview on algorithmic decision-making in social welfare processes

Dear ....,

My name is Aleksander Buszydlik and I am a master's thesis student at TU Delft, jointly supervised by Cynthia Liem (Multimedia Computing group at the Faculty of Electrical Engineering, Mathematics, and Computer Science) and Roel Dobbe (Information and Communication Technology group at the Faculty of Technology, Policy, and Management).

I am writing to you to inquire about the possibility of scheduling an interview on the opportunities and barriers for algorithmic decision-making in your domain.

More specifically, my research focuses on the topic of “algorithmic recourse” (AR), or the challenge of generating actionable recommendations for people that received undesirable predictions from algorithmic models, to help them achieve better outcomes in the future. AR is a relatively new problem in computer science, but with the accelerating pace of adoption of algorithms into decision-making systems, it is often seen as a promising way to improve the safety of algorithmic decision-making processes, trust in their decisions, and the agency of the affected individuals. Moreover, AR techniques can offer the decision-makers a way to audit algorithmic models on criteria such as fairness of outcomes.

Algorithmic recourse is most useful when “black box” models are used to support decision-making, and experts cannot easily understand the grounds of a prediction. For example, a person that unsuccessfully applied for social welfare could receive recourse such as “if you have met with the case worker more often, you would have qualified for the benefits”. AR is both a technical and a social process; my research aims to define the requirements to operationalize these mechanisms in real-world contexts.

I am focusing on social welfare, as a domain where the interests of decision-makers and applicants tend to be aligned, and thus where AR could be particularly applicable.

The main goal of my thesis is to decide if algorithmic models could be introduced into existing decision-making processes, while ensuring the safety of the affected individuals, potentially through the means of AR. Thus, I am interested to learn about the functioning of social welfare in your municipality: the existing processes, the involved roles, and the envisioned place for algorithms.

Of course, our discussion can remain at any level of generality that you deem appropriate, such that it does not put the existing decision-making processes at risk. Moreover, my research is approved by the internal ethics committee of TU Delft as minimal risk.

If you are open to dedicating one hour of your time in the coming weeks for an interview or if you can recommend any of your employees to discuss this topic with me, I would be very grateful to receive a message from you at the (email) address above. I will provide you with more details about the interview ahead of time, including a consent form explaining how the outcomes of our discussion would be used in my research.

Thank you for considering my request. If you would like to know anything else before making your decision, please do not hesitate to contact me.

Yours sincerely,
Aleksander Buszydlik
....

❀ THE END ❀