Humans-in-the-Contestability-Loop: Designing for Social Counselors to Understand and Challenge Algorithmic Decisions

TIMOTHÉE SCHMUDE, University of Vienna, Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science, Austria
DANIEL PAHR, University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis, Austria
LAURA KOESTEN, Mohamed bin Zayed University of Artificial Intelligence, Department of Human-Computer Interaction, UAE, and University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis, Austria, and AIT Austrian Institute of Technology GmbH, Center for Technology Experience, Austria
TORSTEN MÖLLER, University of Vienna, Faculty of Computer Science, Research Network Data Science, Research Group Visualization and Data Analysis, Austria
SEBASTIAN TSCHIATSCHEK, University of Vienna, Faculty of Computer Science, Research Network Data Science, Research Group Data Mining and Machine Learning, Austria

Making algorithmic decisions explainable and contestable is essential to ensure the responsible use of high-risk AI systems. Yet the human-centered design and implementation of explainability and contestability remain underexplored in empirical research. In this paper, we follow a participatory design approach in three stages to develop an interface that supports social counselors in understanding and contesting a welfare fraud detection system. We analyze counselors' information needs, derive functional and usability requirements, and design and evaluate both low-fidelity and high-fidelity interfaces that implement these requirements. Our findings show that contesting administrative decisions is a complex socio-technical process involving social counselors, clients, and social agencies, requiring carefully tuned explanation and contestation elements. Social counselors prioritized procedural information and tools for interactive exploration of model predictions, while emphasizing that explanations must be concise, easy to understand, and sensitive to the emotional impact of information on clients. We find that human intervention remains crucial in contestation processes, that explanations can support counselors by surfacing contestation reasons and facilitating client-side communication, and that conversational explanations enable flexible inquiry. Our insights contribute to trustworthy AI implementations by outlining the potential and challenges of a human-centered explanation and contestation interface for individuals affected by high-risk AI systems.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**.

Authors' Contact Information: Timothée Schmude, timothee.schmude@univie.ac.at, University of Vienna, Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science, Vienna, Austria; Daniel Pahr, daniel.pahr@univie.ac.at, University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis, Vienna, Austria; Laura Koesten, laura.koesten@univie.ac.at, Mohamed bin Zayed University of Artificial Intelligence, Department of Human-Computer Interaction, Abu Dhabi, UAE and and University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis, Vienna, Austria and AIT Austrian Institute of Technology GmbH, Center for Technology Experience, Vienna, Austria; Torsten Möller, torsten.moeller@univie.ac.at, University of Vienna, Faculty of Computer Science, Research Network Data Science, Research Group Visualization and Data Analysis, Vienna, Austria; Sebastian Tschiatschek, sebastian.tschiatschek@univie.ac.at, University of Vienna, Faculty of Computer Science, Research Network Data Science, Research Group Data Mining and Machine Learning, Vienna, Austria.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812271

Additional Key Words and Phrases: explainability, contestability, participatory design, social counselors, welfare fraud detection, public sector AI, high-risk AI systems, socio-technical systems, information needs, usability and functional requirements, interactive explanations, conversational explanations, interface design and evaluation, qualitative methods

ACM Reference Format:
Timothée Schmude, Daniel Pahr, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2026. Humans-in-the-Contestability-Loop: Designing for Social Counselors to Understand and Challenge Algorithmic Decisions. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 35 pages. https://doi.org/10.1145/3805689.3812271

1 Introduction

Algorithmic decision-making (ADM) systems are increasingly used in public institutions to make critical decisions about individuals. These systems rely on data processing and statistical inference to derive information that is deemed useful for decision-making [45]. Examples of such systems were used or piloted for employability scoring [73, 82], recidivism prediction [23], and welfare fraud detection [87]. In these areas, individuals subjected to institutional decisions are vulnerable to harms from erroneous, unfair, or unchecked algorithmic decisions [7]. To reduce these harms, ethical [46] and regulatory [43] frameworks set trustworthiness requirements for ADM systems, including preserving human agency, transparency, and accountability [2]. Explainability and contestability are design principles that can realize these requirements [77] by making systems understandable [6] and responsive to dispute [4], but their implementation in specific contexts remains an open research area.

This study explores the design of explainability and contestability in the context of ADM systems in the public sector, motivated by a real ADM system in the French social insurance that scores welfare recipients' fraud risk [1, 87]. This system serves as an example of a high-risk public sector ADM system, which has been reported to produce automation-driven intersectional discrimination [73], lasting algorithmic imprints on institutions and society [40], and adverse impacts on large groups from faulty decisions [16]. These experiences raise questions about how society can build capacities for explaining and contesting potentially harmful algorithmic decisions, as described in previous work as *contestability loops* [5]. *Social support organizations* are one likely location for these capacities: civil society institutions that provide welfare services such as housing, financial, and asylum assistance [56, 117]. Social counselors in these organizations are in direct contact with individuals impacted by administrative decisions and represent their clients before state agencies [58]. In the context of contesting ADM systems, this gives them a pivotal position as *humans-in-the-contestability-loop* [5, 30].

The roles of social counselors can be compared to those of social workers [57], case workers [19], and frontline workers [58] in that they provide advocacy, bureaucratic navigation, and client support. Similar roles are present in various forms of civil society organizations that make up non-governmental social security networks in welfare states [57]. Our work considers social counselors as domain experts for social work and administration, with tasks composed of advisory, emotional, and procedural support. Examples of their work in the context of welfare benefits include advising clients on their entitlements and assisting with applications, completing and submitting documents to state agencies, offering emotional support during times of hardship, and resolving issues with state agencies, such as missed deadlines or incomplete documentation.

While previous work in human-centered explainable AI [41] has explored design for developers [71] and lay audiences [81] or *AI novices* [78], the roles and requirements of mediating actors such as social counselors are rarely explored in detail. Further, while contestability is studied in a body of conceptual [26, 49, 56] and legal [54, 77] research, few studies advance concrete design solutions situated in the context of public institutions [3, 76, 116]. In this work, we aim to address this gap by answering the following research questions:

[RQ1] *Context*: How do individuals affected by algorithmic systems (social workers and decision subjects) perceive contestation, and what information do they need to act on it?
[RQ2] *Requirements*: What are the design requirements for information hierarchy, explanation, and contestation elements that help social counselors understand and contest algorithmic decisions?
[RQ3] *Design*: How should information hierarchy, explanation elements, and contestation features be designed to meet counselors' requirements?
[RQ4] *Effect*: How do conversational, regular, and no explanations impact counselors' ability to contest algorithmic decisions and influence their understanding, agency, contestation choice, and confidence?

To address these questions, we follow a participatory design approach [85] to develop and test an explanation and contestation interface that supports counselors in understanding, explaining, and challenging algorithmic decisions. The study proceeds in three stages: I. analysis of requirements and context of use, II. design and testing of low-fidelity prototypes, and III. development and evaluation of a high-fidelity interface. Each stage is realized as a task-based interview with social counselors and a client recruited from Volkshilfe Vienna, a social support organization, and analyzed for participants' understanding and contestation experience (Figure 1).

Our findings provide a detailed characterization of counselors' roles as mediators for their clients, whom we identify as *served users*, and highlight counselors' need for brief and easy-to-understand explanations and justifications. The discussion provides four lessons from explanation and contestation design in the public sector, including the importance of considering social constellations, the role of justifications in supporting collective contestation, the challenges of participatory approaches in multi-stakeholder contexts, and the inclusion of legal information within the scope of XAI. We envision these insights to contribute to the realization of trustworthy AI in concrete public sector contexts, and to provoke further discussion on the role of explainability and contestability in making high-risk AI systems understandable and responsive to dispute, particularly for AI novices.

2 Background and Related Work

This section introduces key concepts relevant to the study and describes related work, including: Algorithmic decision-making systems in public institutions (2.1), concept and design foundations of contestability (2.2) as well as explainability (2.3), and the relevance of the trustworthy AI *policy knot* for this study (2.4).

2.1 Algorithmic decision-making in public institutions: The CAF algorithm

The European Parliament defines algorithmic decision-making (ADM) systems as systems that assist human decision-making by using data processing, machine learning, or rule-based reasoning [45]. We use 'AI system' synonymously to capture systems with symbolic and subsymbolic learning components, following XAI, societal, and regulatory usage [67, 83, 117]. This study focuses on an ADM used by *Caisse d'allocations familiale* (CAF), part of France's social security [88], which uses logistic regression to predict welfare fraud likelihood from household investigation data labeled with overpayments. About 13 million households (roughly half of France) are regularly monitored, and 100,000 are flagged for investigation each year [87]. In October 2024, civil society organizations filed charges alleging discrimination against marginalized groups and ineffectiveness [1]. Similar systems exist in Poland [82], the Netherlands [34], and the US [23], making the CAF algorithm a current and generalizable class of public-sector ADM systems. More details about the CAF algorithm are included in the appendix.

2.2 Contestability of algorithmic decisions and design for contestation

Contestability makes systems responsive to dispute [3] by enabling actors (e.g., controllers, decision subjects) to "understand, construct, shape, and challenge" predictions [59]. It aims to increase systems' legitimacy over time [5] and surface embedded values [59]. Contestation can be realized through interactive controls to change inputs or challenge decisions [5], human intervention constellations [30], and assessments and audits [109],

and can be enacted via judicial pathways (appeals, courts) and non-judicial pathways (social or institution-internal interventions) [75]. Contestation mechanisms must fit the intended user audience and domain, as users' preferences and information needs can depend on their AI literacy and prior fairness experience [114]. For example, social platforms may afford contestation through moderation [108], whereas public administrations may rely on fair hearings and judicial review [75]. To this end, human-centered and participatory design methods are a viable approach that has been used in previous work on contestation design [3, 5], also motivating their use in this study. While previous work further explored the conceptual intersection of explainability and contestability [49, 77, 113], their practical intersection remains underexplored, a gap that this study aims to address.

2.3 Designing explanations for AI novices and decision subjects

Explainability means making an AI system understandable in its logic, input features, and purpose [67]. It can be an inherent attribute in interpretable *white-box* models) or provided through post-hoc explanation for non-interpretable *black-box* models [92]. Explanations need to be adapted to user needs [67] in scope (local vs. global), and format (text, visuals, numbers) [103], and may also cover a system's deployment context, including its purpose and embedded values [96]. Compared to explanations for technical experts [6, 79, 103], explanations for lay audiences or *AI novices* [78] are a comparatively new addition to XAI. Tools like the Explainer dashboard [35] offer many methods but are not usually designed for or evaluated with non-experts, creating a "sociotechnical gap" [39], which can be addressed through question-driven [71], participatory [69], and collaborative design [111]. Conversational explanations from LLMs [102] can help AI novices in reaching accurate answers [102] and in reasoning [22]. While prior XAI has designed explanations for non-experts in auditing [66], fairness assessments [81], and deployment decisions [96], few works examine how explanations support AI novices in contesting algorithmic decisions, which is the focus of this study.

2.4 The trustworthy AI policy knot

Policy texts like the EU AI Act [43], Digital Services Act (DSA) [8], GDPR [44], and the proposed AI Bill of Rights [51] aim to ensure high-risk ADM systems are deployed as "trustworthy" [106], respecting human agency and oversight, privacy, non-discrimination, accountability, and transparency [46]. Realizing and safeguarding these principles spans many domains, forming a *policy knot* [53] of legal, institutional, and academic practices. This paper examines the AI policy knot by exploring the realization of trustworthy AI via a participatory design process [85]. Prior work has analyzed the EU AI Act's implications for explainability [83], the role of contestability in procedural justice [68] and society [5], and their regulatory intersection [77]. However, to our knowledge, our study is novel in its use of participatory design to examine the implementation of an explainability and contestability interface in a high-risk public sector context.

3 Methods

This work follows a three-stage participatory design process. This section describes the participant population and recruitment (3.1), the three stages of the overall research and study process (3.2), and summarizes the analysis metrics and methods (3.3). The authors' university's Research Ethics Committee approved this study.

3.1 Participants

All participants in this study were AI novices [78], i.e., stakeholders with no knowledge of the technical foundations of AI systems. They were also affected stakeholders [67], i.e., *decision subjects* who would be affected by algorithmic welfare decisions or *social counselors* who would be responsible for those directly affected (social workers, social pedagogues, adult representatives). Both attributes were elicited through a screening survey that included questions about AI self-efficacy [52], participants' occupation, and their experience with welfare benefits. The

aim of this selection was to explore the lived experiences of individuals in direct contact with the social agency in order to understand the layers of administration and the effect of an ADM system in this context (RQ1-Context), as well as the issues that an explanation and contestation interface could address and, importantly, not address (RQ2-Requirements, RQ3-Design). Table 1 provides an overview of participants' occupation and participation in the three study stages.

Participants were recruited from social support organizations that offer free services to people in need (e.g., welfare recipients, job-seekers, refugees). Social counselors, who are in close contact with the social agency, served as our main informants. Although the study initially focused on clients, counselors anticipated low responsiveness and limited capacity, which was confirmed by recruitment. We therefore focused on counselors, who would represent and support clients if they were algorithmic decision subjects, leading us to frame decision subjects as *served users* who would not use the system directly but would be affected by its use [37]. The sample size followed guidance from formative user testing [107] and from meaning and code saturation [50].

Table 1. Information on the participants and their involvement in the three study stages. Column III details which high-fidelity prototype participants interacted with: regular explanations (A-flow) or added conversational explanations (A-LLM). *P1 and P2 participated together.

ID	Age range	Occupation	I	II	III	ID	Age range	Occupation	I	II	III
P1*	50-59	Client	•	•		P7	30-39	Social worker	•	•	A-LLM
P2*	50-59	Social worker	•	•	A-LLM	P8	30-39	Adult representative		•	A-flow
P3	50-59	Adult representative	•	•	A-LLM	P9	50-59	Social pedagogue			A-LLM
P4	40-49	Social worker	•	•	A-flow	P10	20-29	Ombudsperson			A-flow
P5	30-39	Social worker	•	•	A-flow	P11	30-39	Social worker			A-LLM
P6	30-39	Social worker	•			P12	40-49	Social worker			A-flow

3.2 Research process: Participatory design in three stages

Participatory design (PD) emphasizes cooperation, democracy, mutual learning, and creativity [85]. Acknowledging PD as nuanced and on a spectrum [110], we briefly outline the participatory aspects of our approach. We adopt Bødker et al. [13] and Delgado et al. [32]'s view of PD: engaging people in designing future technology by prioritizing self-determination and agency [112] to influence systems that affect their work and daily life. Prior work further highlights the need for situating PD in specific contexts [48]. We incorporate these principles by focusing on a high-risk public-sector AI use case and aligning the explanation and contestation interface with the values and interests of social counselors and clients. Per Wacnik et al. [110], we describe our design process as *emergent* (responsive to user needs), emphasizing *direct* participation, engaging users *throughout* the process, relying on *iterative* activities, and using *multiple* techniques (personas, scenarios, interviews, prototyping). Unlike user-centered approaches that focus on usability metrics (effectiveness, efficiency, satisfaction) [18, 91], PD also challenges power structures and enables meaningful non-expert contributions [32], a core aspect of this study. Thus, our process aligns with most properties regarded as essential to PD, with potential limitations discussed in Section 7.2. The design process is structured along the following three design stages [99]:

Stage I*: Understanding context of use and analyzing requirements.* Stage I served to understand social counselors' relation to their clients and the social agency, their reaction to the algorithmic welfare fraud detection system, and their information needs and requirements with respect to the interface. In a first interview, social counselors described their work context, reflected on possible reasons and pathways for contesting decisions made by a state agency, listed questions about the ADM system, and created initial sketches of possible interfaces. The product of Stage I was a *requirements definition* describing counselors' context of use for an explanation and contestation interface, its functional requirements, as well as considerations and constraints regarding its design [89].

***Stage II**: Low-fidelity prototyping and formative evaluation.* Stage II aimed to evaluate the functionality and usability of the low-fidelity (provisional, unfinished) interfaces that were created through *parallel* (multiple designs at once) and *iterative* (sequential refining) prototyping [21]. The low-fidelity prototypes were developed and tested in five iterations with peers experienced in interaction design, resulting in three final interactive wireframe designs. The prototypes incorporated elements that supported explanation and contestation, such as a risk calculator, decision procedure flowchart, FAQs, and explanatory videos. In the formative evaluation, participants discussed and evaluated the requirement analysis (Figure 3) and the three low-fidelity prototypes in changing order. The user feedback about the prototypes' functionality and usability then guided the development of the high-fidelity versions.

***Stage III**: High-fidelity prototyping and summative evaluation.* Stage III evaluated a high-fidelity prototype in three versions: the baseline (contestation features, no explanations), A-flow (regular explanations), and A-LLM (A-flow plus chatbot-based conversational explanations). The interface was built with TypeScript/React and GitHub Copilot.[1] The chatbot used GPT-5 with retrieval-augmented generation (details in the appendix). For the summative evaluation, participants contested a fictional decision flagging a fictional client for welfare fraud. They first used the baseline (B), then A-flow or A-LLM. After B and again after A-flow/A-LLM, participants made contestation decisions and reported contestation confidence, and feelings of self-determination while using the interface [11]. Finally, they answered interview questions about their understanding of the system, helpful explanation elements, and, if applicable, chatbot interactions.

3.3 Analysis

To address the research questions introduced in Section 1, we employed a range of analytical methods, which always included inductive or deductive thematic analysis of participants' responses. A detailed summary is depicted in Figure 1.

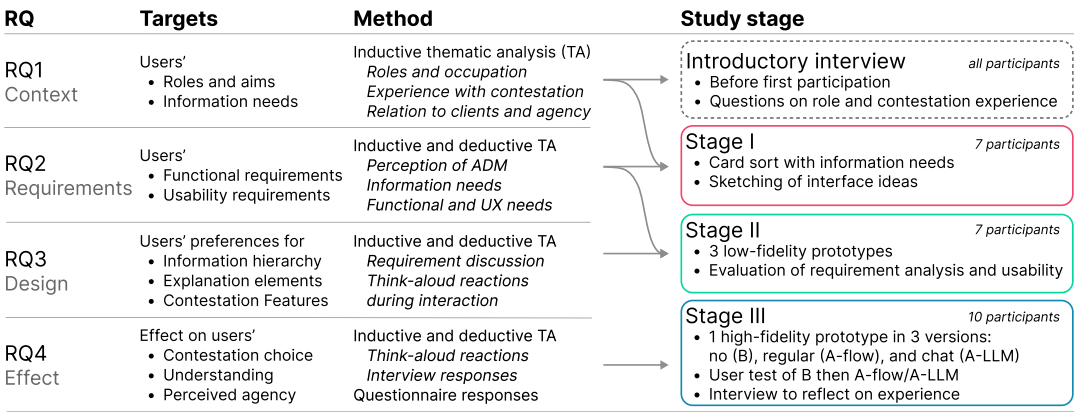

Fig. 1. **Research questions, analysis targets, methods, and study stages.** We used multiple analysis approaches to answer our research questions. For RQ1–Context, we analyzed counselors' responses about their roles and contestation experience in the introductory interviews. For RQ2–Requirements, we conducted a card-sort [93] to elicit and compare their information needs, and used a sketching task as a prompt to discuss the interface requirements. For RQ3–Design, we analyzed participants' think-aloud interaction with the three low-fidelity prototypes to identify useful and superfluous elements. For RQ4–Effect, we analyzed counselors' think-aloud interaction with the high-fidelity prototype, their perceived contestation confidence and sense of competence and autonomy [84], and their responses to a semi-structured interview.

[1]The code of the interface is available under github.com/Stimmot/Explanation_and_Contestation_Interface.git

4 Stage I: Context of use and requirement analysis

This section describes social counselors' role, their relationship with clients and the social agency, and their experience in contesting administrative decisions. It further outlines their informational, functional, and usability requirements for the interface. The findings are compiled into structured personas and a requirement analysis that serves as a starting point for the low-fidelity design (Figure 3). This section is the result of **RQ1-Context** and **RQ2-Requirements** (Figure 1). Quotes and study material were translated from German into English.

4.1 How do counselors and decision subjects handle contestation of administrative decisions?

Institutional context. All participants were affiliated with or employed by social support organizations: non-profit welfare institutions providing counseling, housing, and financial assistance. The study focuses on the organization's branch assisting clients with housing and social insurance, which can also cover labor, asylum, and psychological treatment. Social counselors' work is shaped by interactions with the welfare agency and clients. Because their work relies on navigating administrative structures via negotiations with human agents, they saw the hypothetical ADM system as an added barrier to their workflows and client representation.

Relation between counselors and clients. Social counselors assist clients with housing, agency interactions, legal disputes, and welfare benefit applications. A focus is "*financial security and the submission of applications*" (P5), as interactions with the social agency can be anxiety-inducing and complex for recipients: "*I was scared, I didn't dare to make a claim*" (P1). Clients' reliance on assistance ranges from occasional requests to delegating all administrative tasks. One such administrative task is handling notices informing clients of benefit approvals or rejections. These notices are written in highly formal, bureaucratic language, spanning several double-printed pages with tables, assessment bases, and formalized justifications. Counselors criticized them as "*not easy to understand*" (P3), opaque, and unhelpful for fair hearings or contestation. Clients whose applications for welfare benefits are rejected face severe consequences, including frozen payments and, in suspected fraud cases, lawsuits. Consequently, counselors prioritize timely resolution when rejection notices arrive and help clients understand and respond. To this end, they contact the agency directly, coordinate with teams, and file appeals.

Relation between social support organization and social agency. The relation between the state and support organizations is based on cooperation, but can become tense. Quarterly network meetings of all social organizations' management teams serve to discuss new procedures, foster mutual understanding, and address issues. On the operational level, relations are defined by conflict over recipients' claims. Counselors "*are prepared to fight battles*" (P11) and act as advocates for their clients by promoting their needs to the social agency. This work is complicated by information asymmetry. For example, the social agency may rely on case files (e.g., bank statements, police records) while citing data-privacy rules to limit counselors' advocacy for clients. This institutional opacity is well documented in prior work [73, 98, 117] and is perceived by counselors as creating a power imbalance and reducing options for contestation. While participants were not opposed to automating administration, many doubted that an ADM system could be responsibly deployed in these circumstances.

Participants' experience with contestation. Counselors contest and escalate issues via direct contact with the social agency, internal discussion, cooperation with other support organizations, and, lastly, legal appeals. Non-judicial contestation (contacting caseworkers, internal discussion) is relatively frequent, but legal steps are rarer and more critical. Counselors view legal action as serious due to fees and the burden of tracking and attending proceedings. Such cases are usually backed by institutional support and used to scrutinize broader aspects of the agency's decision-making. Contestation can shift from individual, non-judicial matters to collective, judicial affairs depending on the issues' severity and precedent. We map participants' ADM contestation reasons to categories from prior work [115]: lack of legitimate proof (no proven misconduct), technical inaccuracy, discrimination (stereotype-based decisions), lack of human involvement, and data-privacy violations. These reasons informed the contestation features in the low-fidelity prototypes.

4.2 Summarizing counselors' context of use and compiling the requirement analysis

This subsection summarizes the insights from Stage I and presents the requirement analysis in Figure 3.

Information needs. We categorize the information needs sourced from the card sorting with social counselors by procedural, technical, and legal dimensions, following prior work [49, 115]. Figure 2 provides an overview.

Prioritizing contestation. When receiving a negative welfare decision, social counselors and their clients want to resolve the situation "*immediately, [...] the faster you act, the better*" (P2). Contestation features should thus be available from the beginning, together with information about the decision subjects' rights and possible grounds for contestation.

Providing justifications. Counselors' questions frequently aimed for justifications, meaning information about the norms and values guiding a system [49, 115]. Justifications were requested for, e.g., the system's design rationale, the choice of threshold values, and the principles governing deployment, blending system, procedural, and legal aspects. Designers must decide whether legal and procedural information falls within the scope of XAI.

Many decision subjects are served users. Counselors work as mediators between the social agency and their clients. As clients are accustomed to delegating issues to counselors and may have psychological or physical impediments, they may face challenges in interacting with an interface independently and hence are not direct but *served* users. We thus position the interface design to support social counselors to a) learn about the system in more detail for their own understanding, and b) explain to clients the decision basis and show them available contestation options.

Procedural	Total: 16	Technical	Total: 42	Legal	Total: 6
• Is the decision checked by a human? • Which options for contestation are there and what are the risks involved? • What happens in this process? • On which basis is this claim made in the decision? • How much time is there to contest? • Can someone support me in contesting the decision? • Who is responsible for the system? • Why is this system used? • Who developed this system?		• How does the system work? • How are the features compiled into a risk score? • Is risk increased if the recipient is a single parent? • Why does the number of children influence the risk? • Are there any practical examples to see how the system evaluates someone? • How can risk be reduced? • What does the system know about the recipient? • How does the system receive its data? • Are there any biases in the system? • Is it possible that the results are inaccurate?		• Which rights do I have? • On which law is this decision based? • Which options of legal appeal do I have? • To which benefits am I entitled? • Can the benefits be stopped on the basis of such a decision?	

Fig. 2. **Information needs.** Most of the participants' questions focused on technical aspects such as the system's logic, local prediction instances, counterfactuals, and accuracy. Procedural questions were asked about human involvement, contestation options, risks, and responsibility. Legal questions pertained to rights and underlying laws. The appendix provides the full list.

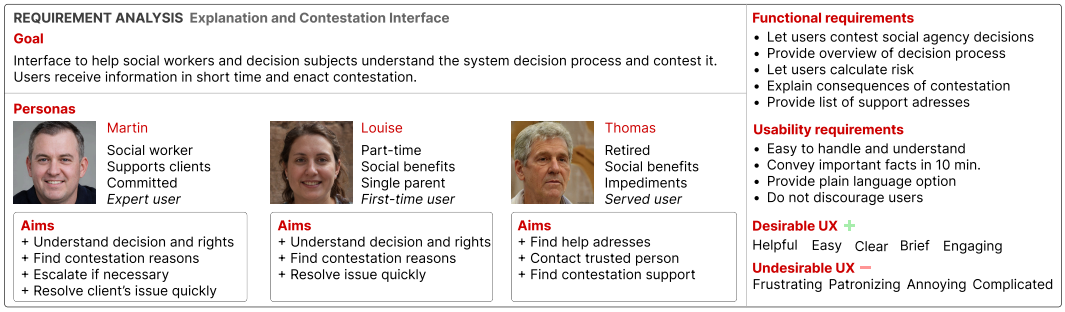

Fig. 3. **Requirement analysis.** The requirement analysis served as the basis for the prototype designs, focusing on three personas: Martin, as a social counselor, and Louise and Thomas, as decision subjects. Thomas is a served user who does not interact with the interface himself; instead, he receives support from Martin, the primary user.

5 Stage II: Low-fidelity prototypes of three explanation and contestation interfaces

This section outlines the design and evaluation of three low-fidelity prototypes to validate requirements and assess whether information needs were met. This section is guided by **RQ3-Design** (Figure 1).

5.1 Low-fidelity prototype design process

We developed low-fidelity interfaces through five iterations and pilot-tested them with peers [99] to refine design and usability. The final three prototypes were interactive wireframes built with Balsamiq [104]. We reviewed existing explanation and contestation designs to identify elements that could meet users' requirements (Figure 3). These elements included process-centric explanations [113], local explanations and counterfactuals [69, 86, 97], system performance depictions [100], and conversational explanations [102]; a detailed list is included in the appendix. These approaches provided intuition for design directions, but most explanation designs achieved a level of detail beyond counselors' requirements: "*It should be helpful, brief, and easy*" (P2). Prior work examined contestation conceptually [5, 75, 109] and theoretically [49, 54], but design solutions for AI novices are underexplored. The prototyping process thus aimed to refine explanation elements to fit social counselors' needs and to draft and iterate features for selecting contestation reasons and pathways. Examples of low-fidelity designs are depicted in Figure 4.

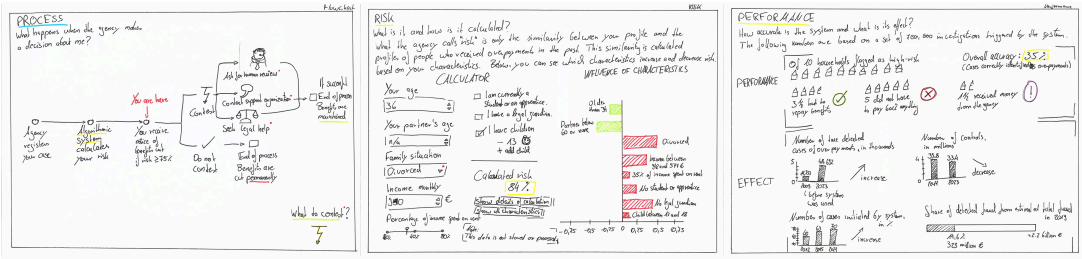

Fig. 4. **Examples of low-fidelity designs.** Single panels taken from five iterations of paper-based low-fidelity interface designs depicting flowchart (left), risk calculator (middle), and system performance (right). See the appendix for additional examples.

5.2 Formative evaluation: Participants' feedback to the low-fidelity prototypes

We tested three wireframe interfaces in the formative evaluation: *Manual and Checklist*, *Chat assist*, and *Builder*. All had four key explanation elements (flowchart, calculator, performance, and feature weight table), six contestation reasons (lack of proof, no mention of algorithmic system, no human intervention, data privacy violation, incorrect decision, and discrimination), and three pathways (write to social agency, contact support organization, file legal appeals). The interfaces differed in their information hierarchies and in how they handled contestation. *Manual and Checklist* featured a landing page with two main options, "Investigation" (explanations) and "Action", where users could select from a reason checklist to receive a suggestion for a contestation pathway. *Chat Assist* featured a chat that flexibly answered questions within a window where users could ask questions freely and use an automated analysis of their notices to receive reason and pathway suggestions. Explanation elements were linked in the chat responses. *Builder* directly presented users with an overview of contestation reasons and pathways and linked to explanation elements as additional information, users could save reasons and pathways into a list and generate a recommendation on how to proceed.

Information hierarchy and navigation. All participants appreciated the split into information and contestation as a structuring element in the *manual and checklist* interface, stating that it was easy to understand

while always leaving open a direct path to contestation. The navigation through conversation in *chat assist* was perceived by many as cumbersome and an impediment to accessing relevant information, and the saving list functionality of the *builder* interface was seen by several participants as unnecessarily complicated. We thus adopted the "Investigation" vs. "Action" information structure as the main navigation for the high-fidelity designs.

Explanation elements. All participants appreciated the flowchart for its explanation of procedures and consequences. The risk calculator was seen as useful for exploring specific decision cases and computing counterfactuals, while the detailed table of feature weights was seen as too much, yet also helpful for escalating contestation. Several counselors worried that disclosing the calculation logic could burden clients, which is explored in detail in Section 6.2. The system's true positive and false positive rates were seen as a good contestation reason and motivated counselors to act on it: "*If that's true, that's awful.*" (P5) All four explanation elements were adopted and refined for the high-fidelity prototypes.

Contestation features. Participants appreciated the contestation recommendations present in all three interfaces, i.e., letter templates, support contacts, and automated notice analysis. However, participants stated that an appeal must be provable to the social agency and doubted that automated analysis of notices would uphold clients' privacy. The selection of contestation reasons was deemed sensible, except for the data privacy violation, which was seen as too difficult to prove. This reason was thus exchanged for one stating that the system used outdated information.

5.3 Key findings of the low-fidelity prototype evaluation

This subsection summarizes the Stage II findings that form the basis for the high-fidelity designs.

Process first, model second. The flowchart was unanimously considered helpful to understand the possible courses of action and situate information about the ADM system. Following this feedback, the flowchart was, therefore, positioned as the first explanation element in the high-fidelity designs.

Contestation is not risk-free. Participants emphasized that the interface should inform about the risks and consequences of contesting, such as the freezing of benefit payments during a pending appeal. Counselors further commented that legal appeals are "*a considerable barrier for many people*" (P10) due to the high knowledge threshold and financial burdens, and that they should be recommended cautiously. We found a tension in the design of contestation features, as counselors cautioned that contestation should be quickly and directly accessible, yet not so effortless as to trivialize the action. In the high-fidelity prototypes, we resolved this tension by requiring users to acknowledge information about their rights and the consequences of contestation before proceeding to the contestation options.

Chatbot introduces interaction challenges. Participants appreciated the option for flexible inquiry, but found the navigation confusing and had difficulties formulating questions. We decided to incorporate conversational explanations in the form of a chat sidebar in the high-fidelity designs, which counselors could use to ask questions that the regular explanations could not answer, but also ignore without issue.

6 Stage III: Design and evaluation of the high-fidelity prototype

This section describes the development of a converged high-fidelity interface and the evaluation with 10 social counselors, focusing on the role of human intervention, the use of explanations for contestation decisions, and the advantages and drawbacks of conversational explanations. This stage was guided by **RQ4-Effect** (Figure 1).

6.1 Design of high-fidelity interface versions and interview process

The high-fidelity prototype featured three versions: a baseline interface with contestation features but no explanations (B), an interface with explanation elements including flowchart, risk calculator, system performance, and detailed feature weights (A-flow), and an interface with an added chat interface for conversational explanations

(A-LLM). Both A-flow and A-LLM also contained the contestation features of B. The high-fidelity interface included all explanation and contestation elements perceived as useful by participants in Stage II. For A-LLM, the system prompt framed the chatbot as a web-based support agent helping recipients understand and contest algorithmic welfare decisions through explanations, information on their rights, and guidance for contestation. The three interface versions were tested and iterated in two pilot studies with peers experienced in interaction design to spot errors and refine usability.

For the evaluation, participants again contested a fictional algorithmic welfare decision for a client. They first interacted with B, then proceeded to either A-flow or A-LLM, alternating according to interview order. After each interaction with B and A-flow/A-LLM, participants selected from six contestation reasons and four pathways, reported their decision confidence, understanding, and rated their sense of autonomy and competence [84]. This split allowed for a two-way comparison: no explanations (B) vs. explanations (A-flow/A-LLM), and regular explanations (A-flow) vs. conversational explanations (A-LLM). The three interface versions are depicted in Figure 5, and the interview procedure is depicted in Figure 6. The appendix provides more information on the interview procedure and chatbot configuration.

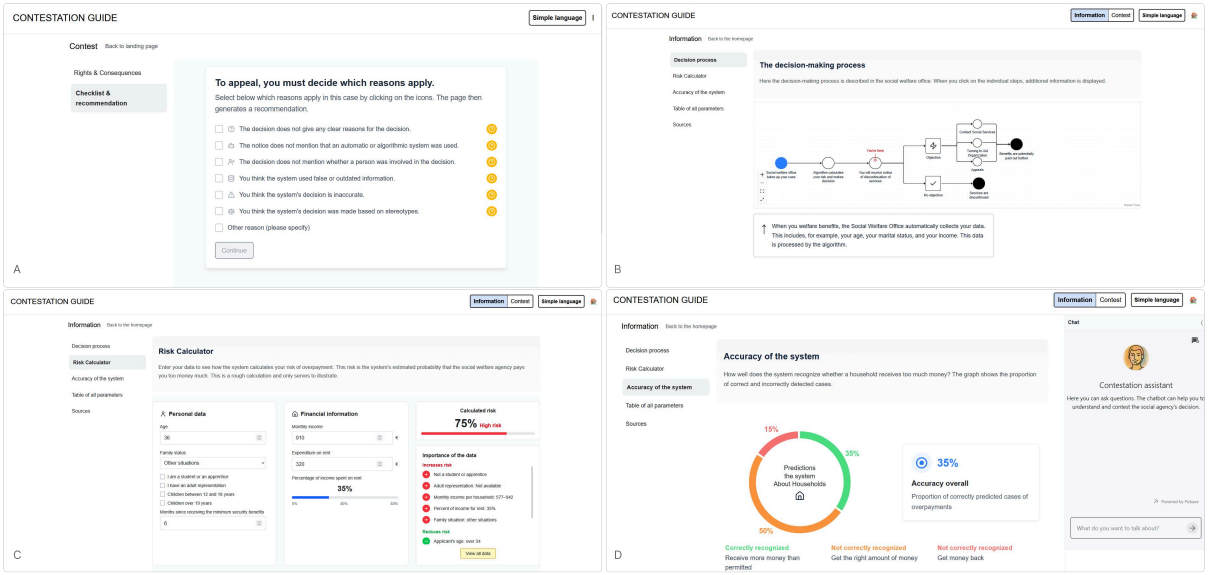

Fig. 5. **Views of the high-fidelity interface.** The baseline interface (A) only included brief descriptions of decision subjects' rights, and a selection of contestation reasons and channels. Interface A-flow included four explanation elements, such as a flowchart (B) and an interactive risk calculator (C). Interface A-LLM (D) additionally provided a chat sidebar for flexible inquiry.

6.2 Summative evaluation: Counselors' understanding, sense of agency, and contestation choices

6.2.1 Human intervention is an essential element of explainability and contestability. Insights from Stage I and Stage II demonstrate that social counselors act as mediators between social agencies and decision subjects, representing clients and navigating the administrative structures of the welfare services: "*We are ourselves like mini-computers*" (P2). Explanations and contestation features must thus be integrated into counselors' work routines, who act as embodied safeguards for transparency, human autonomy, and oversight. Social counselors

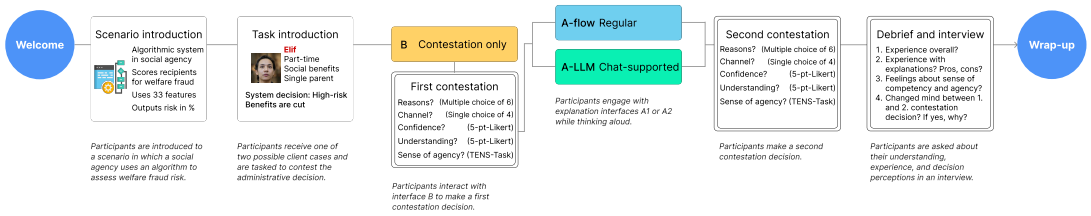

Fig. 6. **Stage III: Interview procedure for the evaluation of the high-fidelity prototypes.**

emphasized that, although the interfaces' contestation features were deemed helpful in reducing workload and speeding up processes, the automation of the contestation process was seen critically. P9 commented: "*If I respond with an automated letter, I can just let the two AIs figure out between them what makes sense or is correct. [...] We shirk our responsibility as humans.*" Importantly, this aversion was two-sided. It pertained both to the logic of making algorithmic welfare decisions about individuals, which participants fundamentally disagreed with, and to the necessity of understanding and tracing this algorithmic logic to build a contestation case, which was described as "*depressing*" (P10). Consequently, participants emphasized two major points in the design of contestation interfaces: prioritizing features that ensure a fair and personal hearing of decision subjects, and providing clients with points for human contact.

6.2.2 Counselors filter information for relevance and emotional impact on decision subjects. Social counselors assume a position of responsibility and control, as they must understand and resolve administrative issues while also selecting the information to show to clients. This produced a two-faceted interaction in our study: Counselors judged each interface element for its contribution to their own understanding, and for its use in client-side explanations. All participants appreciated the flowchart for providing information about the decision procedure and the steps of contestation. The risk calculator and system performance were perceived differently. P11 described the calculator as helpful for explaining scenarios to clients, while P9 perceived it as not truly improving transparency, as the feature weights were not sufficiently justified and remained untraceable. Many participants perceived the system's low true-positive prediction accuracy of 35% as a good justification for contesting and as a way to encourage clients. P7, however, cautioned that this could incite false hopes if the prediction were, in fact, correct. While we thus find unanimous approval of procedural information, the authority and control produced through numbers and technical explanations have to be carefully framed and communicated by the designer. If not, the impact of such information can be devastating, as P10 describes from a client's perspective: *"I now saw how the algorithm works. I understand that I am, in fact, one of those people. And maybe what the algorithm decided is right after all. You just think to yourself: Well, there's a lot of shit in my life."*

6.2.3 Explanations help counselors to assess contestation reasons and pathways. Social counselors consistently contested the agency's algorithmic decisions, seeing it as their duty to try everything for clients. Explanations did not change their overall contestation choices but shaped which reasons and channels they considered: "*Before [in B], it only says 'Here you can contest' [...] but now [in A-LLM] it is supported with data*" (P11). When asked if explanations were necessary, almost all agreed and described that explanations i) yielded new reasons to contest, ii) confirmed intuitions about suitable contestation reasons, iii) provided information to escalate via intervention networks, and iv) would help convey the process to clients. Not all of these purposes for explanations were valid for all participants: P3 and P8 stated that they would not use technical details (e.g., feature weights, performance) as they were outside of their expertise. P9 and P10 criticized the lack of justification for the values and principles guiding the ADM and its design [31], producing a "*pseudo-justification of the decision by bombarding you with technical details*" (P10). Thus, while explanations are helpful, the provided information must balance technical

detail with normative context about the system's sociotechnical setting. Further, explanation elements and contestation features should align with administrative laws and procedures familiar to counselors to give "*basic guidelines within the legal framework for determining what claims I have*" (P8). The implications of including such legal information in XAI design are discussed in Section 7.1.4.

6.2.4 A-flow vs. A-LLM: Conversational explanations can be a flexible support to counselors, if they are adopted. In comparing A-flow and A-LLM, we focus on participants' perceptions of the chat interface, the sole difference between the two interfaces, which was designed to enable flexible inquiry and conversational explanations. Reactions were polarized: of five participants, P2 and P11 did not use the chatbot, assuming it wouldn't provide relevant information; P3 and P7 valued asking flexible questions and found answers helpful; P9 found them superficial, verbose, and imprecise. Participants' questions to the chatbot included: "What happens with the appeal?" (P3), "Why are people affected by poverty at increased risk?" (P3), and "Is the decision based on stereotypes?" (P7). For the latter, P7 pasted the client's profile and the chatbot listed attributes that could indicate discrimination, leading P7 to revise the decision: "Based on that, I would not choose stereotyping as a reason anymore. [...] That it's inaccurate is actually a much better reason for objection." Thus, conversational explanations can provide detailed, targeted justifications that complement existing approaches, but negative expectations may deter users (such as P2 and P11). In summary, conversational explanations can help address knowledge gaps but pose challenges of imprecision and hallucinations. Given risks of misinterpretation and overconfidence, they should be deployed with caution in high-risk contexts.

7 Discussion

In this section, we reflect on our findings to formulate four lessons for designing explainability and contestability in high-risk ADM systems in public institutions. The section argues that XAI and contestation design must account for the social constellation of actor-networks [55], emphasize justification design to support collective contestation, consider adversarial over participatory design in multi-stakeholder contexts, and decide whether legal information is in scope when designing for contestation, followed by a discussion of the study's limitations.

7.1 Lessons learned from designing explainability and contestability in the public sector

7.1.1 XAI design needs to take into account the social constellations that underlie contestation in real-world contexts. Our findings demonstrate that the design of explanation and contestation of algorithmic decisions is not necessarily focused on single users, but can pertain to multiple actors in a sociotechnical constellation. Having started from the goal to design for decision subjects, the design process of this study pivoted to focus on social counselors as they act in a range of humans-in-the-loop [30] roles: as safeguards of affected individuals' dignity; allocators of accountability for administrative failures; advocates for other humans' values; and warm bodies to retain human points of contact. Fulfilling these converged roles makes social counselors *humans-in-the-contestability-loop*: crucial stakeholders for the policy-compliant implementation of explainability and contestability. XAI design needs to account for this convergence of roles by specifying how user roles relate to each other and which information they are expected to share and act upon. Human-centered XAI cannot limit its design scope to the individual level in high-risk settings, as the actions of actors, such as decision subjects, can depend on the support or authority of other actors, such as social counselors and the welfare agency. Tools such as question banks [71, 96], requirement analyses [89], and prototyping [99] can outline users' information needs, functionality, and usability requirements. But power structures and social dependencies are equally important in the context of contestation of high-risk algorithmic decisions [28] and should be considered in participatory explanation design. We see a sociotechnical gap [39] in the lack of exploration of these aspects. Future work should thus aim for a stronger integration of the dimensions of human actors in shared contexts by drawing

on actor-network theory or activity theory [24, 55] and carefully situate design solutions within these shared contexts.

7.1.2 More work on justifications is needed for XAI to support transforming individual into collective contestation. Our findings demonstrate that contestation for social counselors can be realized in two main ways: by contesting the decision of an individual client, and by challenging the system's legitimacy in total. Drawing from previous work [114], we describe these dimensions as individual and collective contestation, respectively. Collectivity here means the coordinated action between counselors and support organizations to build a stronger case of objection. Explanations can support the transformation of individual to collective contestation cases by allowing users to scrutinize the fairness or adequacy of the system's design decisions, human oversight, or deployment circumstances. Whereas traditional explanation methods [79] mostly allow interpretation of the ADM system itself, such as the selection of feature weights or model choice, justifications can enable the assessment of oversight measures and to identify policy embedded in the system design [49]. To illustrate, our paper shows that counselors who assess algorithmic decisions in individual cases simultaneously evaluate whether the system's guiding principles promote the well-being of decision subjects (e.g., ensuring fair hearings and providing human contact). If these principles are not met, counselors can escalate the case by involving other actors and activating their organizational networks. Explainable AI can support this process, but the design of justification mechanisms is underexplored in the current XAI literature [60]. Future work on XAI in high-risk settings should thus put emphasis on exploring how justifications support the transformation of individual into collective contestation issues by enabling the assessment of an ADM system's norms and values: Empirical studies could aim for the creation of question banks collecting normative information needs and the exploration of user-centered processes to design information that enables the identification of systemic shortcomings and value misalignments.

7.1.3 Participatory design is no panacea for multi-stakeholder explanation and contestation design. Early-stage deliberations are seen as pathways to negotiate multiple stakeholders' perspectives in public AI systems [47, 58]. However, when asked whether they would cooperate with the welfare agency to create a trustworthy ADM system, counselors in our study expressed caution. Concerned that their input might be read as institutional buy-in, several participants opposed deliberating ADM deployment with an institution they saw as conflicting with their own values and goals. Social counselors feared their participation would be *instrumental* and amount to "participation washing" [48], rather than *intrinsic* and a means to achieve justice [112]. This conflicts partly with the metaphor of the *agonistic arena* of contestable AI in the public sector: a space of adversarial interactions where conflict is productive [5]. In our study, conflict is an impediment, not a productive force, in early-stage multi-stakeholder deliberations, raising questions about whether participatory design can resolve value misalignments, given its reliance on cooperation [85]. XAI research must decide whether explanation and contestation design should align with individual or multiple stakeholders' needs and, in the latter case, deploy methods that handle dissent and divergent interests to render conflict productive. Design should explicitly acknowledge debate and keep political conflicts visible, an approach that design research calls *adversarial design* and that is well-established in machine learning [36] and red teaming [101]. While XAI research has explored adversarial contexts in technical design [10, 15], integrating these methods into user-centered design is a fruitful avenue for future work on explanation and contestation in multi-stakeholder contexts.

7.1.4 XAI design must consider whether legal information is in scope when designing for contestation. The interface design presented in this study and its observed effects focus on how an explanation and contestation system could be integrated into counselors' work and support their understanding and assessment of the ADM system. As Section 4.2 describes, the design did not consider legal or regulatory dimensions of contestability beyond GDPR rights [44]. Participants in the requirement analysis suggested that administrative laws could be considered, as they often serve as initial entry points for contestation (Figure 2). These legal needs were considered out of

scope because they require detailed knowledge of national legal contexts, would yield findings specific to the national context, and may not directly relate to algorithmic decision-making. The design thus compromised in the provision of legal information by including EU-level regulation (GDPR [44]) but not national law. We find a design tension here: Explanations should effectively support contestation, but judicial contestation requires legal information, which is outside of usual explanation design spaces. Future work should explore whether and how explanations can address this legal informational gap. Policy experiments [80] and abstracted legal contestation features [74] may offer useful avenues to investigate the resulting design trade-offs in XAI.

7.2 Limitations

In the following, we describe the limitations of our study. *First*, the design process initially focused on decision subjects but pivoted to social counselors after identifying them as primary users and their clients as served users. As recruiting clients proved difficult, counselors became the main informants on clients' needs, potentially introducing incongruencies. While closer inclusion of decision subjects was not feasible here, future work should prioritize their inclusion where recruitment is more realistic. *Second*, the study's participatory nature was in parts limited: counselors could not inform development of the CAF algorithm, which was presented as final; they contributed to the explanation and contestation design by producing artifacts in Stage I and acting as idea provides, evaluators, and informants in Stages II and III, thus the "direct design input" could be increased [110]. Further, the long-term impact depends on institutional change, which may proceed more slowly and uncertainly than theory presumes [14]. *Third*, findings on the effect of explanations might differ for post-hoc explanations, which suffer from fidelity gaps [15] and potential unfairness for underrepresented subgroups [10]. In contrast, our explanations used feature coefficients and performance data from a logistic regression, a white-box, interpretable model [92]. Post-hoc explanations may thus differ in their risk to misinform users [17] through overreliance [42] or a false sense of understanding [90], together with imprecision or hallucinations in LLM-generated explanations [102]. While we saw little evidence of such misinterpretation, future work should explore how to prevent these pitfalls for AI novices, especially in high-risk contexts.

8 Conclusion

This paper reported on the development and evaluation of an explanation and contestation interface to help social counselors understand and contest a high-risk ADM system. We conducted a three-stage participatory design study with social counselors, consisting of context and requirements analysis, low-fidelity design and testing, and high-fidelity interface development and evaluation. We find that social counselors act as mediators who filter information and manage contestation between their clients and the social agency. This entails specific requirements for explanations, as their design must support both counselors' understanding and client-facing explanations. Counselors perceived procedural information as essential, while information about model predictions and accuracy was seen as helpful but sensitive due to its emotional impact on clients. We formulate four lessons on the importance of social constellations for contestability, the role of justifications in transforming individual into collective contestation issues, the limits and potentials of participatory design for multi-stakeholder deliberation, and the inclusion of legal information in the scope of XAI. With this work, we contribute empirical insight into the relations between social, organizational, and technological dimensions of explainability and contestability that guides future work on the human-centered design of trustworthy AI systems.

Acknowledgments

This work has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20058] as well as [10.47379/ICT20065].

References

[1] 2024. France: Discriminatory algorithm used by the social security agency must be stopped — amnesty.org. https://www.amnesty.org/en/latest/news/2024/10/france-discriminatory-algorithm-used-by-the-social-security-agency-must-be-stopped/.

[2] Olusegun Agbabiaka, Adegboyega Ojo, and Niall Connolly. 2025. Requirements for trustworthy AI-enabled automated decision-making in the public sector: A systematic review. *Technological Forecasting and Social Change* 215 (2025), 124076. doi:10.1016/j.techfore.2025.124076

[3] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2023. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 8, 16 pages. doi:10.1145/3544548.3580984

[4] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by Design: Towards a Framework. *Minds and Machines* 33, 4 (01 Dec 2023), 613–639. doi:10.1007/s11023-022-09611-z

[5] Kars Alfrink, Ianus Keller, Mireia Yurrita Semperena, Denis Bulygin, Gerd Kortuem, and Neelke Doorn. 2024. Envisioning Contestability Loops: Evaluating the Agonistic Arena as a Generative Metaphor for Public AI. *She Ji: The Journal of Design, Economics, and Innovation* 10, 1 (2024), 53–93.

[6] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (Nov. 2023), 101805. doi:10.1016/j.inffus.2023.101805

[7] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (March 2018), 973–989. doi:10.1177/1461444816676645

[8] and European Economic and Social Committee. 2021. *Digital Services Act and Digital Markets Act – Stepping stones to a level playing field in Europe*. European Economic and Social Committee. doi:doi/10.2864/28842

[9] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 75, 13 pages. doi:10.1145/3411764.3445736

[10] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1194–1206. doi:10.1145/3531146.3533179

[11] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D. Mekler. 2023. How does HCI Understand Human Agency and Autonomy?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. doi:10.1145/3544548.3580651

[12] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 411, 21 pages. doi:10.1145/3544548.3581314

[13] Susanne Bødker, Christian Dindler, Ole Iversen, and Rachel Smith. 2021. *Participatory Design*. Springer Nature. doi:doi.org/10.2200/S01136ED1V01Y202110HCI052

[14] Susanne Bødker and Morten Kyng. 2018. Participatory Design that Matters—Facing the Big Issues. *ACM Trans. Comput.-Hum. Interact.* 25, 1, Article 4 (Feb. 2018), 31 pages. doi:10.1145/3152421

[15] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike Von Luxburg. 2022. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 891–905. doi:10.1145/3531146.3533153

[16] Maarten Bouwmeester. 2023. System failure in the digital welfare state: Exploring parliamentary and judicial control in the Dutch childcare benefits scandal. *Recht der Werkelijkheid* 44, 2 (Dec. 2023), 13–37. doi:10.5553/RdW/138064242023044002003

[17] Clara Bove, Thibault Laugel, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2024. Why do explanations fail? A typology and discussion on failures in XAI. http://arxiv.org/abs/2405.13474

[18] John Brooke. 2013. SUS: a retrospective. *J. Usability Studies* 8, 2 (2 2013), 29–40.

[19] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. doi:10.1145/3290605.3300271

[20] Caisse nationale des allocations familiales (Cnaf). 2016. L'enquête d'évaluation du paiement à bon droit et de la fraude de la branche Famille. https://www.cnis.fr/wp-content/uploads/2017/10/DC_2016_6e_reunion_GT_Travail_Dissimule_CNAF_enquete_PBDF.pdf. Présentation au groupe de travail du CNIS, 6e réunion du groupe de travail "Travail dissimulé".

[21] Bradley Camburn, Vimal Viswanathan, Julie Linsey, David Anderson, Daniel Jensen, Richard Crawford, Kevin Otto, and Kristin Wood. 2017. Design prototyping methods: state of the art in strategies, techniques, and guidelines. *Design Science* 3 (2017), e13. doi:10.1017/dsj.2017.10

[22] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, Greenville SC USA, 103–119. doi:10.1145/3640543.3645199

[23] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. doi:10.1089/big.2016.0047

[24] Torkil Clemmensen, Victor Kaptelinin, and Bonnie Nardi. 2016. Making HCI theory work: an analysis of the use of activity theory in HCI research. *Behaviour & Information Technology* 35, 8 (Aug. 2016), 608–627. doi:10.1080/0144929X.2016.1175507

[25] Pierre Collinet. 2013. Focus – Le data mining dans les Caf : une réalité, des perspectives:. *Informations sociales* n° 178, 4 (Aug. 2013), 129–132. doi:10.3917/inso.178.0129

[26] Robert Patrick Collins, Johan Redström, and Marco Rozendaal. 2024. The right to contestation: Towards repairing our interactions with algorithmic decision systems. (2024). doi:10.57698/V18I1.06 Publisher: International Journal of Design.

[27] Commission Nationale de l'Informatique et des Libertés. 2010. Délibération n°2010-086 du 25 mars 2010 autorisant la mise en œuvre par la Caisse nationale des allocations familiales (CNAF) d'un traitement de données à caractère personnel ayant pour finalité l'amélioration du ciblage des comptes allocataires à contrôler par les Caisses d'allocations familiales (CAF). https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000022205702. Accessed May 3, 2025.

[28] KATE CRAWFORD. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. http://www.jstor.org/stable/j.ctv1ghv45t

[29] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. doi:10.1145/3531146.3533108

[30] Rebecca Crootof, Margot E Kaminski, and W Nicholson Price Ii. 2023. Humans in the Loop. *VANDERBILT LAW REVIEW* 76 (2023). https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=2659&context=law-faculty-publications

[31] Karl de Fine Licht and Jenny de Fine Licht. 2020. Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations Are Key When Trying to Produce Perceived Legitimacy. *AI Soc.* 35, 4 (dec 2020), 917–926. doi:10.1007/s00146-020-00960-w

[32] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. doi:10.1145/3617694.3623261

[33] Caisse Nationale des Allocations Familiales (CNAF). 2023. Datamining Données Entrantes – Documentation Technique. https://www.documentcloud.org/documents/24177825-datamining-donnees-entrantes-documentation-technique/. Accessed May 2, 2025.

[34] Sam Desiere and Ludo Struyven. 2021. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. *Journal of Social Policy* 50, 2 (2021), 367–385. doi:10.1017/S0047279420000203

[35] Oege Dijk, oegesam, Ray Bell, Lily, Simon-Free, Brandon Serna, rajgupt, yanhong-zhao-ef, Achim Gädke, Hugo, and Tunay Okumus. 2022. oegedijk/explainerdashboard: v0.3.8.2: reverses set_shap_values bug introduced in 0.3.8.1.

[36] Carl DiSalvo. 2015. *Adversarial design*. Mit Press.

[37] Anke Dittmar and Maximilian Hensch. 2015. Two-Level Personas for Nested Design Spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 3265–3274. doi:10.1145/2702123.2702168

[38] Vincent Dubois, Morgane Paris, and Pierre Édouard Weill. 2018. Targeting by Numbers: The Uses of Statistics for Monitoring French Welfare Benefit Recipients. In *Creating Target Publics for Welfare Policies*, Laurence Barrault-Stella and Pierre Édouard Weill (Eds.). Logic, Argumentation & Reasoning, Vol. 17. Springer International Publishing, 93–109. doi:10.1007/978-3-319-89596-3_5 Accessed May 2, 2025.

[39] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O. Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 34 (apr 2023), 32 pages. doi:10.1145/3579467

[40] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1305–1317. doi:10.1145/3531146.3533186

[41] Upol Ehsan, Philipp Wintersberger, Elizabeth A Watkins, Carina Manger, Gonzalo Ramos, Justin D. Weisz, Hal Daumé Iii, Andreas Riener, and Mark O Riedl. 2023. Human-Centered Explainable AI (HCXAI): Coming of Age. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–7. doi:10.1145/3544549.3573832

[42] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo Japan, 211–223. doi:10.1145/3172944.3172961

[43] European Commission. 2024. Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations. .

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

[45] European Parliament. Directorate General for Parliamentary Research Services. 2019. *Understanding algorithmic decision-making: opportunities and challenges.* Publications Office, LU. doi:10.2861/536131

[46] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (June 2019), 261–262. doi:10.1038/s42256-019-0055-y

[47] Michael Gleicher. 2016. A Framework for Considering Comprehensibility in Modeling. *Big Data* 4, 2 (6 2016), 75–88. doi:10.1089/big.2016.0007

[48] Lara Groves, Aidan Peppin, Andrew Strait, and Jenny Brennan. 2023. Going public: the role of public participation approaches in commercial AI labs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1162–1173. doi:10.1145/3593013.3594071

[49] Clément Henin and Daniel Le Métayer. 2022. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* 37, 4 (Dec. 2022), 1397–1410. doi:10.1007/s00146-021-01251-8

[50] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. 2017. Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research* 27, 4 (March 2017), 591–608. doi:10.1177/1049732316665344

[51] Emmie Hine and Luciano Floridi. 2023. The Blueprint for an AI Bill of Rights: In Search of Enaction, at Risk of Inaction. *Minds and Machines* 33, 2 (01 Jun 2023), 285–292. doi:10.1007/s11023-023-09625-1

[52] Marie Hornberger, Arne Bewersdorff, Daniel S. Schiff, and Claudia Nerdel. 2025. A multinational assessment of AI literacy among university students in Germany, the UK, and the US. *Computers in Human Behavior: Artificial Humans* 4 (May 2025), 100132. doi:10.1016/j.chbah.2025.100132

[53] Steven J. Jackson, Tarleton Gillespie, and Sandy Payette. 2014. The Policy Knot: Re-Integrating Policy, Practice and Design in Cscw Studies of Social Computing. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 588–602. doi:10/gg5g9w

[54] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021), 1957–2048.

[55] Victor Kaptelinin. 2006. *Acting with Technology: Activity Theory and Interaction Design.* MIT Press ; Ebsco Publishing [distributor], Cambridge, Ipswich. OCLC: 904661880.

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

[57] Anne Kaun and Gabriela Taranu. 2020. Automating Society Report 2020 / Sweden. automatingsociety.algorithmwatch.org/report2020/sweden/.

[58] Anna Kawakami, Luke Guerdan, Yang Cheng, Anita Sun, Alison Hu, Kate Glazko, Nikos Arechiga, Matthew Lee, Scott Carter, Haiyi Zhu, and Kenneth Holstein. 2022. Towards a Learner-Centered Explainable AI: Lessons from the learning sciences. http://arxiv.org/abs/2212.05588 arXiv:2212.05588 [cs].

[59] Daniel Kluttz and Deirdre K. Mulligan. 2019. Automated decision support technologies and the legal profession. *SSRN Electronic Journal* (2019). doi:10.2139/ssrn.3443063

[60] Klára Kolářová and Timothée Schmude. 2026. Start Using Justifications When Explaining AI Systems to Decision Subjects. In *Digital Humanism*, Ludger Hagedorn, Ute Schmid, Susan Winter, and Stefan Woltran (Eds.). Springer Nature Switzerland, Cham, 190–202.

[61] Gilles Kounowski. 2002. L'informatique et le système d'information des Allocations familiales. *Recherches et Prévisions* 68-69 (2002), 49–72. doi:10.3406/caf.2002.1018 Accessed May 4, 2025.

[62] La Quadrature du Net. 2010. CAF Transmission Variables Odds Mod 2010. https://git.laquadrature.net/la-quadrature-du-net/algo-et-controle/caf/-/blob/main/caf_transmission_variables_odds_mod_2010.xls?ref_type=heads. Accessed May 3, 2025.

[63] La Quadrature du Net. 2010. CAF_code_2010_recu.sas. https://git.laquadrature.net/la-quadrature-du-net/algo-et-controle/caf/-/blob/main/CAF_code_2010_recu.sas?ref_type=heads. Accessed May 3, 2025.

[64] La Quadrature du Net. 2014. CAF_code_2014_recu.sas. https://git.laquadrature.net/la-quadrature-du-net/algo-et-controle/caf/-/blob/main/CAF_code_2014_recu.sas?ref_type=heads. Accessed May 3, 2025.

[65] La Rédaction. 2024. *Les algorithmes de la CAF pour contrôler les usagers.* https://lvsl.fr/les-algorithmes-de-la-caf-pour-controler-les-usagers/ Accessed May 3, 2025.

[66] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 2022. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–34.

[67] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (July 2021), 103473. doi:10.1016/j.artint.2021.103473

[68] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[69] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–35. doi:10.1145/3359283

[70] Leem. 2022. The French Social Security Finance Bill (PLFSS) & Medicines: An Explanation. https://www.leem.org/sites/default/files/2022-10/Kit-PLFSS-VersionAnglaise.pdf Accessed May 3, 2025.

[71] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[72] Lighthouse Reports. 2023. How We Investigated France's Mass Profiling Machine. https://www.lighthousereports.com/methodology/how-we-investigated-frances-mass-profiling-machine/. Accessed May 2, 2025.

[73] Paola Lopez. 2019. Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. In *Proceedings of the 18th Annual STS Conference*. Graz, 289–309. doi:10.3217/978-3-85125-668-0-16

[74] Henrietta Lyons, Tim Miller, and Eduardo Velloso. 2023. Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 764–774. doi:10.1145/3593013.3594041

[75] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Designing for Contestation: Insights from Administrative Law. http://arxiv.org/abs/2102.04559 arXiv:2102.04559 [cs].

[76] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. doi:10.1145/3491102.3517606

[77] Winston Maxwell and Bruno Dumas. 2023. Meaningful XAI based on user-centric design methodology: Combining legal and human-computer interaction (HCI) approaches to achieve meaningful algorithmic explainability. *SSRN Electronic Journal* (2023). doi:10.2139/ssrn.4520754

[78] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4, Article 24 (Sept. 2021), 45 pages. doi:10.1145/3387166

[79] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). https://christophm.github.io/interpretable-ml-book

[80] Nadia Nahar, Jenny Rowlett, Matthew Bray, Zahra Abba Omar, Xenophon Papademetris, Alka Menon, and Christian Kästner. 2024. Regulating Explainability in Machine Learning Applications – Observations from a Policy Design Experiment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2101–2112. doi:10.1145/3630106.3659028

[81] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 3 (2022), 1–30.

[82] Jedrzej Niklas, Karolina Sztandar-Sztanderska, and Katarzyna Szymielewicz. 2015. Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making. panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf.

[83] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez. 2023. The Role of Explainable AI in the Context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1139–1150. doi:10.1145/3593013.3594069

[84] Dorian Peters, Rafael A. Calvo, and Richard M. Ryan. 2018. Designing for Motivation, Engagement and Wellbeing in Digital Experience. *Frontiers in Psychology* 9 (May 2018), 797. doi:10.3389/fpsyg.2018.00797

[85] Xiang Qi and Junnan Yu. 2025. Participatory Design in Human-Computer Interaction: Cases, Characteristics, and Lessons. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 804, 26 pages. doi:10.1145/3706598.3713436

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

[87] Manon Romain, Adrien Senecat, Soizic Pénicaud, Gabriel Geiger, and Justin-Casimir Braun. 2023. How We Investigated France's Mass Profiling Machine — lighthousereports.com. https://www.lighthousereports.com/methodology/how-we-investigated-frances-mass-profiling-machine/.

[88] Manon Romain, Adrien Sénécat, Elsa Delmas, Thomas Steffen, Léa Girardot, and Lighthouse Reports. 2024. Comment l'algorithme de la CAF prédit si vous êtes « à risque» de frauder — lemonde.fr. https://www.lemonde.fr/les-decodeurs/visuel/2023/12/04/comment-l-algorithme-de-la-caf-predit-si-vous-etes-a-risque-de-frauder_6203836_4355770.html.

[89] D.T. Ross and K.E. Schoman. 1977. Structured Analysis for Requirements Definition. *IEEE Transactions on Software Engineering* SE-3, 1 (Jan. 1977), 6–15. doi:10.1109/TSE.1977.229899

[90] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26, 5 (2002), 521–562. doi:10.1207/s15516709cog2605_1

[91] Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: How to plan, design, and conduct effective tests.* John Wiley & Sons.

[92] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (01 May 2019), 206–215. doi:10.1038/s42256-019-0048-x

[93] Gordon Rugg and Peter McGeorge. 2005. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems* 22, 3 (2005), 94–107.

[94] République Française. 1996. Loi organique n° 96-646 du 22 juillet 1996 relative aux lois de financement de la sécurité sociale. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000548068. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000548068 Accessed May 4, 2025.

[95] République Française and Caisse Nationale des Allocations Familiales (CNAF). 2013. Convention d'objectifs et de gestion 2013–2017 entre l'État et la CNAF. https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/COG/2013/CONVENTION/CONVENTION_D-OBJECTIFS_ET_DE_GESTION-2013-2017_ENTRE_L-ETAT_ET_LA_CNAF.pdf. https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/COG/2013/CONVENTION/CONVENTION_D-OBJECTIFS_ET_DE_GESTION-2013-2017_ENTRE_L-ETAT_ET_LA_CNAF.pdf Accessed May 4, 2025.

[96] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2025. Information that matters: Exploring information needs of people affected by algorithmic decisions. *International Journal of Human-Computer Studies* 193 (2025), 103380. doi:10.1016/j.ijhcs.2024.103380

[97] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "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*. ACM, Seoul Republic of Korea, 1616–1628. doi:10.1145/3531146.3533218

[98] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, Seoul Republic of Korea, 2138–2148. doi:10.1145/3531146.3534631

[99] H. Sharp, J. Preece, and Y. Rogers. 2019. *Interaction Design: Beyond Human-Computer Interaction.* Wiley. https://books.google.de/books?id=UDeQDwAAQBAJ

[100] Hong Shen, Haojian Jin, Angel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. doi:10.1145/3415224

[101] Ranjit Singh, Borhane Blili-Hamelin, Carol Anderson, Emnet Tafesse, Briana Vecchione, Beth Duckles, and Jacob Metcalf. 2025. *Red-Teaming in the Public Interest.* Technical Report. Data & Society Research Institute and AI Risk and Vulnerability Alliance. https://datasociety.net/wp-content/uploads/2025/02/Red-Teaming-in_the_Public_Interest_FINAL1.pdf Accessed: 2026-03-20.

[102] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* 5, 8 (July 2023), 873–883. doi:10.1038/s42256-023-00692-8

[103] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 2239–2250. doi:10.1145/3531146.3534639

[104] Balsamiq Studios. 2026. Balsamiq. https://balsamiq.com Accessed: 2026-01-03.

[105] Sécurité sociale. 2025. *Conventions d'objectifs et de gestion (COG).* https://www.securite-sociale.fr/la-secu-en-detail/gestion-financement-et-performance/cog Accessed May 3, 2025.

[106] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy Artificial Intelligence. *Electronic Markets* 31, 2 (June 2021), 447–464. doi:10.1007/s12525-020-00441-4

[107] Carl W Turner, James R Lewis, and Jakob Nielsen. 2006. Determining usability test sample size. *International encyclopedia of ergonomics and human factors* 3, 2 (2006), 3084–3088.

[108] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. doi:10.1145/3415238

[109] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–28. doi:10.1145/3476059

[110] Peter Wacnik, Shanna R. Daly, and Aditi Verma. 2025. Participatory design: a systematic review and insights for future practice. *Design Science* 11 (2025), e21. doi:10.1017/dsj.2025.10009

[111] Katharina Weitz, Ruben Schlagowski, Elisabeth André, Maris Männiste, and Ceenu George. 2024. Explaining It Your Way - Findings from a Co-Creative Design Workshop on Designing XAI Applications with AI End-Users from the Public Sector. In *Proceedings of the*

CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 745, 14 pages. doi:10.1145/3613904.3642563

[112] Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday* 29, 4 (Apr 2024). doi:10.5210/fm.v29i4.13642

[113] Mireia Yurrita, Agathe Balayn, and Ujwal Gadiraju. 2023. Generating Process-Centric Explanations to Enable Contestability in Algorithmic Decision-Making: Challenges and Opportunities. In *2023 Human-Centered XAI Workshop at CHI Conference on Human Factors in Computing Systems (CHI '23)*.

[114] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Kars Alfrink, Ujwal Gadiraju, and Alessandro Bozzon. 2025. Identifying Algorithmic Decision Subjects' Needs for Meaningful Contestability. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–29.

[115] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Kars Alfrink, Ujwal Gadiraju, and Alessandro Bozzon. 2025. Identifying Algorithmic Decision Subjects' Needs for Meaningful Contestability. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (Oct. 2025), 1–29. doi:10.1145/3757415

[116] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia C. Pont, and Alessandro Bozzon. 2025. Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1028, 21 pages. doi:10.1145/3706598.3713145

[117] Theresa Züger and Hadi Asghari. 2023. AI for the public. How public interest theory shifts the discourse on AI. *AI & SOCIETY* 38, 2 (April 2023), 815–828. doi:10.1007/s00146-022-01480-5

9 Endmatter

9.1 Generative AI Disclosure Statement

During the preparation of this manuscript, we utilized the following AI-assisted tool to assist with formatting, grammar, and fluency of writing: ChatGPT (version GPT-5, released by OpenAI in 2025) and Grammarly. These tools were used solely to improve clarity and readability without altering the paper's intellectual content, methodology, or findings. We further used GPT5 through GitHub Copilot for assistance in the development of the high-fidelity prototype.

9.2 Ethical Considerations Statement

For this study, we conducted a series of three interviews with participants recruited from social support organizations. Of these participants, one was a client, and all other were social counselors. The client participated in the interviews together with a social counselor as per stated request by the client. Every participant was provided with information about the study's aims, process, responsible persons, and data collection and storage prior to participation, and signed a confirmation form attesting that they understood and agreed with the participant information. Participants were eligible to and received monetary compensation for their participation, if they participated outside of their working hours, and else received alternative compensation in the form of snacks during the interviews.

In a briefing before the beginning of the first interviews, social counselors recommended us to explicitly mention that the ADM system presented in the study was used in France, and to emphasize that the system's deployment was a fully hypothetical scenario in Austria, to avoid discomfort among participants. Hence, our scenario description is transparent about the origin and nature of the system.

The first University of Vienna's Research Ethics Committee approved this study under reference 01433.

9.3 Use and attribution of icons

The icons used in Figure 6 as "algorithm" is provided by Freepik through Flaticon.com.

Appendix

The following sections provide supplementary information on aspects of the study design and analysis. Section A provides more details about the CAF algorithm, B describes a typical workflow of client-counselor interaction, C gives additional information on the study's questionnaire, D lists participants' questions created in the card sort, E provides an overview of explanation and information elements used in the low-fidelity prototypes, F provides more images of the low-fidelity designs, G lists the interview questions asked to participants to evaluate the high-fidelity interface, H provides the system prompt used for the chatbot in interface A-LLM, I describes the files used as information base for the retrieval-augmented generation, and J provides the complete codebook of the inductive thematic analysis.

A More information on the CAF algorithm

History. Algorithmic risk scoring in French welfare emerged from broader social policy transformations in the 1990s under the "Plan Juppé", which tightened state control over welfare financing [70, 94, 105]. The plan's goals (expense consolidation and efficiency) created new commitments to combat social fraud, prompting development of the DMDE[2] risk-scoring system as a technical response [38]. The 2001 to 2004 performance agreement prioritized risk-based monitoring and audits based on beneficiary data [38]. Locally, accounting officers, who were personally liable for missed fraud, required tools to support the controls. In 2004, the Caisse d'allocations familiales (CAF, the French social insurance) launched DMDE based on logistic regression models and a dataset of roughly 3,000 known fraud cases [38]. The system was then scaled nationally and sourced local data to build a broader learning base [20, 25]. By 2010, seventeen branches of the CAF were piloting versions of the system. After authorization by the French data protection authority (CNIL[3]), nationwide deployment began in 2011 [25, 27, 38].

Architecture. Implemented in SAS [63], the 2010–2014 system comprised five independent logistic regression models: SCOREPRO (professional status), SCORESIT (familial status), SCORERESS (financial resources), SCORELOG (housing), and SCOREGLOB (global features associated with "indus"[4]), with the final score taken as the maximum score between the five models [38, 62, 63]. Algorithmic scoring soon became central to fraud control, reinforced by oversight from the Court of Auditors and the National Anti-Fraud Unit and institutionalized by the 2013–2017 performance agreement that tied metrics to detected overpayments and recovery rates [38, 95]. Algorithm-triggered audits rose from 23% (2011) to 63% (2016), staff-initiated checks fell from 51% to 20%, and while home checks declined from 280,000 (2009) to 166,000 (2015), their hit rate increased from 17% to 44% [38, 65]. CNAF's statistics directorate introduced a single-model version in 2014 retaining most of the 33 model variables [64]. The 2018 version of the system was piloted in 2019 and deployed nationwide in 2020 [33, 72].

Development. The logistic regression model operates within a broader IT ecosystem centered on a central decision-support data warehouse that supplies statistics to local branches [61]. Key system components include a relationship management platform with recipient and household data; a document and communications manager that tracks letters, forms, and calls; an appeals and settlements application; and a social assistance system that contains financial aid information [61]. Additional inputs include municipal-level socio-economic data and an anonymized repository of confirmed fraud cases, which both feed the construction and refinement of risk profiles [27, 33]. Together, these components form the informational basis for the modeling and development of the algorithmic welfare fraud scoring system [33, 61].

[2]Dispositif de Maîtrise des Dépenses et des Erreurs.
[3]Commission nationale de l'informatique et des libertés.
[4]In the CNAF taxonomy, "indus" or undues, signify overpayments from the side of CAF to benefit recipient as a result of an irregularity.

B Typical workflow in counselor-client interaction

In the following, we describe a typical interaction between clients and counselors to situate the explanation and contestation within the counselors' workflow.

Clients often contact Volkshilfe Vienna through one of several local counseling offices across the city. There, they register, state their issues, and receive help from social counselors. Clients in this study typically seek help with housing, welfare benefit applications, or financial issues. The client is invited to a bilateral meeting where the problem, personal circumstances, and current exchanges with the welfare agency are discussed. If administrative notices have been received, clients bring them for assessment. Depending on the case, the counselor may help directly with online or postal benefit applications, drafting response letters, or finding housing through internal channels. In agency conflicts, counselors prioritize direct written or telephone contact with the responsible case worker. Interaction may be limited to a few meetings if resolved quickly, or become long-term if the organization provides permanent housing.

The interface in this study comes into play when clients receive administrative notices containing an algorithmic decision and contact Volkshilfe's counselors. In bilateral meetings, counselors take inventory, assess documents, and use the interface to understand decision parameters and background, and to lodge a complaint. The interface can serve as a counselor's work tool or as a demonstration during meetings with clients.

C Questionnaires

Prior to study participation, participants filled out a questionnaire asking about demographical data, their relation to the social agency (as counselor, as client), their AI literacy, and for participants interacting with A-LLM their previous experience with chatbots, including which models they used how often and for which purposes. AI literacy was elicited through the five-point self-efficacy scale developed by Hornberger et al. [52].

In Stage III, participants further answered a version of the TENS-Task scale, designed to capture "the experience of engaging in a technology-specific task" [84] with a focus on participants' sense of autonomy, competence, and relatedness. After repeated pilot testing, we decided to remove the question area targeting relatedness (feeling connected to others), as these questions confused pilot testers and were not perceived to relate to the task or interface at hand.

D Stage I: Participants' full list of questions

Table 2. All questions created by participants for the card sort in Stage I. The questions capture participants' information needs about the ADM system and were ranked by participants according to their perceived priority.

P1	P2	P3	P4	P5	P6	P7
Is the decision reviewed by a human?	Which law is this decision based on?	On what basis is this fraud claim made in the decision?	What is the process, how does the system evaluate this?	What complaint/appeal options exist and what are the associated risks?	How is the risk determined?	Do biases appear in the system? For example, does someone with a refugee background have a higher risk?
What rights do I have?	How can you prove to the agency that you are not receiving too much money?	Does the fact that someone is single and a single parent increase the risk?	Can it be that the results are not accurate, that they deviate?	Is the decision still reviewed by a human?	Why are benefits stopped immediately and not checked internally first?	Can welfare benefits even be discontinued because of such a system?
What happens in this process?	Does someone support me in filing legal appeals?	Is the risk higher when someone has been receiving benefits for 5 years?	Is it possible to see practical examples with which the system was tested?	What information is entered exactly? Which information goes into the calculation?	Is the decision reviewed by a human?	How is something like 0.27 calculated? What does that mean?
	What legal appeals can I file?	What else does the system know? What can it access?	Does wealth count as income as well?	In which cases is there a risk and why?	Why is the risk determined this way for this client? What reasons lead to it?	Why does income reduce the risk? Why are these attributes weighted like this?
	Which parameters are at play in this decision?	Who is responsible for the system?	Are alimony/child support included in income?	How can the risk be reduced?	Why do the mentioned parameters increase risk?	What is contained in this profile? What data does the social agency have access to?
	How much time do I have to file an objection?	Does the system capture my lifestyle habits?	Who developed this system?	Which benefits can I claim?	What options do I have for lodging an objection?	How are household/benefit units assessed?
	How come that the number of children influences the risk score?	Who operates the system?	What do the percentage values for the attributes mean?	How is the risk calculated from the weights of the attributes?	What rights do I have?	Why use such a system at all?
	So you have to prove your innocence instead of the agency proving guilt?	Why does age play a role?		Where does the 75% limit come from?		What are these old fraud cases? How is fraud defined?
	What does the 5% increase for age mean?	Where does the information come from, where does the system get my data?		What are the backgrounds for the algorithmic risk assessment?		
		How does this system recognize me?		Why do age and marital status increase risk?		
		Can the system also perform calculations?				

E Stage II: Explanation and information hierarchy elements

Figure 7 shows a selection of design elements that were identified from previous work or created and tested in the low-fidelity prototypes.

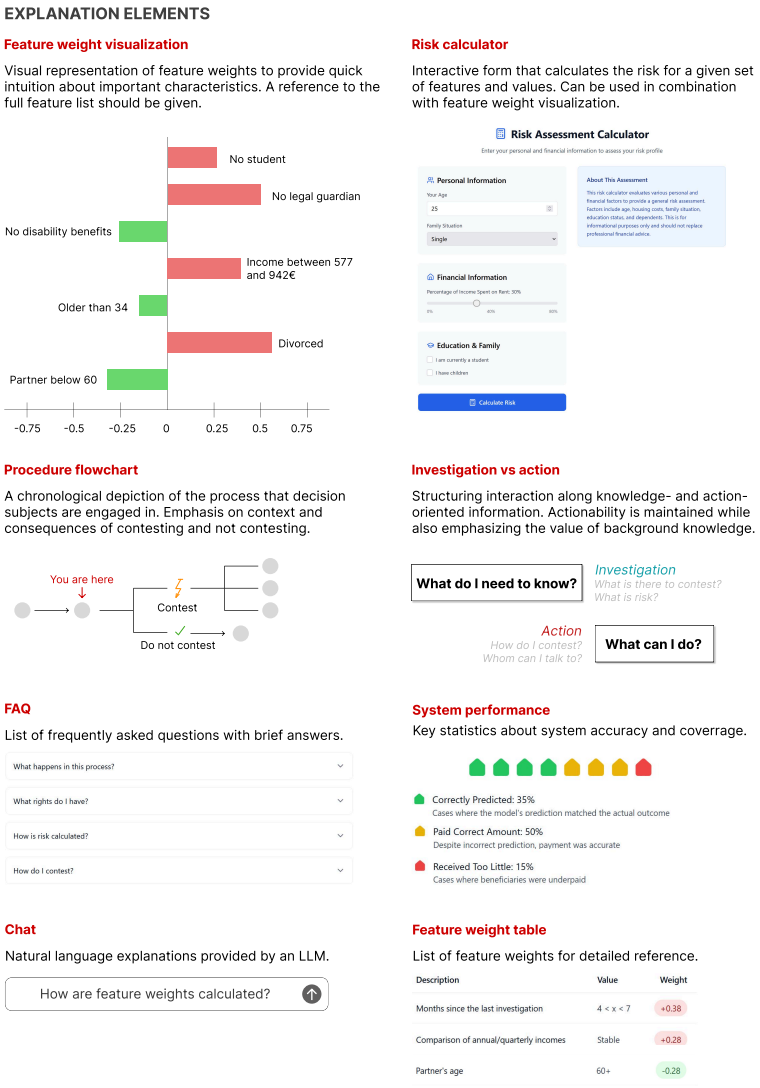

Fig. 7. **Selection of explanation elements incorporated in the designs.** Feature weights were inspired by explanations of local prediction instances [69, 86, 97], the risk calculator by interactive explanation designs [12, 29], the flowchart by process-centric explanations [113], the FAQ by question-driven explanation approaches [9, 71, 96], the system performance metrics confusion matrix design for lay audiences [100], the chat interface by conversational explanations [102], and the table of feature weights by explanation dashboard designs for non-experts [81]. The separation into investigation and action pages originated from own design ideas.

F Stage II: Images of the low-fidelity prototypes

The low-fidelity interfaces underwent five iterations of parallel and iterative prototyping. Figure 8 provides images from different iterations.

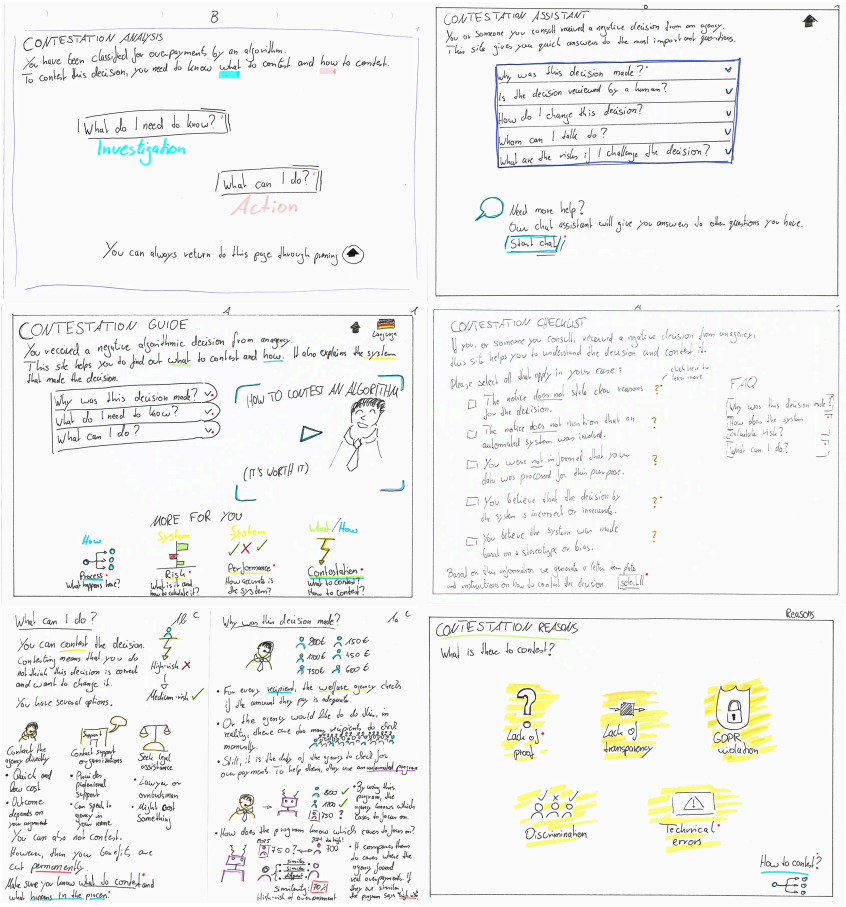

Fig. 8. **Selection of panels from the low-fidelity prototyping process.**

G Stage III: Interview guide

After the interaction with the baseline interface and A-flow/A-LLM, participants were asked a selection of concluding questions to reflect on the interaction experience, their understanding, contestation choice, and sense of agency.

(1) Selection of contestation reasons and pathways
- Why did you choose these reasons for the appeal?
- Why did you choose these avenues for the appeal?
- Did your selection of reasons and avenues change between the first (B) and the second (A-flow/A-LLM) website?
- If yes, why?

(2) Influence of explanations on contestation choice
   - Which of the elements influenced your decision the most?
   - Which of the elements would have helped you the most in explaining the system to a client?
   - Did the information in the second website (A-flow/A-LLM) make you feel more confident or less confident in contesting?
   - Did you feel that you understood the system better? Why?
   - Were there things that remained unclear to you? Why?
(3) Feeling of autonomy and competence
   - Do you feel you have control over the situation with the social welfare office? Why?
   - Did your sense of control change between the first (B) and the second (A-flow/A-LLM) website? Why?
   - Would you need the second (A-flow/A-LLM) website at all, or would the first one be sufficient?
(4) Experience with chatbot (only A-LLM)
   - What was your impression of the chatbot?
   - What did it help you with, and what did it not help you with?
   - What can the chatbot tell you that the rest of the website cannot?
(5) From your perspective, would this website be helpful for your work? Would anything need to be changed?

H Stage III: Chatbot's system prompt

The chatbot was meant to serve as a web-based social support agent that helps welfare recipients understand and contest algorithmic decisions. Its goal was to provide accessible and flexible explanations, rights information, and guided contestation options. The system prompt was tested and refined in pilot studies with peers as well as with participants' collected questions from Stage I. Below, we provide the chatbot's system prompt.

> **Your application context**
> You are a social support agent who helps users understand and contest administrative decisions made by the social agency, which is responsible for paying welfare benefits. You are not yourself working for the social agency, but for an external social support organization.
> For each benefit recipient, the social agency uses an algorithm that is based on logistic regression and 33 parameters to calculate a risk score that this person is receiving too much money. The process is called risk scoring of welfare fraud.
> You are implemented in the context of a website that helps users to explain and contest these decisions. Questions of users might refer to the website content and functions. The website is structured into two main sections: "Information" and "Contestation".
> Under "Information", the website explains the decision procedure in the social agency, provides a calculator to compute one's own risk score, shows the system performance, and a table of all features that you know from the parameters pdf file.
> Under "Contestation", users are informed about their rights and the consequences of contesting. They can subsequently select from several reasons to contest the decision and choose a way (direct contact to social agency, contact to support organizations, contact to legal offices). The website then simulates that the contestation is sent in the way chosen. Users can contest for the following reasons: Unsufficient reasons given in decision; decision does not mention that an automated system was used; decision does not mention if a human was involved; system uses wrong or outdated data; system might have made an inaccurate decision; system discriminates against person in question; other reasons (free text). Not all of these reasons are necessarily correct.
> Users have the following rights: right to know that an algorithm is involved; right to know which data of them is processed; right to contest the decision; right to get a human involved.

Contesting can have these consequences: Review of documents by social agency; conversation with caseworker; new decision; legal dispute if not resolved otherwise.

Your behaviour
For each question, you scan the knowledge base for an answer. You aim to give brief answers that are easy to understand. You do not use technical jargon unless the user explicitly asks you about it. That means you should avoid words like "logistic regression" and "algorithm" and instead use words that are easier to understand. Even though you have the precise logit weights from the parameter PDF file, you do not provide the precise weights unless explicitly asked, instead you translate the weights to "high impact", "low impact", and so on.
You have an output token limit of 300 words, but you try to keep your answers below 150. You suggest follow-up questions after providing information. If you suggest contesting a decision, you suggest that the user can use the website for that, by selecting the appropriate reasons in the site "Contestation".
If you do not know the answer to a question, you must say so. If the user asks for the same thing twice, provide an easier answer than before.

I Stage III: Chatbot's information base

The chatbot was complemented with retrieval-augmented generation to allow it to more provide more precise information sourced directly from information about the ADM system. The chatbot's knowledge base was compiled from information sources on the welfare fraud detection system in France [87] and contained information on the deployment context, logistic regression parameters, performance data, and decision process information. The same knowledge base was also provided for interface A-flow for direct reference by participants.

J Codebook

The following tables list the themes, subthemes, and codes developed from the inductive thematic analysis in Stage III. Table 3 and **??** contain codes on how the interface supported social counselors. Table 4 contains codes on what participants suggested to improve in the interface. Table 5 contains codes that capture the effect of explanations on counselors and clients. And Table 6 contains codes capturing considerations concerning the implementation of an explanation and contestation interface in a real-world context. Each table is split into subthemes and codes that are grounded in example quotes from the interviews.

Below are your tables converted to longtable environments with continued headers and "continued on next page" footers. You can drop the surrounding table environment; longtable handles paging and captions itself.

Table 3. Codebook for the theme **How the interface helps counselors**. The subthemes capture categories of contestation, understanding, client-side explanation, scrutiny of the ADM system, and benefits specific to the chat interface.

Subtheme	Code	Description	Examples
To contest	Confidence through information	More knowledge has led to more confidence	"The most helpful are the flowchart and the explanations with the justifications. That gives certainty that this could apply; if I had to write something myself it would be harder." (P3), "Before I couldn't name it; now I know the percentages [of accuracy] and I can bring them in." (P8), "Because it lets me better assess whether my appeal will actually have the effect I want to achieve." (P11)

Continued on next page

Codebook for the theme How the interface helps counselors (continued)

Subtheme	Code	Description	Examples
	Good reason to contest	Feels that this point strongly supports the contestation	"That is a strong argument that there are no reasons in the notice." (P4), "Yes, I would already file an appeal here from a legal perspective; every notice must be justified, here it's only a pseudo-justification." (P10), "That is very bad, that would be fatal, it wouldn't hold up in any court in the world." (P12)
	Practical relevance	Circumstances in which the website becomes relevant	"The website itself could become more relevant if these kinds of algorithmic systems become more relevant." (P8)
	Time is crucial	Things need to move fast for contestation	"Well, of course it's more practical and simpler. I don't have to struggle with formulating appeals and squeezing out texts that sound legal." (P2), "It's practical that it's click click click and it's gone." (P3)
To understand	Appreciates simplicity	Remarks positively that the website is easy to handle	"I'd rather choose the basic version. The simpler, clearer, punchier and faster, the better." (P2), "Clear, straightforward, easy to use, not much frills, [...] and usable for clients too." (P3), "I like the website; it's also low-threshold." (P10)
	Interactive reasoning	Developing understanding through interaction	"Yes, that was very helpful for me with the risk calculator, and with the system's accuracy, the statistical values." (P4)
	More is more	Prefers having the second site (A-flow/A-LLM) rather than only the first (B)	"The second website is indeed good and helpful." (P4), "The second website helps for explaining it to clients." (P2), "The first would certainly suffice, but the graphical representation is important." (P3)
	Nice to have	Interesting but not essential	"The accuracy is interesting, also to reassure the client, but it doesn't affect my work." (P7), "Technical details are nice to have, but not the level on which I would file an appeal." (P8), "The calculator is a toy; I find myself asking, 'So what?'" (P9)
	Priority information	Something deemed very important	"Accuracy is very relevant; it could be displayed more prominently." (P5), "[With the calculator] Now it gets exciting, that's brilliant, and that's what we should insist on." (P11)
	Secondary information	Not strictly necessary to look at	"I also find it secondary how the system works in detail." (P7), "I didn't look at the factors, they were secondary, too." (P8)
	Working understanding	Understanding sufficient for the task	"Yes, I understand the system better now, including why Elif was scored low." (P3), "Yes, definitely, because of the list in the calculator. Even if I don't agree with the assessment, it enables me to file a better appeal and gives me more arguments." (P7)
To explain to clients	Key information	What to show to a client sitting on the side	"The flowchart, I'd use that to explain the steps. Also the calculator and the accuracy, but not the table." (P2), "I'd enter everything into the calculator, also together with the client, to run through cases." (P11)

Continued on next page

Codebook for the theme How the interface helps counselors (continued)

Subtheme	Code	Description	Examples
To scrutinize	Assumed discrimination	Belief that decision is based on discriminating stereotypes	"Stereotypes here already mean discrimination, because they preempt the individual assessment and don't give the person a chance to present their case properly." (P10), "It depends on the consequences of the stereotypes, but here yes, because of the normative consequences." (P11)
	Authority of numbers	Numbers produce power and control	"Numbers also confer authority in situations with clients." (P2), "The risk is not factual but statistically predicted; that should be phrased more cautiously." (P5), "Before I couldn't name it; now I know the percentages [of accuracy] and I can bring them in." (P8)
	Bad design	Algorithm badly designed, but not on purpose	"A system, an algorithm simply cannot capture every lived reality." (P7)
	Demands justification	Designers should defend this choice of system design	"I would also like information about the intention behind the system, and right at the beginning, to frame the rest of the information." (P9)
	Human involvement	Whether and how a person is involved in the decision	"The caseworker … yes, what are they actually there for?" (P3)
	Sanctioning design	Algorithm designed to be bad	"Was it deliberately designed so that the accuracy is so low?" (P5), "Oh man, that's depressing; I imagine being affected and trying to understand my risk, and the risk calculator says, 'yes, that's right, you are crap.'" (P10), "For me, it was shockingly precisely spelled out how the evaluation is done. […] That's impressively negative." (P11)
Through chat	Chat helps	Instances where the chatbot was useful	"It can answer more detailed questions, justify much better, and it can also answer 'why' questions." (P3)
	Chat prejudice	Assuming chatbots are not much help	"I never use it. […] That was probably 27 years ago, it didn't really help and I haven't clicked on it since." (P2)
	Chat trade-offs	Pros and cons of a chatbot	"It is of course error-prone, but it also creates a subjective feeling of safety, that's good." (P7), "Assistants are of course limited; they must not disclose the data and you mustn't enter sensitive data." (P9)
	Chat usage	How someone usually uses chatbots	"I don't use ChatGPT that often, but colleagues do, for administrative law questions." (P7)
	Interaction change	Behavior changes with conversational explanations	"It delivers quick answers; it's less 'studying,' more fast info." (P7)
	Rarely uses chatbots	Aversion or disinterest	"I rarely use chatbots; I often feel it's a waste of time." (P11)

Table 4. Codebook capturing participants' statements on **What the interface could improve**, organized by the subthemes contestation, understanding, and scrutiny.

Subtheme	Code	Description	Examples
Contestation	Consequences of contestation	What can happen after contesting	"The flowchart could also point out that money can be lost." (P5), "The website is efficient in the sense that it's fast, but it's missing: Which entitlements you have or not. If the appeal fails, you may have wasted your chance." (P8)
	Contestation procedure	How and whether the contestation is filed	"So far I've only been referred to a counseling center. [...] The appeal process is ultimately not yet completed." (P11)
	Difficult to contest	Shaky grounds for contestation	"It will be difficult because it was checked by a system and additionally by a caseworker; I don't know whether you can achieve anything with an appeal. You can try, no question, but I almost think it will be fruitless in this case." (P4)
	Feature request	Would improve the site	"Missing: the ability to print, a response from the recipient, an explanation of what happens after the appeal." (P3), "A calculator for minimum income would also be helpful." (P5), "A free-text field where the person can tell their story would be important." (P10)
	Legal particularities	Legal nature of contestation	"Further rights beyond GDPR would be important, to know what is possible." (P8), "You do have an administrative right to appeal; exactly where does the decision go in court?" (P5), "Printing the template would be important, or some confirmation that the appeal has been received, because of the deadline." (P7), "I would link the Minimum Income Act here to see the exact wording." (P7)
	Not enough information	Lacking info to judge contestation	"I would perhaps add how it is in this case. [...] Has child support already been demanded from the child's father? [...] How good are her language skills?" (P4), "[After B] Feeling confident about filing an appeal, rather not. Because these are all very automated questions with relatively little explanation at the beginning and relatively little now for me as a layperson. Well, okay, I'm a lawyer, but even for me as a lawyer these few sentences didn't give me much to go on." (P10)
	Regulatory grounding	Lacking connection to concrete laws	"On the website there is a tab that says 'your rights in relation to the GDPR'. But it is also about the administrative rights, [...] if you are even able to establish personal contact, or if everything has to be done in writing. [...] For me, it would be important to have some basic guidelines within the legal framework for determining what claims I have."

Continued on next page

What the interface could improve (continued)

Subtheme	Code	Description	Examples
Understanding	Contest on suspicion	Assuming a reason is valid for contesting	"I checked that on suspicion, just to be able to state more reasons. I checked 'the system's decision is inaccurate' because, firstly, one should challenge a lot and, secondly, that just can't be right." (P3)
	Information structure proposal	How to organize information	"I think the [table] doesn't need its own heading. It could simply be the risk calculator and then you click here and something separate opens." (P5)
	Missed aspect	Something not understood properly	"[Goes back to the calculator and enters the case, but overlooks two inputs (marital status and months since receiving benefits).]" (P8)
	No gain through information	Did not benefit from information	"The risk calculator is a fun toy, but so what? To me that still has nothing to do with transparency, because I don't know why the risk rises here or falls there, it's not traceable for me." (P9)
	Too much information	More information than necessary	"[The table] is more for nerds. Effect on risk… that's where I drop out. I wouldn't look at it anymore." (P2)
Scrutiny	Pseudo-justification	Information that only pretends legitimacy	"There's a lot of information that you can't really place unless you read into it. And it creates a pseudo-justifiability of the decision by throwing all the technical stuff at you and saying 'here, that's why.'" (P10)

Table 5. Codebook for the theme **Explanations' effects on perceptions and emotions**, ordered by effects on clients and on counselors.

Subtheme	Code	Description	Examples
On Clients	Emotional context	Belief that clients' emotions will come into play	"Clients often come to us under a lot of pressure." (P7), "Such clear-cut pluses and minuses [the feature weights], I think that suggests right and wrong behavior. [...] And as someone who is affected, I don't know how much I would criticize this." (P5)
	Legal anxiety	Going to court means stress for everyone	"You could lose trust here because of the mention of legal proceedings." (P4), "Young people don't feel comfortable with the legal route, especially because they have so few resources and are intimidated." (P10)
On Counselors	Improved control	Site gives a firmer grasp of the situation	"Yes, in the sense of being able to access information that I wasn't aware of before." (P11)
	Lack of responsibility	Not acting responsibly	"That's exactly what I criticize in the application process when people have a cover letter written by AI and then it's evaluated by AI in the online portal. […] We shirk our responsibility as humans. That can help some and hinder others, but I think the ultimate responsibility lies with the human." (P9)

Continued on next page

Explanations' effects on perceptions and emotions (continued)

Subtheme	Code	Description	Examples
	No human trace	Dehumanized; machines interact with machines	" 'This website gives me the freedom to file however I want.' No, rather not. Because I still can't indicate here that I'm completely different and don't need all these stupid checkboxes, I just want to be looked at as a human being." (P10)
	Sense of agency	Feeling in or out of control	"Exactly, because the authority… it's an administrative act that infringes on my rights because my benefits are being cut. That means the authority has to explain why the benefits are being cut. […] I would say that's not my problem." (P10)

Table 6. Codebook for theme **Real-world implementation of interface**, ordered by topics involving administration and social support organizations.

Subtheme	Code	Description	Examples
In administration	Administration quirks	Oddities of public administration	"I'd be more likely to say we've already come across cases where certain systems interact. It's all in the data protection paperwork, you sign it as soon as you apply for benefits with the social services office." (P11)
	Institutional opacity	Official procedures block important information	"A lot is said verbally that isn't in the letters, so there's no transparency for those affected." (P12)
	Procedural roadblocks	Structures that make contestation difficult	"I think complaints are made more often than I hear about. From my own social work operational experience, such complaints mean a lot of work." (P11)
In social support organizations	AI competence network	Places that help with AI questions	"I'm not an opponent. We even have an AI office hour once a month, appointments run by the company's IT department. We've also had talks on how we should deal with it." (P3)
	Always contest	Would challenge decision no matter what	"These numbers don't help me even if it's right 99%. But in 1% it's wrong. […] I file an appeal anyway." (P2)
	Contestation for a healthy system	Contestation as a sign of working processes	"A healthy system of 'okay, yes, we made a mistake, then a complaint is filed,' in the sense of an error culture. But complaints are also a high administrative burden for them, especially if they feel everything was calculated correctly, they have to reopen it so that the same decision is made again." (P11)
	Continued interest	Would engage with information beyond the study	"I'd be interested, maybe what's still missing are the training parameters the system was trained on. In the end only the criteria are given, but not the raw material, so to speak." (P10)

Continued on next page

Real-world implementation of interface (continued)

Subtheme	Code	Description	Examples
	Escalated contestation	Organizing opposition at management level	"This isn't a proper notice for me. If it really looked like that, my very first step would be to escalate it all the way up to management, forward the notice, redacted. [...] On one hand seek cooperation, and on the other try to get a commitment at a higher level." (P11)
	Intervention network	Places that help with contestation	"If we say we're now getting such automated benefit terminations, my first step would be to see how counseling centers react to that." (P11)
	Interaction with agency	How it feels to interact with the social agency	"There are meetings where the social agency is present and from their perspective they also address difficulties with social organizations or other partners, and together we look for solutions. Often those meetings are about understanding why things are done a certain way [...] and you might be angry about certain decisions if you don't know the details behind them." (P11)
	Legal aid	Supporting counselors in legal aid	"It would be a great help to me if I worked in the legal department. And you could perhaps use it for other things as well, there are countless proceedings and rulings." (P3)
	Reality of social work	How social workers operate	"I sometimes feel like a mini-computer, a bureaucratic compass, like a mother sending her children to school." (P2)
	Required knowledge	What one must know to use the website	"It would be useful in the counseling center, yes, but it's often about much more, entitlements to different benefits. We already tried integrating that into the website, but it was too complex; in any case you have to acquire some basic knowledge." (P3)
	Trust in site	Whether to trust the website	"Who created the site?" (P11), "Because of the information about the calculator and the flowchart I also have more trust in the site." (P11)
	Trust per peers	Would trust it if others do	"Well, you have to file an appeal; it will probably spread quickly that this is how you have to do it." (P2)