

Better Together? The Role of Explanations in Supporting Novices in Individual and Collective Deliberations about AI

TIMOTHÉE SCHMUDE, University of Vienna, Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science, 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

Deploying AI systems in public institutions can have far-reaching consequences for many people, making it a matter of public interest. Providing opportunities for stakeholders to come together, understand these systems, and debate their merits and harms is thus essential. Explainable AI often focuses on individuals, but deliberation benefits from group settings, which are underexplored. To address this gap, we present findings from an interview study with 8 focus groups and 12 individuals. Our findings provide insight into how explanations support AI novices in deliberating alone and in groups. Participants used modular explanations with four information categories to solve tasks and decide about an AI system's deployment. We found that the explanations supported groups in creating shared understanding and in finding arguments for and against the system's deployment. In comparison, individual participants engaged with explanations in more depth and performed better in the study tasks, but missed an exchange with others. Based on our findings, we provide suggestions on how explanations should be designed to work in group settings and describe their potential use in real-world contexts. With this, our contributions inform XAI research that aims to enable AI novices to understand and deliberate AI systems in the public sector.

CCS Concepts: • **Human-centered computing** → **User studies**.

Additional Key Words and Phrases: explainable AI, understanding, deliberation, qualitative methods, focus groups

1 INTRODUCTION

A growing number of AI systems¹ are deployed in the public sector to decide about critical issues, such as employment, migration, and criminal justice [4, 24, 108, 132]. These systems can have consequences for all stakeholders but tend to have the largest impact on their decision subjects (people the system decides over), such as discrimination or misclassification [13, 102]. These harms intensify when decision-making is opaque and uncontestable [1, 3, 32]. For these reasons, public AI systems should be considered as ‘matters of public interest’ [132], meaning that they need to be explainable, justifiable, and open for public deliberation [13, 61, 93]. Explanations can make AI systems more understandable and easier to assess and control [71]. Much of explainable AI (XAI) research is focused on individuals, but research has shown that group settings can facilitate the understanding of complex topics [90, 94, 97]. Further, group settings encourage the exchange of views and arguments [113, 116], which are vital when engaging in deliberation

¹We use the term ‘AI system’ to describe algorithmic systems with machine learning components. The nomenclature follows research on explainable AI [71] and research on AI in the context of society [27, 132] and regulation [99].

(collectively finding a solution to a problem [48]). XAI has not explored in detail how explanations can be combined with group settings to leverage these benefits. In this paper, we aim to address this gap.

As arguably many people affected by decisions of AI systems have no technical background in developing or analyzing these systems, this work focuses on ‘lay people’ [34, 77, 112] or ‘AI novices’ [88]. Explanations for AI novices naturally have different requirements than explanations for AI practitioners, as they have different expertise [36], interests [71], and prior knowledge [21, 107, 120]. Explanation formats (e.g., visual, textual, dialogue) are known to impact AI novices’ understanding [10, 21, 120] but show inconsistent effects [10] due to contextual factors such as participants’ perceptions of the use case domain. A possible solution is the adaptation of explanations with personalization [28, 92, 112] and interactivity [7, 21, 47]. To this end, previous research has analyzed AI novices’ information needs [107] and developed explanations in collaboration with end-users [73, 126]. However, designs that can adapt to AI novices and that support their deliberation, especially in groups, are still rare [23, 93]. But these explanations are essential to provide opportunities where AI novices can learn about and discuss AI systems and to realize the principles of human-centered AI (engaging stakeholders and empowering people [111]).

Designing explanations that support understanding and deliberation for AI novices in both group and individual settings meets multiple challenges. Group composition and dynamics place special demands on explanation design [91], as explanations need to cater to a diverse set of information and format preferences [7, 11, 107]. They must further support a joint understanding process and collaborative interactions [78], such as sharing and combining, all while providing comprehensive information and remaining clear and navigable. We address these challenges by proposing a modular explanation design that spans four information categories (*data, system details, usage, and context*) from which users can select. Another challenge consists in validating explanation approaches qualitatively with the relevant stakeholder groups. Specifically, XAI research does not always include people from marginalized population groups, who are most likely to be affected negatively as decision subjects [13]. To address this, we conducted two focus groups with decision subjects to include their perspectives and voices on AI systems in the public sector.

To examine the role of explanations in supporting AI novices’ understanding and deliberation we present the findings of a task-based interview study with 43 participants, involving 8 focus groups and 12 single interviews. For this study, we used an explanation design comprising 36 single explanations in question-answer pairs. These explanations are organized into the four categories *data, system details, usage, and context* and further assigned to subtopics and levels of detail (Figure 1). Participants used these explanations to solve the study tasks and decide about deploying a public AI system (Figure 3). We used an employment scoring algorithm that connects to previous work on AI systems in employment [79, 96, 108]. Our analysis examines participants’ self-reported understanding, decision confidence, and perceptions of key information. We further conducted a thematic analysis of how participants interacted with explanations in both settings. The following research questions guide the analysis:

[RQ1] *Explanations:* How does a question-driven, modular explanation design support AI novices’ understanding in groups and individual settings?

[RQ2] *Deliberation:* How do AI novices use explanations to form opinions and make decisions about AI systems?

Our contributions include i) an explanation design that builds on a question-driven and modular design to accommodate different levels of completeness and soundness and that is suitable for both individual and group settings; ii) an in-depth description of how explanations support participants’ understanding and deliberation processes that identifies salient differences between the two settings; iii) an analysis of which type of explanations participants requested most often and perceived as most important; and iv) recommendations regarding the design and use of explanations in group

settings. We envision that this work can provide valuable starting points for future XAI research that aims to connect explanations to deliberation on public AI systems.

2 BACKGROUND AND RELATED WORK

This section describes how our work is embedded in human-centered explainable AI and outlines the main challenges and approaches to designing explanations for AI novices. It further introduces the two main lenses of analysis to answer our research questions: understanding and deliberation.

2.1 Human-centered explainable AI

Explainability is often described as a cornerstone of responsible AI systems [121], as explanations can enable stakeholders such as users and decision subjects to understand [71] and contest AI decisions [1]. A similar focus is set by the domain of human-centered AI [17], which proposes to build AI systems that 1) are based on user-experience design and stakeholder engagement, and 2) empower rather than replace people by being controllable and autonomy-preserving [110, 111, 130]. These principles become especially important in high-risk settings [40], such as employment [42, 108], immigration [4], or criminal justice [24], where erroneous or non-transparent algorithmic decisions can cause severe harm to those affected [102]. In response to these risks, the domain of *human-centered explainable AI* (HCXAI) examines how explanations can contribute to “equitable and ethical Human-AI interaction” [38]. It assumes that transparency alone is not enough to make AI systems explainable [3], but that explanations need to consider the system’s social context [127], its lifecycle [35] and its different stakeholder groups [38]. In the context of this work, human-centered explainability is realized by testing and validating a design approach intended to support AI novices in understanding AI systems and deciding about their deployment in public institutions [132].

2.2 Designing explanations for AI novices

The majority of people who interact and are involved with AI systems in public institutions are lay people or *AI novices*, here defined as “users who [might] use AI products in daily life but have no (or very little) expertise on machine learning systems” [88]. Established explanation methods, like LIME [103], SHAP [80], and surrogate models [89] are tailored to experts and require technical knowledge; hence, they do not address the needs of AI novices. To better cover these needs and match information to them, it is necessary to understand how non-experts conceive of AI systems. Previous HCI research has analyzed lay understandings to explore user perception and understanding of several algorithmic systems [34, 39]. Similarly, XAI research has begun to explore the information needs of AI novices to design suitable explanations for a broader audience [107]. However, few explanation designs have been proposed that truly assume the perspectives of AI novices [21, 120]. In the following, we summarize current approaches with respect to AI novices’ information needs and current practices of explanation design.

Regarding **information needs**, previous qualitative research outlined that AI novices value information about the context and intention of a system’s deployment [61, 107] as well as about the responsible institution [13]. In contrast to traditional XAI approaches, which focus on descriptive information about the system’s workings and outputs, explanations for AI novices thus also require normative information, such as justifications [8] for design choices. Regarding **information coverage**, previous work posits that transparency does not equal understanding [3] and that simply making all information about a system available is no valid explanation strategy. Empirical evaluations of

this claim showed that “white-box” explanations (transparent models) can improve “objective”² understanding but may overwhelm non-expert users and reduce perceived understanding [21]. However, later work [11] used similar explanations and found that they had the opposite effect on understanding, attributed to a difference in the studies’ use case domains (student admission vs. finance). This indicates that the amount of information should be adaptable to the given context. Regarding **explanation format**, Szymanski et al. [120] examined how expert and lay users rated explanations of different formats and found that while lay users favored visual explanations, they performed better with textual ones. Other studies confirm this discrepancy and posit that comprehension varies with demographic factors and domain knowledge [36, 112, 123]. These issues are assumed to be addressed with **personalization of explanations** [28, 112], meaning that they are selected and designed according to the user’s stakeholder role [71], prior knowledge [107], beliefs [87], and explanatory stance [16, 64]. Further aspects to be considered include the explanation’s purpose [43] and the user’s familiarity with AI [66]. These approaches guided the conceptual development of the explanation design presented in this study.

We compiled information from different sources documenting employment prediction algorithms [2, 108], producing an extensive collection of “scavenged” [128] material. To structure this collection, we drew from work on intelligibility types [76], question-driven explanation design [74], and the separation of information categories into *data*, *system details*, *usage*, and *context* [107]. We further applied the principles of explanation soundness (fidelity, complexity) and completeness (coverage, density) [19, 47, 68] by introducing a structure of sub-topics and a hierarchy of explanation levels. This combination of question-driven explanations, levels of detail, and user-controlled selection of information aims to support modularity and interactivity [21, 47, 106] as well as the adjustment of explanations to users’ needs [28, 112]. Section 3.2 describes how these principles were realized in the explanation design.

2.3 Analytical lenses: Understanding and deliberation

In the following, we introduce understanding and deliberation to serve as the main analytical lenses for this paper. Section 3.5 then operationalizes them for the evaluation of the explanation design and settings.

2.3.1 Individual and collaborative understanding of AI systems. Improving understanding of an AI system is the primary purpose of explanations, as understanding is thought to enable assessment (e.g., of a system’s fairness) [71] and action (e.g., contestation) [50] for the system’s stakeholders. However, understanding can be defined in numerous ways [6, 46, 64, 131]. This work draws from research in learning sciences, cognitive sciences, and explainable AI to define understanding as i) connecting and applying information [6, 46], ii) being the attempt to grasp the underlying structure of a phenomenon by way of simplification [131], iii) consisting of several “facets” that include both the analytical and the emotional connection to information (explain, interpret, apply, take perspective, empathize, self-reflect) [129], and iv) being a “working” mental model that is attained by recognizing and filling gaps until the learner deems it sufficient [64]. Due to the challenge of defining and measuring understanding [105], recent research has proposed an “abilities-based” approach [115], connecting to comparable operationalizations by the learning sciences [129]. We examine understanding by analyzing which facets of understanding participants use to answer the study tasks and make a confident deployment decision (Section 3.4).

While individual understanding has been the subject of many studies in XAI [20, 21, 106, 123], understanding in group settings has been less explored [22]. We thus draw from disciplines that have investigated collaborative understanding: The cognitive sciences have examined distributed cognition (sharing cognitive load) and outsourcing [64] (delegating

²We use the term in reference to Cheng et al. [21] and Bove et al. [10], it means to describe factual or testable understanding.

understanding) as fruitful mechanisms for collaborative settings, such as the navigation of a ship [63]. The prerequisite is that groups achieve “cognitive symbioses with mutually supporting roles” [64], i.e., a constructive working dynamic. Similarly, educational psychology has found that peer discussion [113], collaborative reasoning [90], and aggregated knowledge [94] leads groups to perform better than individuals on the same tasks. However, whether groups perform well depends on their interactions, which can be described with cognitive and social mechanisms of collaborative success and failure [97]. When groups perform worse than individuals, the associated mechanisms include increased memory load and retrieval disruption (losing train of thought). In contrast, members tend to have established common ground and shared task-related information when they perform better. Thus, while it is not clear from the outset if groups are better for learning than one-on-one settings [9], their advantages, such as sharing of cognitive load and exchange of views, likely support finding solutions to complex problems and present a valuable testing ground to deliberate deployment of AI systems.

The field of computer-supported cooperative work (CSCW) has long examined collaborative settings about group composition and interactions [29, 41, 117]. However, work in XAI has only begun to consider how explanations for group interactions could be approached, describing that “many-to-one” interactions (multiple people interacting with an explanation) will likely differ from “one-to-one” interactions due to “complexities in group dynamics, cognitive bias amplification, trust issues within the group, and group-centric evaluation” [93]. Lastly, previous work in XAI has examined individual versus group understanding in AI-assisted decision-making but surprisingly found little effect on understanding [22]. Following up on these findings, we use a ‘triangulation of methods’ [18], as described in Section 3.5), to empirically explore and compare the effects of explanations on the understanding processes of AI novices in groups and individual settings.

2.3.2 Deliberating on AI systems. Deliberation, in the sense of informed reasoning and decision-making, is based on understanding [32] and is key in enabling citizens to debate public sector AI proposals and their potential consequences [61, 132]. Habermas [48] describes deliberation as the exchange of rational-critical arguments on a problem to the end of finding a solution. These rational-critical arguments are grounded in truth or a *shared understanding* of reality, are open for judgment, and can be defended. This connection between shared understanding and deliberation is central to our examination of explanations’ effects. Deliberation takes place in many areas that shape politics and life in society [81]. Examples include public referendums that let inhabitants vote on jurisdictional changes (such as Swiss federal and state laws [118]), citizen forums addressing matters of public importance (such as water supply in California [54]), and community-based grassroots formats where citizens support each other (such as the right to repair movement [27]). These settings have in common that they involve “social entities made up of people who are in one way or another engaged with their environment” and who use deliberation and productive conflict to negotiate and change policy issues [49]. While these participation formats are not perfect and potentially incur cognitive biases such as *groupthink*³ [5, 55, 93], they create spaces where the general public can gather, discuss, form opinions, and decide on public interests. We argue that AI systems in public institutions constitute such public interests, encapsulated in the term *public AI*. Züger and Asghari [132] employ the term to make explicit that AI systems in public institutions must fulfill obligations to prove their benefit. These obligations include being justifiable, equal, open to validation, technically secure, and the result of a *deliberation or co-design process*. Identifying formats supporting this deliberation on public AI systems is an open research challenge. Prior work has investigated how ‘mini-publics’ [44] can be used to support the co-design of algorithmic policy [73] and procedural justice in algorithmic resource allocation [72]. HCI

³Prioritizing group harmony over real argumentation and discussion.

research has further shown that participatory formats can connect communities and institutions in public service transformation [31]. And in XAI, studies showed that group discussions can facilitate the critical analysis of an AI system's recommendations and that supplying information on both pros and cons of an AI's recommendation lead to more frequent and more productive group deliberation [22].

However, settings that allow participants to deliberate in person on the deployment of high-stakes public AI are underexplored. We aim to address this gap by implementing mini-publics as focus groups with three different compositions (domain experts, decision subjects, and members of the general public⁴), thus including stakeholders of different backgrounds and degrees of involvement. We further compare group deliberation processes with those in single interviews, which can be described as “internal deliberation” [85]. On this basis, we aim to provide insight into suitable explanation designs and social formats to support deliberation on public AI systems. Section 3.5 describes the concrete analysis approach to this end.

3 METHODS

In this section, we describe our methods and study procedure. We conducted a task-based semi-structured interview study with 43 participants (Section 3.6), structured into 8 focus groups with 3–5 participants each and 12 single interviews (Section 3.3). Participants were presented with the study’s employment prediction use case (Section 3.1) and a collection of explanations about this system (Section 3.2) before solving four tasks and deciding about the system’s deployment (Section 3.4). The study closed with an interview, lasting 90–120 minutes for focus groups and 60 minutes for single interviews. We analyzed individual and collective interactions with the explanations, self-reports, and deliberation processes (Section 3.5). The university’s research ethics committee approved this study.

3.1 Use case: The AMS employment prediction algorithm

3.1.1 Description. The *AMS algorithm*⁵ is a system developed to calculate the employability of job-seekers in Austria. It was created by a private company for the Austrian Public Employment Agency between 2015 and 2021 but was never used as a live system and put on hold in 2021 due to privacy objections [2]. The system uses a logistic regression model trained on historical data to predict job-seekers’ employment chances based on demographic features (such as age, education, nationality, etc.) and work history. The outputs are a short-term and long-term employment score for each job-seeker [45]. These scores would serve as recommendations for the job-seekers’ counselors at the employment agency to assist in deciding about suitable support measures. Counselors could overwrite the system’s predicted scores of job-seekers but would need to give a reason for doing so [2, 53]. More information is provided in the appendix.

3.1.2 Choice of use case. Algorithmic tools that assist in assessing job-seekers and resource allocation have been deployed in various countries, including Germany [14], Austria [2], Poland [96], and the Netherlands [33]. However, the introductions of these applications also repeatedly led to sociotechnical conflicts [108]. The deployment of the *AMS algorithm* was motivated by three overarching goals: a) increasing consultation efficiency, b) increasing support measure effectiveness, and c) reducing arbitrariness [45]. Detailed reports warned that counselors might over-rely on the algorithm or hesitate to overrule its suggestions [2]. Further, the algorithm’s model and underlying data structure were predicted to discriminate against marginalized groups, who would lack the option to contest the system itself [79]. Transparency and ongoing scrutiny of the algorithm were listed as necessary measures to prevent these risks [2]. As the

⁴Participants who were neither directly affected as job-seekers nor were potential users of the system.

⁵AMS stands for the Public Employment Agency (Arbeitsmarktservice).

AMS algorithm represents a larger class of algorithmic decision-making systems that spark public debate around their deployment in public institutions [102], it exemplifies how AI systems become matters of public interest and presents a suitable use case for our study.

3.2 Explanation design

3.2.1 *Description.* The explanation design comprised 36 question-answer pairs about the *AMS algorithm*. Each question belonged to one of four categories, *data* (format, content, limitations), *system details* (features, model, examples), *usage* (operation by and interaction with users), and *context* (intention of deployment, target group, responsible actors). Each category was further divided into topics with three levels of increasing detail (base level, level 2, level 3). Every explanation was printed on an A5 paper sheet and contained a question (e.g., "Who operates the system?") answered with a brief text or image (cf. Figure 2). Participants first received an explanation overview (Figure 1) and the four base explanations and could request levels 2 and 3 at any time during the explanation phase (as depicted in Figure 3). The explanations were presented in an analog paper format that allowed participants to interact with them physically and that facilitated social interactions, such as sorting, exchanging, pointing, and reading to each other. The full collection of explanations is depicted in the appendix.

3.2.2 *Design foundations.* The explanation design was intended to allow the users to explore information in a flexible and self-directed manner. To this end, the design used a *modular structure*, meaning that the explanations were divided into four information categories, which were again subdivided into topics and three levels of detail (base level, level 1, level 2). The four information categories, *data*, *system details*, *usage*, and *context* were based on research on AI novices' information needs and covered both technical and sociotechnical system aspects [107]. The subdivision of explanation categories into topics and levels of detail organized this broad supply of information while accommodating different needs of information completeness and soundness [69] and introducing a degree of personalization [19]. The goal of the modular design was thus to create explanations that offered information on every aspect of the AI system, from which participants could select the most relevant according to their information needs and preferences. It further aimed to avoid limitations of "groupware" systems [83] by supporting both individual and collaborative interaction, providing multiple user perspectives (e.g., user, decision subject), and synchronizing interaction with the material.

The question-answer style was motivated by explanation design research [74, 75] and was intended to improve user engagement and understanding by matching their thought processes. For example, when users learn that the system uses features to calculate job-seekers' employment scores, they might ask what the exact weights of these features are and how they are calculated. This corresponds to the three levels of detail in topic A of *system details* (Figure 1). The explanations further used different explanation methods [114] such as feature importance, local and global explanations, examples, counterfactuals, and argumentative approaches (Figure 2). Information was presented in different formats but mostly relied on textual information and used highlighting, colors, and illustrations to emphasize key points.

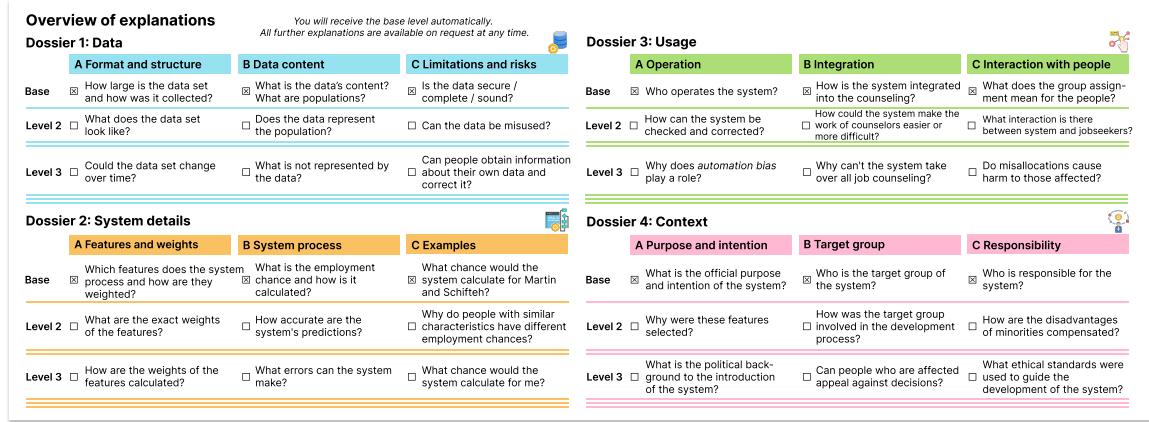


Fig. 1. Overview of explanations. Explanations were designed as a collection of 36 question-answer pairs. The questions were assigned to 4 categories, *data*, *system details*, *usage*, and *context*, each containing 9 questions. Participants received the base explanations at the beginning of the explanation phase, as indicated by the ticked boxes, and could request all other explanations at any time during the explanation phase using this overview.

Dossier 1: Data Base

A Format and structure

How large is the data set and how was it collected?

Scope: 860,277 entries on business cases. A business case refers to the period in which a person is unemployed and registered with the employment agency. This means that several business cases can exist for one person.

Period: The data describes a period over the last four years.

Storage location: Data warehouse on the employment agency's server.

Collection: The data was taken from the database of the Association of Social Security Institutions. Consultants at the agency can supplement or correct the data if necessary.

Dossier 3: Usage Level 2

B Integration

How could the system make the work of counselors easier or more difficult?

Easier

- Provides an **overview** of any relevant information
- Provides **guidance** for assessing the chances of jobseekers
- Enables a judgment that is not based **solely on the view of the advisor**
- Can be used to **legitimize** decisions

More difficult

- Key figures can **limit the advisor's scope** of discretion
- Obligation to give reasons when correcting the decision can be **time-consuming**
- Interaction with jobseekers could become **too formalized**
- Requires appropriate **training**

Dossier 2: System details Level 2

A Features and weights

What are the exact weights of the features?

The exact weights depend on the population. Here are examples of the weights for the population with complete information:

Feature	Value	Weight	Feature	Value	Weight
Gender	M	0	Only care for children or family	No	-0.15
	F	-0.14		Yes	0
Age	<30	0	Professional group	Production	+0.17
	30-49	-0.13		Service	0
	50+	-0.7	Regional labour market	1/2/3/4/5	0/-0.34/-0.18/-0.83/-0.82
Nationality	Employment country	0	Employment history	<75% employed in year	-0.74
	EU	-0.08		75% employed in 4 years	0
	Others	-0.05	How often job-seeking in 4 years?	0/1 time / 2 times / 3 times	0/+0.63/+1.19/+1.08
Education	Compulsory school	0	Job-seeking for how long?	Less than 6 months	0
	Apprenticeship	-0.28		At least 1 time more than 6 months	-0.8
	High School or higher	-0.02	Participation in support measures	None	0
Impaired health	Yes	-0.67		1 time supporting	-0.57
	No	0		1 time qualifying	-0.21
				1 time employment-promoting	0.43

Dossier 4: Context Level 3

B Target group

Can people who are affected appeal against decisions?

Affected persons **cannot legally appeal to be reclassified** to a higher category by advisors or to have control over the decisions made. However, they can address the group allocation in dialog with the advisors and **request a correction**.

In order to guarantee a right of appeal, a corresponding legal basis would have to be created. Another solution could be the establishment of an ombudsman's office, which those affected can visit to receive help or legal advice.

Fig. 2. Four explanation examples. Examples for explanations in the categories *data*, *system details*, *usage*, and *context*. Each question was printed on a sheet of A5 paper with a short answer to the question. Answers could be fully textual or complemented with visual elements like charts or colored shapes. Each category was given a different color and icon to facilitate navigation.

3.3 Study procedure

We describe the procedure in the focus group and single interview setting (depicted in Figure 3). Like previous work in XAI [22], we conducted individual and group settings to compare how the social setting would affect participants' understanding and deliberation processes.

3.3.1 Focus group procedure. Throughout the study, participants sat together with the investigator and could freely interact with each other. They first completed consent forms and questionnaires about demographics and knowledge about employment (domain knowledge) and AI systems (technical literacy). A round of introductions followed, where each group member stated their name and last interaction with AI to break the ice. The investigator then explained the study procedure and distributed a mock newspaper article introducing the AI use case (included in the appendix). Participants indicated their understanding, deployment decision, and decision confidence for the first time (3.4). The study's explanation phase followed (including orientation, task, and decision phase), throughout which participants received and kept access to all explanations. In the beginning, the group received an overview of the explanations and all base level explanations, all other explanations could be requested at any time. After 15 minutes, participants received task sheets and had another 15 minutes to complete them, deciding independently whether they wanted to work together or individually. Finally, the group had 10 minutes to make a joint deployment decision (yes/no, with conditions allowed). A second round of individual reports followed, described in the next section, and then the investigator concluded the study. Focus group studies took around 90 minutes.

3.3.2 Single interview procedure. The majority of the study procedure remained the same in single interviews. However, the explanation phase did not include a group decision phase; instead, the orientation and task phases were prolonged to 20 minutes each to provide individuals with the same total time as focus groups to interact with the explanations and individual reports about the group setting were omitted. The individual study setting was included to analyze how understanding processes changed depending on whether participants worked in a group or alone and whether the explanation design would support both settings. Single interviews took around 60 minutes.

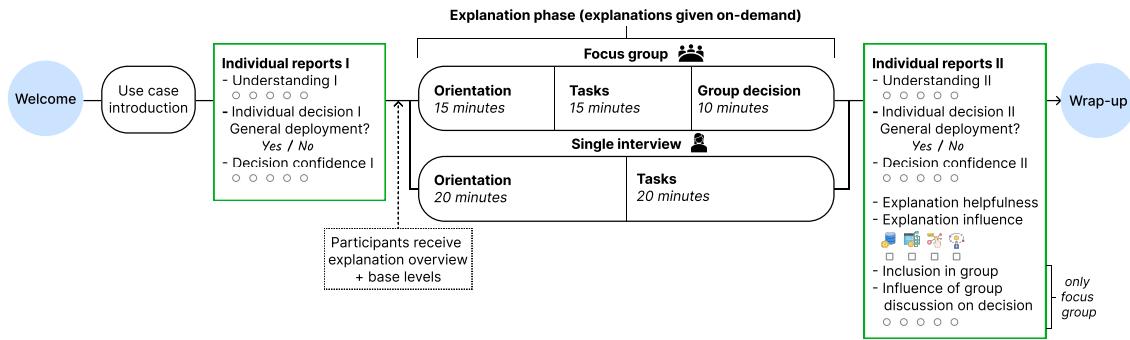


Fig. 3. Overview of the study procedure. Focus groups and single interviews differed only in the explanation phase and the questions for the second individual reports.

3.4 Study elements

This section provides descriptions and motivation for the study elements: the introduction of the use case, the explanation phase comprising orientation, tasks, and the group decision, and participants' individual reports.

3.4.1 Use case introduction. Participants received initial information about the *AMS algorithm* in the form of a mock newspaper article inspired by an Austrian newspaper publication from 2019 [119]. The article provided key information about the system's basic workings, goals, and deployment context and featured the opinions of employers and employee associations about its merits and risks. The presentation format was chosen to provide an introductory summary of the AI system using a familiar layout and non-technical language while highlighting both the pros and cons of the system's deployment. Thus, the article served as a basic introduction to the use case, which aimed to approximate the amount of information participants might receive through the media. This way, participants received a baseline of information with which to assess their initial understanding, deployment decision, and decision confidence. Further, this introduction served to outline relevant aspects that could be explored using the explanations.

3.4.2 Explanation phase (orientation, tasks, and group decision). For orientation, participants received the explanation overview (Figure 1) and base-level explanations (the first row of explanations in each category) and had 15 minutes (focus groups) or 20 minutes (single interviews) to become familiar with the structure and explore topics of interest. Explanations were provided as single A5 sheets to promote the physical sharing and exchanging of explanations. Participants could freely decide which explanations to request and read, and whether to share and discuss information with others.

For the tasks, participants received the case of Mr. Harald G.⁶, a fictional job-seeker with a brief backstory and a list of features. Participants had 15 minutes (focus groups) or 20 minutes (single interviews) to solve four tasks pertaining to this case. All questions could be answered with information from explanations in different categories and levels of detail. Whereas tasks 1, 3, and 4 required locating information, task 2 could be solved in two ways (aside from guessing): by either giving an estimate based on the rough weightings in the *system details* base explanations or calculating the precise employment score. Participants could access and request all explanations and discuss possible solutions. These were the four tasks (correct answers underlined):

Task 1: Can Harald change the data stored about him (e.g., to correct it)? (yes / no)

Task 2: In which group of employment chance does the system categorize Harald?

(high (>66%) / medium (<66% & >25%) / low (<25%))

Task 3: Which support measures will Harald receive?

(qualifying measures, such as courses and training / stabilization and increased supervision / none)

Task 4: Can Harald appeal against this decision? (yes / no)

For the group decision, focus group participants had 10 minutes to discuss the system's deployment and were asked to collectively decide whether to accept or reject it. This was meant to simulate a small referendum in which each participant's voice counted for the final outcome. If no consensual decision was reached in time, participants were asked how the situation should be resolved (e.g., majority vote). They were further made aware of the option to state conditions for the system's deployment.

⁶The case example was inspired by Allhutter et al. [2] and adapted to this study, as depicted in the appendix.

3.4.3 Individual reports I and II. Participants were asked for individual reports before and after the explanation phase. At both points, participants reported understanding (5-point scale), deployment decision (yes / no), and decision confidence (5-point scale) to examine the effect of the explanation phase. In report II, participants also reported the explanation categories that were most helpful for their understanding (multiple choice) and most influential to their decisions (multiple choice); focus group participants additionally reported perceived inclusion in the group and the discussions' influence on their decision (5-point scales). During report II, the investigator asked participants interview questions about their interaction with the explanations, understanding processes, prioritized information, and situational aspects. The list of interview questions is included in the appendix.

To prevent influence between participants' reports, individual reports in focus groups were conducted anonymously and re-assigned by the study examiner using a color-coded reporting system (Figure 4). Each participant was assigned a color and received the material for all ten individual reports. Participants took the corresponding paper slip for each report, wrote their answer, and threw it in a gathering container that hid it from view. The gathered reports were then collected and recorded by the investigator at the end of the study.

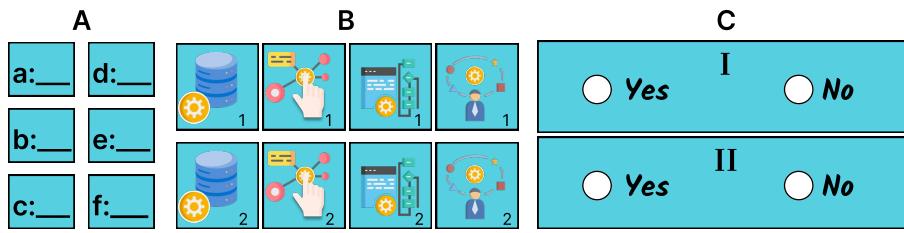


Fig. 4. Material for individual reports of participants. Participants received the materials for individual reports on laminated paper slips in different colors (blue, yellow, red, green, and grey) and used them to answer questions individually. Slips that were numbered with letters a to f (A) served as 5-point scales for understanding, confidence, inclusion in group, and influence of group discussion and were answered by writing a number (1–5); slips with icons and a corresponding textual description (B) served as selection of the most helpful and influential explanation categories and were answered by selecting any number of icons; slips with decisions (C) served as voting ballots for deployment decisions and were answered by ticking yes or no.

3.5 Analysis

All focus groups and interviews were audio-recorded and transcribed. These transcripts provided the data basis for the thematic analysis. Participants' individual reports, task solutions, decisions, and the investigators' field notes provided further data for within- and between-subject comparisons of understanding and decision-making.

3.5.1 Thematic analysis. For both research questions, we conducted thematic analysis [12] of participants' articulations to develop a qualitative account of their understanding and decision-making processes. To this end, we developed inductive codes in the first pass and refined them in the second pass on the transcriptions. The resulting inductive code base was structured along the overarching categories of understanding, deliberation (decision-making processes, arguments), opinions (e.g., about AI and policy choices), and experiences (e.g., anecdotes and lived situations). The full code-book is provided in the appendix. We further highlight that while the quantitative items in participants' self-reports serve to characterize the diversity of participants' perceptions and facilitate qualitative exploration [125], they are not intended to invoke "inference [...] of greater generality" [84] nor impose a mental model based on variance theory [101].

3.5.2 RQ1-Explanations. RQ1 focuses on the explanations' impact on participants' understanding of the study's use case and the differences between individual and group settings. We employ a triangulation approach [18] by using three ways of analyzing the explanations' effect on participant understanding. Firstly, comparing participants' individual reports before and after the explanation phase was a subjective indicator of changes in their understanding and decision confidence. Secondly, participants' answers to the four study tasks indicated their factual or testable understanding. Lastly, participants' verbal reports during and after the explanation phase were used to analyze their understanding processes and barriers thematically. In a deductive analysis, we further compared their interactions to mechanisms of "collaborative success and failure" [97] and the "six facets of understanding" [129]. With this three-part combination, we examined participants' subjective understanding, their information gain, and the cognitive processes of their understanding. This choice was motivated by educational psychology research outlining that understanding cannot only be elicited through test questionnaires [105] but involves emotional [129] and meta-cognitive processes [122] that are equally important. Our focus on understanding is motivated by previous XAI research, which highlights the importance of understanding in decision-making processes [52, 71, 110].

3.5.3 RQ2-Deliberation. RQ2 focuses on how participants formed opinions about the *AMS algorithm*, weighed the pros and cons of its deployment, and settled on a deployment decision. To this end, we compare participants' decision confidence before and after the explanation phase. We further conduct an inductive and deductive thematic analysis of participants' interactions in both settings to connect them to the "elements of deliberation" [116] – a set of characteristics that outline deliberation processes. Based on this, we analyze when participants used arguments (grounded, defensible positions), opinions (personal judgments on things, values, states), and personal experiences [85, 116] to consider the system's deployment. For single interviews, we examine participants' responses to interview questions during the study to examine their reasoning process and "internal deliberation" [85]. Lastly, to account for one of the most prevalent cognitive biases in group settings, we examine focus groups for occurrences of groupthink [5, 55] – an effect that sets in when concurrence-seeking in groups overrides realistic argumentation and discussion.

3.6 Participants

3.6.1 Recruitment. Table 1 and 2 provide an overview of the study participants. Participants were recruited through cooperation with civil society organizations, an employment agency, public calls for participation, and the authors' extended network. For focus groups, the authors contacted staff from these organizations known from previous studies or used channels of general inquiry to describe the study and invite participation. Interested organizations all offered to support the recruitment process by coordinating with the authors on selecting and inviting potential participants and finding a place and time to conduct the studies. Groups A, B, C, E, F, G, and H were organized this way. Group D was recruited through the authors' network and was equally composed of people who had previously been job-seeking. For individual studies, participants were recruited using the same channels, and calls for participation were further advertised on screens and information boards throughout different city districts. All studies were conducted in person in office or public spaces. Participants were compensated with 30€ for participation in focus groups (90–120 minutes) and 20€ for single interviews (60 minutes). Our approach for organizing, composing, and moderating focus groups was informed by Krueger [67]. Concerning the participant sample size, we are guided by research on qualitative methods, which suggests that the number of participants should be determined by code and meaning saturation [51].

3.6.2 Recruitment criteria. Participants were required to be of full legal age and AI novices, i.e., to have no technical knowledge or expertise about machine learning systems as described in Section 2.2. These criteria were screened

in a pre-questionnaire before invitation to the study using two questions: “How would you rate your knowledge of algorithms?” and “How would you rate your knowledge of artificial intelligence (AI)?”. Each question could be answered on a scale corresponding to “no knowledge at all” to “professional and detailed knowledge”. Here, the first question elicited technical knowledge, as participants familiar with AI tools might have rated their AI expertise as high but were unlikely to know about algorithms without a strong technical interest or background.

3.6.3 Group composition. Participants were further selected to be representatives of one of three roles: domain experts, decision subjects, or members of the general public. We define domain experts as people who are competent in the field that the AI system is used in, such as job counselors or advisors (groups B, C, G, and H). We define decision subjects as people who would potentially be impacted by the system’s decision, such as job-seekers and people who had previously been job-seeking (groups E and F). All remaining participants are considered members of the general public and were included to test changes in explanation effects and participants’ perceptions (groups A and D). The study was conducted with separate participants in three pilot groups to test and refine the explanation design and study procedure.

Table 1. Details on the study participants in the focus groups.

Group	ID	Age	Education	Occupation	Group	ID	Age	Education	Occupation
Group A	A1	63	University	Retired	Group F	F1	48	A-levels	Job-seeking
	A2	69	Secondary school	Retired		F2	35	n/a	Job-seeking
	A3	63	Vocational school	Retired		F3	49	A-levels	Job-seeking
	A4	70	Vocational school	Retired		F4	50	Vocational school	Job-seeking
Group B	B1	46	University	Social counselor		F5	48	A-levels	Job-seeking
	B2	76	A-levels	Retired	Group G	G1	37	University	Executive staff
	B3	46	University	Social counselor		G2	49	University	GDPR officer
	B4	70	A-levels	Retired		G3	44	Secondary school	Training counselor
Group C	C1	60	Apprenticeship	Personnel counselor		G4	58	University	Executive staff
	C2	60	University	Personnel counselor	Group H	H1	37	University	Team lead
	C3	51	Apprenticeship	Job trainer		H2	56	Apprenticeship	Job trainer
Group D	D1	65	University	Business consultant		H3	45	University	Job trainer
	D2	53	University	Retired		H4	43	University	Job trainer
	D3	52	University	Business consultant		H5	60	University	Administrative staff
Group E	E1	36	University	Graphic designer					
	E2	32	Apprenticeship	Job-seeking					
	E3	40	Apprenticeship	Job-seeking					

Table 2. Details on the study participants in the single interviews.

ID	Age	Education	Occupation	ID	Age	Education	Occupation
S1	74	University	Retired	S7	40	University	Job trainer
S2	29	A-levels	Nurse	S8	43	University	Rehabilitation counselor
S3	28	University	Social counselor	S9	44	University	Social center manager
S4	29	University	Doctoral student	S10	52	University	Rehabilitation counselor
S5	37	University	Administrative staff	S11	59	University	Social center manager
S6	28	University	Job-seeking	S12	39	University	Education program manager

4 RESULTS

In this section, we present our results as answers to our research questions: How question-driven, modular explanations⁷ support understanding in individual and group settings (RQ1, Section 4.1) and how AI novices used explanations to form opinions and decide about the system’s deployment (RQ2, Section 4.2). Participant labels denote the study setting (focus group: A–H / single interviews: S) and the participant ID, as listed in Tables 1+2. To distinguish themes in the analysis, *inductive themes* are italicized, while ‘deductive themes’ are put in quotes.

4.1 RQ1-Explanations: How does a question-driven, modular explanation design support AI novices’ understanding in groups and individual settings?

To examine how AI novices used the explanations to understand the study’s use case, we analyzed their self-reports, articulations, and interactions in both settings. We found that each setting supported different aspects of understanding, suggesting a trade-off. We first describe how the explanations contributed to *shared understanding* and ‘collaborative success’ in groups (4.1.1) and continue with the explanations’ role in instances of ‘collaborative failure’ [97] (4.1.2), summarized in Figure 5. We then describe individuals’ interactions with the explanations (4.1.3), participants’ feedback on the explanation design (4.1.4), and summarize the benefits and drawbacks of both settings for XAI (4.1.5).

4.1.1 Groups’ benefits: Shared understanding and increased engagement. In the best cases, groups leveraged the modular explanation structure to use distributed cognition [63], meaning that participants processed information in parallel and then combined it. We use the term *shared understanding* to capture interactions that realized distributed cognition. Examples of such interactions included *locating information together*, *sharing information with others*, *discussing interpretations*, *debating task solutions*, and *querying and explaining* (a question by a group participant invites other participants to contribute). The explanations only afforded this set of interactions to groups, as they required social interaction with other participants. For example, in Group C, participant C1 read the first study task aloud and asked for input (*querying*), after which the group discussed solutions (*explaining*):

- C1 Can Harald change the data stored about him? Yes, he can certainly change it, can’t he? [...]
- C3 Which stored data, the one down there? [points at Harald’s demographic features]
- C1 Yes, just that.
- C3 49 – no, male – no. The apprenticeship – no, Austria – he can still change that. Duty of care – he could get married or have children. He could change his service sector. He could change his career. Impairment...
- C1 Well, what is meant by ‘change’? When he enters the data, he can change the data. He doesn’t have to specify the knee problem. [...]
- C3 So he can change it.
- C2 Yes.

Interactions such as *querying and explaining* and *discussing interpretations* rely on collaboration between participants to ‘share working memory resources’, ‘complement others’ knowledge’, ‘re-expose information’, and ‘correct errors’. These aspects of collaboration are described as cognitive mechanisms of ‘collaborative success’ [97] and provide groups with multiple ways to tackle explanations. For example, participants tended to work through information about *usage* and *context* alone or in pairs but raised explanations with the group when they were difficult or piqued their interest. This was often the case with explanations about *system details*, which included the most numerical information but also were

⁷We note again that with “explanation” we mean a question and answer pair and with “explanations” we mean the collection of all 36 explanations (Section 3).

an important key to understanding the system's calculations. We describe the process of using other's understanding to close gaps in one's own as *outsourcing* [64]. As understanding AI systems involves interacting with a variety of different information categories (e.g., technical, political, social), outsourcing provides a way to hand information to the team member most competent in this category. For example, in Group B, B2 expressed their appreciation for B3's help in solving study task 2: "*It was a math problem. You [B3] filtered it out well. It was very analytical. With your help, we were able to recognize these weak points.*" In contrast to *querying and explaining*, which participants used to invite input or spark conversation, *outsourcing* was thus used for the active delegation of an impeded understanding process.

Whether groups used this collaboration depended on participants' relationships and the group's social dynamics. The explanations also served to support these 'social mechanisms' of collaboration, by drawing the group's attention to certain aspects of the AI system and encouraging them to share their experiences and opinions. We present an excerpt from Group G as an illustration. Here, participant G4 *shared an explanation* that documented the algorithm's impact on two job-seekers ('joint management of attention'), which prompted G2 and G3 to *discuss interpretations* ('increased engagement'). This interaction established 'common ground' that the group later used for deliberation:

G4 *That's bad, the two of them. Look, "What chances would the system calculate for Martin and Schifteh?"*

G2 *Schifteh is probably worse off, isn't she?*

G4 *Schifteh has a 30% chance of employment and Martin 52%, even though Schifteh has a degree and would be working in the IT sector. And Martin has compulsory schooling and works in the cleaning sector. Martin's chances of employment are almost twice as high as Schifteh's. [...]*

G3 *I think that's a bit weird. [...] Because if she can speak English very well and has the specialist knowledge that our IT sector needs...*

G4 *She even gets two minuses for living in Favoriten [a city district]. [...]*

G1 *Yes, and here you have it in writing, I'll have to look at that too.*

A later excerpt in Section 4.2.3 further shows that explanations also led participants to 'negotiate multiple perspectives'. Addressing both sides of collaboration, cognitive and social, here is an important goal when supporting group understanding. It can be assumed that when explanations succeed in doing both, they can provide participants with more comprehensive and more complex information than individual settings. In line with this, previous work in HCI has found that group interaction boosts task performance compared to individual settings [59]. However, in our study, individuals surprisingly performed better in the study tasks than groups, as discussed in Section 4.1.3. Even so, we argue that the set of interactions (*shared understanding*) enabled through the combination of our explanation design and the group setting presents important pathways to help AI novices understand algorithmic systems. These interactions can be especially useful when group members have different domain expertise and information needs, as they can use complementary knowledge, memory, and perspectives to make sense of information. At the same time, as noted earlier, these interactions are partly dependent on the group dynamic. Positive interactions like those described were especially frequent in Groups G and H (job counselors and trainers), where participants knew and trusted each other. In contrast, the next subsection describes instances where groups encountered challenges in understanding, illustrating the importance of social mechanisms.

4.1.2 Groups' drawbacks: Process loss and susceptibility to social dynamics. In some of the focus groups, participants lost track of information, forgot their train of thought, or abandoned understanding altogether. We summarize these effects under the term *impeded understanding* and its final result as *abandoned understanding*. We found that *impeded understanding* occurred due to *explanation design flaws* and co-occurred with adverse social dynamics, resulting

in ‘process loss’ (groups falling short of their potential performance [65]). For some participants, the benefits of the explanations’ modular structure turned into disadvantages when it hampered them in navigating and retrieving information. Such impeded interactions included *cumbersome information uptake*, being *overwhelmed by information*, and *relying on intuition over information*. Further, for some participants, the group setting contributed to these impediments. For example, participant B3 stated that “*For me, it doesn’t make sense to [...] split up [the explanations], and everyone reads a part, that’s actually not enough.*” As before, these impediments can be connected to cognitive mechanisms of ‘collaborative failure’ [97]. When groups had difficulties in interacting with the explanations, they also incurred ‘memory coordination cost’ (increased cognitive load) and ‘retrieval strategy disruption’ (losing train of thought). We illustrate these mechanisms with an exchange from Group G. Although this group was composed of participants with university education, it did not succeed in calculating the employment chance for study task 2, in contrast to single interview participants with the same education.

- G4** *We can go through the features briefly. Where is the piece of paper with this terrible matrix? [...]*
- G2** *I still don’t understand which value to put. To calculate it, I need an exact value for the weighting.*
- G1** *You can calculate it with this. The apprenticeship has 52%. I believe that he [Harald] has over 25%.*
- G3** *Yes, definitely, I mean, roughly speaking...*
- G1** *She [Shifteh] has over 30%. And she also has 2 minuses [...] and a plus.*
- G3** *That’s also how I estimated it. [...]*
- G2** *But how do you calculate it? [...] And why are there differences between the general weighting and the exact calculation? That doesn’t click for me right now.*

Here, the levels of detail in the explanations acted against participants’ understanding by obscuring the actual feature weights, which were only accessible in level 2 of *system details*. While improved navigation might solve this issue, it also shows the difficulty of simplifying information about AI systems without omitting key aspects. In the intention to provide an easier reading of the feature weights, which was an advantage in other cases, the explanations’ clarity was reduced, and information was obscured. Addressing all information needs of AI novices [107] thus leads to problems with information overload, as observed in previous work on ‘white-box’ explanations [21]. However, *impeded understanding* alone did not mean that collaboration failed; rather, it depended on how groups dealt with these issues. Here, the key aspects were group cohesion [65] and constructiveness [95]. When the group dynamic was unfamiliar, it gave room to negative social mechanisms, such as ‘social loafing’ (group loses motivation) and ‘fear of evaluation’ (being criticized by others), and participants began to *abandon understanding*. This suggests that when interaction between participants stopped, interaction with the explanations stopped as well. Limiting these negative dynamics and promoting positive ones must be a goal of both explanation design and setting.

Notably, these adverse social dynamics occurred most often in Groups E and F, which were composed of job-seekers. Participants had trouble engaging with the explanations and abandoned interactions and understanding by saying: “*Probably [you can solve it] with that, but I don’t know, I’m too stupid for that.*” (E3) or “*I don’t know what I should say. Everything has already been said.*” (F2). Here, two things failed: The explanations failed to make crucial information accessible, and the group setting failed to uplift members who were discouraged. Interactions that offset this discouragement, such as *locating information together* and *outsourcing*, were not realized in Groups E and F. We thus propose to use co-design approaches to make explanations viable for decision subjects, as has been done with public servants [126]. Further, XAI should employ methods that create a productive social dynamic, which we identify as the second key aspect to support ‘collaborative success’ and *shared understanding* (Figure 5).

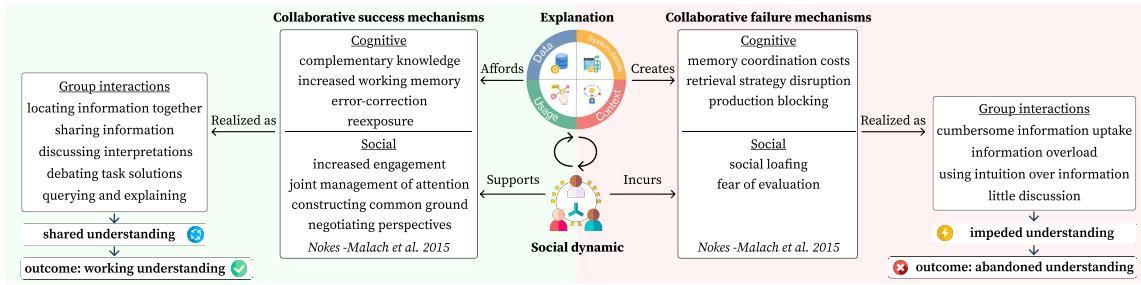


Fig. 5. Both explanation and social dynamic have an impact on collaborative performance. In focus groups, both explanations and social dynamic were key factors for the understanding outcome. If participants could engage easily with the explanations and each other, their interactions realized mechanisms of ‘collaborative success’ [97] and led to *shared understanding*. In contrast, if participants had trouble using the explanations and could not outsource or discuss these issues, interactions rather realized mechanisms of ‘failure’ [97] (Section 3.5) and showed *impeded understanding*. Depending on these intermediary steps, groups experienced the outcomes as *working* or *abandoned understanding*. From the perspective of XAI, both explanation and social dynamic are thus important aspects to keep in mind when designing explanations for groups in collaborative settings.

4.1.3 Why individuals performed better in the study tasks but still felt the absence of collaboration. Despite the different advantages and disadvantages that group and individual settings offer in learning contexts (Section 2.3.1), these differences are rarely examined empirically in XAI⁸. We address this gap by comparing in-person focus groups and single interviews to examine whether the social setting impacted participants’ understanding. We use a triangulation⁹ approach by investigating participants’ understanding with respect to three aspects: interactions with the explanations, task performance, and self-reported understanding.

Overall, the explanation afforded its core functionality to both individuals and groups. Interestingly, participants in single interviews tended to request the same number or more explanations than focus groups (Figure 6), despite having seemingly less working memory. S5 described: “Well, maybe [it was overwhelming] at the very beginning [...] But I then realized that I could get through it to some extent.” Single interview participants also regularly *consumed explanations in bulk*, i.e., read through the whole of an explanation category rapidly. This interaction was nearly nonexistent in focus groups. However, even though participants in single interviews performed better in the study tasks, they often stated that they missed the “*exchange with people, with other perspectives*” (S3). S8 explained that this exchange would allow for a different form of understanding:

I think that, on your own, you can think about it very intensely and [...] make up your own mind. But that’s also the disadvantage, making up your own mind. Others may have completely different thoughts and a different professional background. And that would probably have been an exciting exchange. (S8)

Participants in individual settings further performed better in the study tasks than groups (Table 3). A possible explanation is that participants in single interviews engaged differently with the study tasks, as they often calculated the exact employment chance in study task 2. None of the focus groups completed this step, but rather made educated guesses. This might be explained by the degree of focused attention the settings afforded participants. As single interviews incurred no distractions, participants could immerse themselves in the explanations.

Notably, this difference is not represented in self-reported understanding. Most participants in both settings reported unchanged understanding after the explanation phase (Table 3). Paradoxically, they verbally stated that it improved. E3

⁸In one study that made this comparison, participants met in a chatroom and decided *in collaboration with the AI* and not about it [22]

⁹Using multiple methods of data collection (here: self-reports, task performance, articulations/actions) to explore a phenomenon [18].

commented: “*I don’t think I understood it the way you can understand it yet, but it’s definitely better than before.*” And S3 explained that: “*I would still say my understanding is ‘good’, but this ‘good understanding’ is much more informed now than the first superficial one.*” This indicates that participants tended to judge their understanding relative to the information available, not necessarily in relation to their previous report. We describe this process as *calibrating understanding*. Previous research in cognitive science has documented similar effects [64], which were also reproduced in an XAI study on white-box explanations [21]. However, participants’ verbal reports, their feedback on the explanations (Section 4.1.4), and the calibration process itself indicate that the explanations improved understanding. Including additional measures, such as information gain, could further capture the calibration process, which is discussed in Section 5.2.

To compare individual and group interactions with the explanations, we lastly draw from the ‘six facets of understanding’ [129]. The framework describes that understanding is represented by the ability to ‘explain, interpret, apply, take perspective, empathize, and self-reflect’ with respect to a topic. The more facets are covered, the better the understanding. Seeing that individuals had a clear advantage in solving the study tasks suggests that the individual setting supported the ‘apply’ facet. In contrast, the group settings often led participants to ‘explain’ information to others and ‘interpret’ it (Section 4.1.1), and to ‘take perspective’ and ‘empathize’ through the exchange of views and experiences (Section 4.2). As explanations aim to improve understanding of a given AI system, combining both settings to cover more facets of understanding could thus be a fruitful approach. Further, explanations for individuals in particular can benefit from information covering facets usually dependent on social interaction. Our design aimed to implement this through explanations such as *What chance would the system calculate for me?* (interpret – making it personally relevant), and *How could the system make the work of counselors easier or more difficult?* (take perspective: provide multiple angles and arguments).

Based on these findings, we argue that both group and individual settings can contribute to participant understanding and should ideally be combined. In particular, focused attention can facilitate the application of information, while *shared understanding* and the exchange of opinions and arguments (Section 4.2) aid encouragement, reflection, and collective action. Considering this trade-off between settings can inform how explanations can be combined with social settings to cover as many understanding facets as possible.

4.1.4 Reflections on the explanation design: Modularity, levels of detail, and most important information. To examine how the explanation design was received, we asked participants for feedback on the explanations’ structure, content, style of expression, and information coverage. We report and summarize the participants’ criticisms as a basis to formulate design improvement suggestions in Section 5.

Strengths and weaknesses of the design. Positive comments described the explanations’ structure as “nicely presented” (A2, C2) and “good to get an overview” (C3, H4) while being “active and controllable” (S8). Critical comments described the information coverage as “too much” (D1, S4), and the structure as “confusing” (B1, D1) and “demanding” (D2). Participants saw the design’s strengths in its four-category structure, question-driven presentation, active selection, and information scope. However, the scope and depth of information also led to information overload and loss of overview. Further, the explanations’ many and complicated texts were described as “very difficult” (E2, F2). E2 compared the language to “*letters [...] from the court. I understand every single word, but I don’t understand the context.*” Previous work has found that textual explanations can effectively convey information but tend to raise aversion with users [106, 120]. However, Weitz et al. [126] paradoxically found that users preferred textual explanations. This points to the need for further research on textual formats in XAI, like the automated adaptation of text to different difficulties.

Table 3. Reported understanding and task performance. This table shows participants' two understanding (und.) self-reports and performance in the four study tasks (Section 3). Increases are colored green, decreases are colored red in reported understanding. The number of filled circles (●) indicates the number of correct study tasks. Participants in single interviews generally performed better in the study tasks. Group H showed a particularly high task performance as participants efficiently located the relevant information together (and made an educated guess for task 2). Most other groups had a lower task performance, but still the setting acted against participant discouragement and addressed specific facets of understanding (Section 4.1.5).

Focus groups					Single interviews				
ID	Und. I	Und. II	Change	Task Performance	ID	Und. I	Und. II	Change	Task Performance
A1	2	2	0	●○○○	F1	2	4	+2	●○○○
A2	1	2	+1	●○○○	F2	1	4	+3	●○○○
A3	4	4	0	●○○○	F3	5	2	-3	●○○○
A4	4	4	0	●○○○	F4	3	3	0	●○○○
B1	4	2	-2	●●○○	F5	4	4	0	●○○○
B2	4	2	-2	●○○○	G1	5	5	0	●○○○
B3	5	5	0	●●○○	G2	3	2	-1	●●○○
B4	2	3	+1	●●○○	G3	5	2	-3	●○○○
C1	4	4	0	●○○○	G4	5	4	-1	●○○○
C2	2	2	0	●●○○	H1	4	5	+1	●●●●
C3	1	3	+2	●○○○	H2	4	5	+1	●●●●
D1	4	3	-1	●●○○	H3	4	4	0	●●●●
D2	4	3	-1	●●●●	H4	4	5	+1	●●●●
D3	4	4	0	●●○○	H5	4	4	0	●●●●
E1	3	4	+1	●●○○					
E2	3	3	0	●○○○					
E3	4	4	0	●○○○					
Single interviews									
ID	Und. I	Und. II	Change	Task Performance	ID	Und. I	Und. II	Change	Task Performance
S1	5	5	0	●●●○	S7	4	4	0	●●●●
S2	4	2	-2	●●○○	S8	3	4	+1	●●●●
S3	4	4	0	●●●●	S9	5	5	0	●○○○
S4	4	4	0	●●○○	S10	4	3	-1	●●●○
S5	5	4	-1	●●●●	S11	4	4	0	●●●●
S6	1	4	+3	●●●○	S12	2	4	+2	●○○○

Most helpful and influential information. Participants in focus groups stated that all explanation categories helped their understanding and influenced their decision evenly (Figure 7), often mentioning that “*all of them [are relevant]... I don't think you can leave anything out, really*” (D3). In contrast, participants in single interviews found *data* much less helpful and less influential, stating, e.g., that they prioritized another category in the time available. Notably, participants emphasized that two categories were central: *system details* and *context*. *System details* were perceived as “tangible” (S6) and “concrete” (S8), and explanations about the features and weighting were perceived as especially important: “*That is the central point, the basis of the whole system.*” (G4) In turn, explanations from the category *context* were requested the most (Figure 6). Here, participants appreciated explanations that described decision subjects’ inability to contest decisions and the system’s political background. Drawing from the concept of ‘intelligibility types’ [76], we argue that *system details* provided descriptive information to the question “What did the system do?”, while *context* provided normative information to the question “Why did the system do [this]?”. Future research should investigate how both information types can be integrated into explanations for AI novices.

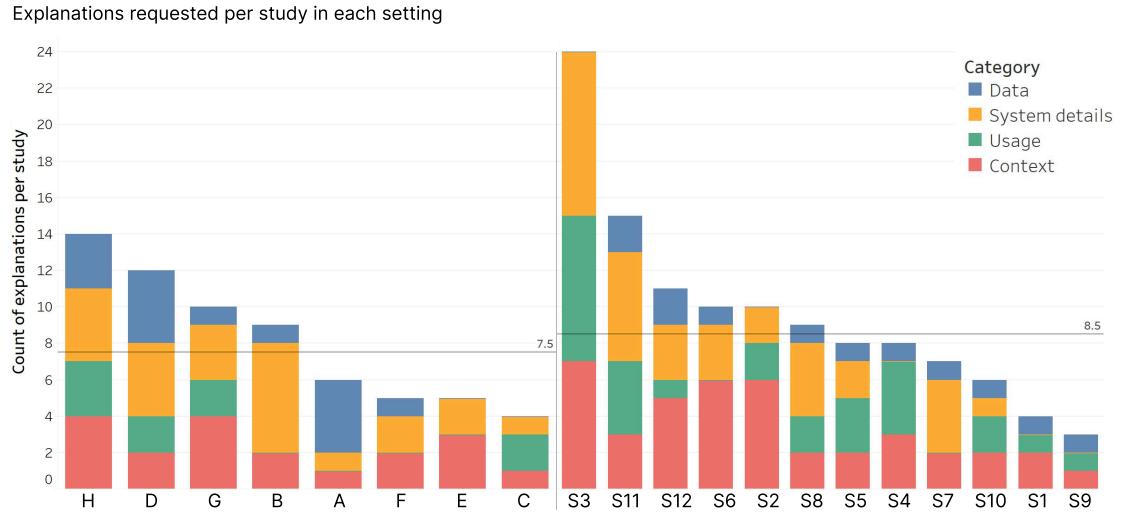


Fig. 6. Number of explanations requested. The left side shows explanations requested by focus groups, the right side by participants in single interviews. The horizontal lines indicate the median. While groups were able to process many explanations by splitting the reading, several single interview participants went through equal or even higher counts. Note that *context* explanations were the only category requested in every study.

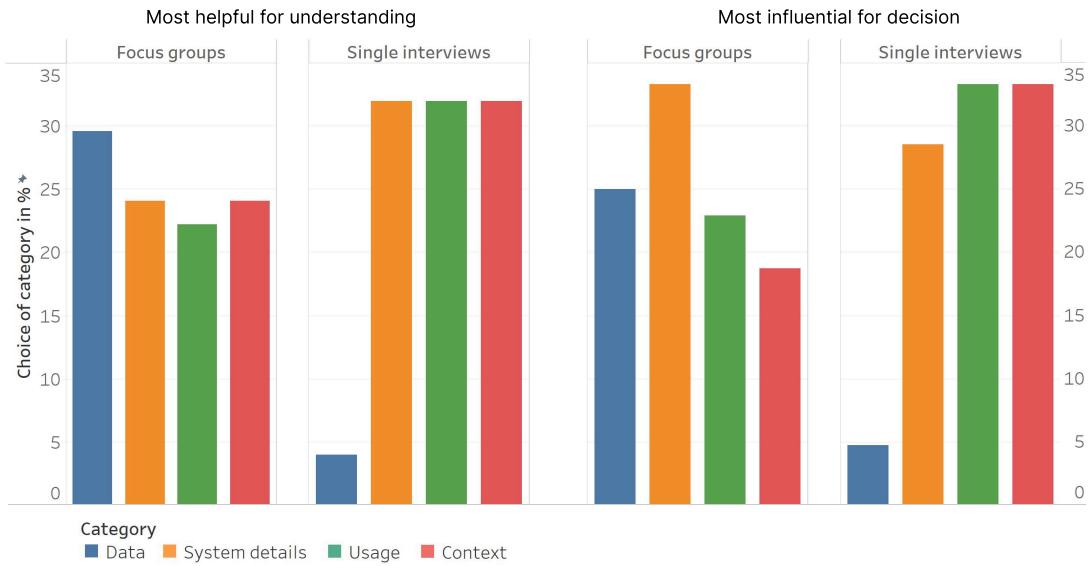


Fig. 7. Most helpful/explanatory/influential categories for understanding and most influential categories for participants' decisions. Participants could select any number of explanation categories for both questions, including none and all four. Focus group participants found all categories helpful for understanding, but reported *system details* to be more influential for their decisions. Participants in single interviews found *data* both less helpful and less influential and prioritized other categories in the given time. (Section 3.2)

4.1.5 Summary RQ1: *Group and individual settings address different understanding facets in explainable AI.* In RQ1-Explanations, we asked how a question-driven, modular explanation design supports AI novices' understanding in groups and individual settings. We found that the explanations support both settings but differ in how understanding develops. In groups, we found that explanations facilitated interactions that produced *shared understanding* and involved cognitive and social mechanisms of 'collaborative success' [97]. In groups with a trusting social dynamic, participants tackled explanations together, which acted against discouragement. When groups had a negative social dynamic, the explanation design could become overwhelming and understanding issues were left unchecked, leading to 'process loss' [65] and *abandoned understanding*. Participants in single interviews interacted with the explanations in a more focused and self-directed manner. This had advantages for task performance and explanation engagement, which aligns with research on tutoring methods [9]. However, the positive effects of aggregated knowledge [94] and peer discussion [113] in group settings should not be disregarded. Our findings showed that group settings can bridge understanding issues by boosting morale and letting participants share knowledge, interpretations, and experiences. We thus argue that individual and group settings support different *understanding facets* [129], meaning that they provide different grounds for understanding AI systems. While individual settings can make it easier to understand technical and numerical details that require much attention ('apply'), group settings can support understanding of the deployment context and consequences through the exchange of expertise and lived experiences ('interpret', 'take perspective', 'empathize'). Consequently, individual and group settings should be combined to leverage their different modes of interaction and understanding facets when explaining AI systems. In cases where both social settings cannot be provided, explanations of AI systems should aim to reinforce facets that are not covered in the corresponding setting.

4.2 RQ2-Deliberation: How do AI novices use explanations to form opinions and make decisions about AI systems in groups and individual settings?

Before and after the explanation phase, participants decided if the study's case should be deployed, and groups additionally made a collective decision (Table 4). We compared participants' decisions and decision confidence and how they deliberated deployment to examine the impact of explanations and the social setting. We first describe participants' confidence and decision changes in single interviews (4.2.1) and then present three cases of group discussion illustrating 'elements of deliberation' [116], including reasoned arguments (4.2.2), disagreement (4.2.3), and groupthink (4.2.4). Section 4.2.5 summarizes the findings.

4.2.1 Explanation phase led to increased decision confidence and decision swings. For most participants, deciding about the *AMS algorithm*'s deployment was a clear choice: 7 out of 8 groups and the majority of single participants voted "No" (Table 4). Many participants reported increased decision confidence after the explanation phase and stated that they felt better informed due to the explanations and, where applicable, the group discussion. Reasons for these increases included a better understanding of the system's "*fundamental idea*" (S4) and the "*exchange of different opinions and things that catch your eye*" (G2). Participant B3 emphasized that the explanations, although they only contained factual information, provided a stark contrast to *public narratives*:

Well, I changed my mind – you think you understand something when you see it in the media. You have a political opinion about it. But you don't know the background information. And when you get to the background information, you can have a completely different opinion. (B3)

This contrast highlights how explanations of an AI system can impact decision-making by correcting lay understandings [34] and is in line with previous work that advocates for making algorithmic complexity visible to users

to tackle lay understandings [39]. It further directly connects the explanations and changes in participants' decision confidence. Few participants reported decreased decision confidence, and only S3 and S11 reported a stronger decrease of -2 (Table 4). S3 explained that the system might have benefits, but too much depended on the *conditions for deployment*. In particular, it should be deployed “*responsibly, with a pilot project, in a selected group, for three months*”, and not haphazardly, where “*you sit around for a day or eight hours and then training is finished*” (S3). S11, who together with S3 requested the most explanations out of all participants, paradoxically stated that their confidence decreased because “*I'm still missing so much information. Especially [...] how tedious it is for the counselors if they have to disagree with the system.*” In consequence, S11 changed their decision from ‘Yes’ to ‘No’. Notably, this *fear of an algorithmic imprint* was a prevalent theme throughout all studies and was often connected to past experiences with digitization projects, and the corresponding *institutional deficiencies*. S5, who also changed from ‘Yes’ to ‘No’, similarly stated that the explanations helped them to *scrutinize the system*: “*You don't have to introduce anything that's extra bad*”. While the explanations thus made the *decision more uncertain* for some, they undoubtedly encouraged critical reflection about the use case and triggered decision changes. Despite only having “*their own mind*” (S8), single participants could make use of the explanations to *weigh pros and cons* and *adjust their mental model*. This form of “*internal deliberation*” is supported by exposure to different views, as provided through the argumentative explanations in *usage* and *context*. This suggests that explanations can substitute at least small parts of public deliberation, which is thought to be the more salient driver for “*reasoning towards good outcomes*” [85]. To contrast these findings with focus groups, we illustrate this public deliberation with three conversation excerpts that showcase elements of deliberation in focus groups.

4.2.2 Case 1 - reasoned arguments: Group B discusses whether to deploy the system. Group B was composed of staff members and volunteers of a civil society organization. Three participants in this group changed their votes from ‘Yes’ in the first report to ‘No’ in the second report. We found this change to be driven by three main deliberation elements [116]: ‘sourcing’ information, ‘reasoned arguments’ (opinion claims grounded in the information), and ‘engagement’ with the topic and between participants. In the excerpt, B2 and B3 *weigh pros and cons* of deployment. B3 grounds their arguments in explanations about the system’s features and weightings (*system details*), changing the discussion’s course:

B2 *I'm skeptical, but I'm still in favor of introducing it. Because it could be an aid and a relief for the staff working there.*

B3 *I was originally in favor for these reasons, but since I've seen these parameters, I would be very much against it. Because I think there's a lot of ideology in it. I think it's no longer acceptable that men are favored over women and that duty of care only applies to women. This comes from a time that should be long gone.*

B2 *Those are strong arguments.*

B3 *The things that come out are so absurd as well. For example, Harald's apprenticeship was rated positively, but he can't even use the apprenticeship for retraining. [...] As much as I like the idea, I don't like the parameters.*

B1 *Did you vote yes first?*

B3 *I ticked yes at first, but I was really shocked at what was in there [in the system]. [...]*

B1 *What I'm wondering is, what would be the real benefit of introducing the system? [...]*

B4 *It's a grid, a structure for the people who work at the agency, so that they can quickly find a box.*

The excerpt highlights how the explanations led B3 to *change their deployment decision* and served as *discussion triggers*. In the resulting discussion, participants state both arguments (*discrimination, what's the benefit?*) and opinions (*disagreement with policy choices, AI can assist in decisions*). Note that there is a difference between arguments (expression

Table 4. Individual and collective decisions about deploying the study's use case. Participants were asked for their decision about the deployment of the employment prediction system before and after receiving explanations and discussing them (Section 3.3). Focus groups further made a collective decision about the deployment. Instances in which participants changed their votes between the first and second decision are colored red. In most focus groups, decision confidence increased after the explanation phase, with the exceptions of Groups E and F, in which participants had trouble engaging with the explanations and did not collaborate with each other (Section 4.1.2). In single interviews, participants reported similar increases, except for S3 and S11, who explained their confidence decreases with strong adjustments to their mental models of the use case (Section 4.2.1).

Focus groups					Single interviews				
ID	Decision I	Group decision	Decision II	Decision Conf.	ID	Decision I	Group decision	Decision II	Decision Conf.
A1	No		Yes	0	F1	Yes		Yes	0
A2	No	Yes	Yes	+1	F2	Yes		Yes	-1
A3	Yes		Yes	+1	F3	Yes	No	No	+1
A4	No		Yes	0	F4	No		No	-1
B1	No		No	+3	F5	Yes		Yes	-1
B2	Yes	No	No	0	G1	Yes		Yes	+1
B3	Yes	No	No	+2	G2	No	No	No	0
B4	Yes		No	+3	G3	Yes	No	No	0
C1	No		No	-1	G4	No		No	+1
C2	No	No	No	0	H1	No		No	+1
C3	No		Yes	+2	H2	No		No	+1
D1	No		No	+1	H3	No	No	No	+3
D2	No	No	No	0	H4	No		No	0
D3	No		Yes	0	H5	No		No	+1
E1	No		No	0					
E2	No	No	No	0					
E3	No		Yes	-1					

ID	Decision I	-	Decision II	Decision Conf.	ID	Decision I	-	Decision II	Decision Conf.
S1	Yes	-	Yes	0	S7	No	-	No	+3
S2	No	-	No	-1	S8	No	-	No	0
S3	No	-	No	-2	S9	No	-	No	+1
S4	No	-	No	+1	S10	Yes	-	Yes	0
S5	Yes	-	No	+2	S11	Yes	-	No	-2
S6	No	-	No	+2	S12	No	-	No	+1

of reasoning processes that can be defended against critique) and opinions (expression of the speaker's belief) [85, 116]. While conceiving arguments to persuade interlocutors can result in confirmation bias (interpreting evidence such that it confirms existing beliefs) [86], the fact that B3 changed their attitude, in fact, indicates that the explanations acted against this bias. We argue that the excerpt thus shows a positive synergy in that the explanations provided grounds for 'arguments', which entered the discussion and provoked 'collective reasoning' and three decision swings. However, considering the large argumentative influence of B3, it should also be considered how the discussion would unfold if B3 had advocated for deployment. A case with comparable dynamics is described in 4.2.4.

4.2.3 Case 2 - disagreement: Group D debates normative positions regarding the algorithmic representation of people. Group D was composed of participants who had been job-seeking in the past. When the group discussed the AI system's deployment, the conversation shifted to how features that represent job-seekers' profiles should be selected and weighed. This produced disagreement, an "important marker for deliberation" [116] that displays heterogeneity of viewpoints, acts against polarization, and involves close examination of others' reasoning. In the excerpt, D3 argues for the system's deployment, while D1 argues against, and D2 acts as a mediator:

D3 I believe that the system can form the initial basis, based on the unalterable facts, which are of course weighted, but then it has to be enriched by a human being. [...]

D1 But I don't believe that there are unalterable facts – well, not in this area. It's all a question of representation and the lens through which you see the world.

D3 When the job-seeker says, 'I only have four years of elementary school', then that's four years of elementary school... [...]

D2 That doesn't mean that he can't still be a very educated person.

D3 But that is hard to sell to an employer, right? [...]

D2 I'm skeptical about the data. You [D3] said it's the 'basis', I think there are cracks in this basis. And I'm afraid [...] that something will be pre-determined...

D3 But the human decision is always subjective.

D2 That has to be weighed up. On the one hand, you have the arbitrariness of the individual employee, yes, and on the other hand, you have an incomplete picture of a person.

D1 Or a false image.

D2 An incomplete one, I would say.

The group here discusses diverging views and expresses opinions. While these opinions are meant to persuade and defend, they are grounded in *lived experiences* rather than in the explanations. The central conflict develops between D1's belief that the system *misrepresents reality* and D3's viewpoint that it can *increase objectivity* and *assist in decisions*. The discussion here did not lead to a consensus on the deployment decision in the given time, but resulted in a majority vote for 'No'. We argue that it still illustrates an important process in the deliberation on public AI systems: Participants again 'sourced' information that was turned into arguments, but the debate led to a more fundamental topic that surfaced discrepancies which would impede finding a collective decision. The fact that participants then engaged in 'disagreement' is a sign of productive deliberation, as it shows that there were diverse viewpoints, that no polarization or 'groupthink' [55] occurred, and that the proposal was closely examined based on the information given [116]. In a real setting, this form of debate could serve as a fruitful basis to investigate whether the system is in the 'public interest' [132] and to host 'early-stage deliberations' [61] on the system during development. The merit of this debate was further later acknowledged by D1, who found the explanations confusing but stated that these exchanges were the study's "centerpiece" and most intriguing part. We argue that the interplay between explanations and group discussion here supported a (simulated) evidence-informed policy-making process [82].

4.2.4 Case 3 - groupthink? Group A follows a minority position and votes for system deployment. Group A was composed of volunteers from a civil society organization. Three participants in this group changed their decisions, shifting from 'No' to 'Yes' after the explanation phase. We explain these changes with three aspects: First, Group A focused on the explanation category *data* and did not interact much with other categories (Figure 6). This meant less attention was paid, for example, to the system's feature selection and weightings that were decisive in Cases 1 and 2. Second, participants of Group A stated that they were not directly affected by the system, as they were retired, implying low 'engagement': "It doesn't affect me anymore and I think to myself, yeah..." (A1). Third, participants prioritized group concurrence above a "careful, critical scrutiny" [55]. The following excerpt illustrates the tipping point for the collective decision:

A3 I still think the system is better, even if there are still mistakes in it, than sitting opposite someone [a counselor] who doesn't like you... [...]

A2 So rather 'no'?

- A1** Yes, as A3 says, it's... I don't know.
- A3** Yes and no... [...]
- A2** I mean, it can't be avoided, it will happen. I'm convinced of that, whether we like it or not, it's done.
- A1** It won't affect us anymore, at least not in the employment office. [...] I agree with the majority.
- A2** But that's difficult now.
- A3** I'll stick with 'yes'. My daughter would say I shouldn't think so negatively, especially with AI. [...]
- A2** I say 'yes' too. [...] You, A1 and A4, can tip the scales.
- A1** I say 'yes' now too, but not because I've changed my mind, but because I want an overall solution.
- A4** I say 'yes' but I'm leaning towards 'no'.

Despite articulated reservations, all participants ultimately decided to vote for deployment. We compared this excerpt with characteristics of 'groupthink', a "mode of thinking" in which people value concurrence higher than consideration of alternative courses of action [55]. This mode produces defective decision-making processes due to three key aspects: strong social identification with the group, salient norms, and a perceived low self-efficacy to make the decision [5]. The excerpt clearly demonstrates two of these aspects: A1 changes their decision due to a desire for group harmony, and A4 follows suit (group identification). Further, both A1 and A3 express their uncertainty and sway between options (low self-efficacy). While A2's statement that *AI is inevitable* is an opinion rather than an argument (neither 'sourced' nor the product of evident reasoning), it triggers the group to make a quick decision that disregards any remaining 'disagreement'. Although the process can be connected to aspects of 'groupthink' [55], such as rationalizations of flawed logic and self-censorship, these aspects are not nearly as pronounced as in the literature [5, 55, 56]. For example, the group did not share an illusion of unanimity, and the uncertainty among participants suggests no guiding salient norms. Still, as participants avoided 'disagreement' and instead *followed decisions of other*, the excerpt presents a suboptimal deliberation process [5]. In part, this can be attributed to the explanation's failure to make all fundamental information easily available and to not encourage analytical thinking over intuitive, heuristical thinking [15]. In addition, the group might have missed a role that explicitly takes the opposing viewpoint to fuel discussion, which was identified to benefit deliberation in previous XAI research [23]. The implications of these findings are discussed in Section 5.

4.2.5 Summary RQ2. The findings in this Section demonstrate how explanations supported deliberation in focus groups and single interviews. Many participants reported improved decision confidence and changed their deployment decisions based on the explanations, often due to a disillusionment regarding the *AMS algorithm's* assumed merits. These changes occurred in both settings, suggesting that the explanations supported public and internal deliberation [85]. In group settings, participants used explanations when discussing deployment, as illustrated in Case 1. Case 2 further highlights that explanations surfaced discrepancies in personal beliefs and produced productive conflict. In contrast, Case 3 shows a deployment decision based more on concurrence-seeking than on 'collaborative reasoning' [90]. However, we hesitate to label the exchange as 'groupthink', as it does not align with all factors that characterize the phenomenon [5]. Based on these findings, we argue that explanations can support people in considering if AI systems are in the public interest and to discuss "*whether and under what conditions* to move forward with developing or deploying" them [132]. To achieve this, both the explanations and the group setting need to i) be designed so that they allow for the easy sourcing of information for arguments, ii) make all relevant information available as soon as possible, and iii) include mechanisms that encourage participants to examine both the proposal and their positions closely. Matching explanations and social setting to support 'elements of deliberation' [116] thus presents promising starting points for future research on how explainable AI can promote public deliberation on AI.

5 DISCUSSION

In this section, we discuss how our findings answer our two main research questions: Whether a question-driven, modular explanation design supports AI novices' understanding in groups and individual settings (RQ1) and how AI novices used these explanations to deliberate about AI systems (RQ2). We describe the advantages of both social settings for explainable AI, outline which real-world use cases would benefit from our explanation design, discuss whether the explanations improved understanding, and provide suggestions for their design improvements. We summarize the implications of our findings in Figure 8.

5.1 Do AI novices learn and deliberate about AI better together or individually?

In Section 4.1, we described that explanations produced *shared understanding* in groups, involving both cognitive and social mechanisms of “collaborative success” [97]. Section 4.2 further showed that explanations improved participants’ decision confidence and provided grounds for different elements of deliberation [116], such as reasoned arguments and disagreement. In the best cases, focus groups in our study had a familiar [57] and solution-oriented [95] atmosphere that facilitated sharing and discussing information. In these settings, the modular explanation structure showed its strengths by allowing for the distribution of tasks among group members, providing high levels of detail and breadth if needed, and offering different viewpoints that could be used as argumentative and conversational starting points. In this sense, the explanations fulfilled their aim of supporting learning and deliberation about a public AI system [61]. The interaction between group members is the differentiating factor compared to “one-to-one” [91] explanation settings. In our study, single interviews allowed for more focused engagement with explanations and a form of “internal deliberation” [85] but lacked the exchange of knowledge and perspectives with others that is deemed central for deliberation about public AI [132]. Regarding learning and deliberation, XAI would thus benefit from researching how group settings can be used to leverage collective reasoning [90], wisdom of the crowds [94], and performance increases through peer discussion [113]. However, the benefits of group settings have several preconditions, such as the containment of cognitive biases (groupthink [55], equality bias [93]) and, crucially, a trusting social dynamic [22].

The importance of the social dynamic became evident in groups where members were not familiar and had trouble engaging with the explanations. In groups G and H, for example, the social dynamic bridged understanding issues of individual participants and acted against discouragement. In groups E and F, in contrast, these understanding issues eventually led participants to abandon understanding, as the social atmosphere did not support them in overcoming them. Here, a lack of trust or simply unfamiliarity between participants likely amplified effects such as social loafing and the fear of being evaluated [97]. This underscores the importance of creating trust between group members in collaborative XAI settings [57]. Intuitive measures could be the introduction of a simple task that the group solves collaboratively before engaging with explanations, such as the Wason card selection task [124]. Another measure could be the introduction of roles (e.g., proponents and opposition), as has been done with the “devil’s advocate” in previous work [23], to facilitate discussion and close examination of the proposal. Future work should examine how such measures can be incorporated into explanation design to support interaction in groups of comparable compositions. Lastly, regarding cognitive biases, we observed an effect resembling some aspects of “groupthink” [55] when participants in Group A changed their vote to “Yes” to reach a group decision. We argue that this effect originated in the lack of detailed interaction with explanations and, possibly, a perceived low degree of personal affection by the system’s deployment. However, this is contrasted by participants in Group D, who debated at length about the system’s deployment without reaching a consensus, despite not being directly affected. Potential measures to avoid groupthink

in discussion could thus be to encourage debate, which again could be the introduction of roles to improve the “dialectic argumentation” [86], and to explain the system in a way that makes it more personally relevant to participants [129], e.g., by emphasizing connections to their own experiences.

5.2 Did the explanations improve participants’ understanding?

In Section 4.1.5, we described that the explanations helped participants develop different ‘facets of understanding’ [129]. In groups, participants were encouraged to ‘explain’ information to each other and ‘empathize’ with others’ experiences, while individuals could better ‘apply’ information in the study tasks. We further described that groups’ interactions with explanations realized mechanisms of “collaborative success” [97]. We thus conclude that the explanations had a positive effect on understanding. However, a more complete answer requires that we consider the difference between measurement methods and true cognitive states. In Section 4.1, we described that a majority of participants reported unchanged understanding after the explanation phase (Table 3) but, paradoxically, described verbally that their understanding improved; two seemingly incongruent pieces of evidence. We explain these contradictory findings with a process we call *calibrating understanding*. The term describes that participants tend to report understanding not in absolute terms, or even in relation to past understanding, but in relation to the currently available information. Participants explicitly stated that they calibrated their interpretation of ‘good understanding’ according to their knowledge of the information basis, which differed before and after the explanation phase. The calibration process can be traced by using concepts from the cognitive sciences: Participants i) reported their initial understanding after reading the use case introduction, they then ii) saw the explanations and realized that they had understanding gaps [104], which they iii) proceeded to locate and close [64], however, they iv) also realized that they could not look at every available explanation and would develop at most a “partial understanding” [62], which they v) rated accordingly in the second self-reports.

Previous studies in XAI documented similar effects caused by white-box explanations [21], which, due to their high information density, led to increased ‘objective’ understanding but decreased self-reported understanding. Importantly, [100] found similar discrepancies when participants who received no explanation gave higher understanding scores than participants who received faithful explanations; a discrepancy our findings might explain. In the same study, similar discrepancies also occurred between trust self-reports and observations of behavior [100]. In line with these findings, we argue that the calibration effect should be accounted for when measuring understanding, for example, by eliciting an additional metric that captures the perceived scope of available information. A potential reporting question could be “How much information do you feel you currently have about the presented AI system?”, combined with a 5-point scale ranging from “very little” to “very much”. Self-reported understanding could then be compared with self-reported information scope and verbal responses to acquire a more complete picture. Recent work in XAI has further proposed understanding measurement based on participants’ abilities [115]. This approach appears promising, as the ability to calculate study task 2 was a relevant metric in our study. We thus see eliciting understanding via multiple measures and exploring how these measures can be combined in individual and group settings as a direction for future research.

5.3 Which real-world settings would benefit from explainable AI in groups?

In Section 2.3.2, we described several settings where citizens gather to discuss and form opinions on matters of public interest. These included referendums, forums, and community-based spaces. This paper investigates settings suitable for deliberating the deployment of public AI systems, an issue that we frame as a matter of public interest due to the scope and severity of its potential consequences. Having established that using explanations in group settings benefits

participants' understanding, decision confidence, and decision-making processes, it is worthwhile to consider how this setting could be employed in real-world contexts. One answer can be given based on participants' feedback, who stated that training in their job agency should employ a similar format to educate about AI. Notably, this feedback was given by domain experts (Groups C, G) and decision subjects (Group E), suggesting that the setting would suit both stakeholder groups for an educational intervention. Similarly, P3 explained that the explanation approach could be helpful if a similar system were used in their care facility by embedding it in the team's regular meetings, in which difficult cases are discussed and joint decisions are made. These insights are in line with previous work on XAI in public institutions. Notably, Lee et al. [73] and Weitz et al. [126] conducted participatory workshops to design explanations with end-users in the public sector, finding that co-designing explainable AI helps in considering the needs of both clients and end-users. We envision that collaborative settings and 'mini-publics' [44] could be useful in many contexts that aim to strengthen participatory democracy with respect to AI. Potential areas of application could be professional consultation workshops for citizens affected by algorithmic decisions, comparable to legal clinics [98], community-based education and training interventions, such as "contestation cafés" [27], or union forums that inform and organize employees' voices about the use of AI in their institution [60]. On a different note, Crivellaro et al. [31] found that participatory formats that aim to connect communities to public institutions can suffer from a lack of crucial information (e.g., budgets), which could be alleviated by an information structure such as the presented explanation design. In short, we propose that explainable AI in collective settings could be a valuable engagement format for contexts in which public AI could impact people's lives. Future work could explore how collective XAI settings could be implemented in these contexts as part of responsible AI initiatives and in connection to both institutionalized [30] and user-based [109] auditing practices.

5.4 What's missing from the explanation design and how could it be improved?

In Sections 4.1.4, we described that participants appreciated the explanations' comprehensive and flexible information selection and self-directed and active exploration. However, they also noted that the explanations have a high access threshold and require adjustment to the modular structure, making oversight difficult. A digital version of the explanation design could improve the overview through summaries and navigation while allowing for simple language options and cross-references. As the simple awareness of the scope of information also seemed to overwhelm participants, approaches to condense the scope and selection would be beneficial. An example could be a recommendation system that would suggest to participants explanations from different categories and levels of detail based on their stated interests and technical knowledge. However, we also emphasize that identifying explanation subtopics and splitting them up into levels of detail is challenging. Our explanations' structure provides different levels of soundness and completeness [69], but deciding how the available information is allocated into this structure requires a subjective choice. In our study, detail levels 2 and 3 were supposed to convey more detailed but also more difficult explanations than the base levels. Still, some participants stated that the most critical information for them resided in level 3. We see avenues for future work in selecting and hierarchically structuring information to be included in explanations for AI systems, such that they allow for exploration while not obscuring essential information. This essential information should balance textual and visual design to ensure its uptake does not rely only on textual understanding.

Due to the modular structure and the complexity of some of the information, several participants used intuition in answering the study tasks instead of truly searching for the explanations for solutions. Framing this in the dual-process theory of cognition, we observed that participants in these moments used System 1 (intuitive heuristics) rather than System 2 thinking (analytical reasoning) [15, 58]. Lambe et al. [70] listed strategies to counteract this tendency and encourage analytical thinking in the medical domain, including checklists, cognitive forcing mechanisms (consideration

of alternative diagnoses, reconsideration of diagnoses), guided reflection, and use of particular reasoning approaches. Buçinca et al. [15] further tested cognitive forcing mechanisms in an AI-assisted decision-making scenario, in which they used three interventions: showing participants an AI decision only on demand, showing the AI decision only after the participants made their own predictions, and letting them wait before showing them the AI decision. These mechanisms improved the performance of the human-AI teams but led users to dislike the interface's usability, presenting a trade-off. Nonetheless, we concur that these strategies should be considered in future explanation design to "ensure that people will exert effort to attend to those explanations" [15].

5.5 Summary of explanation design suggestions

We summarize the implications of our findings in the form of suggestions for the design of explanations suited to AI novices in individual and group settings in Figure 8.

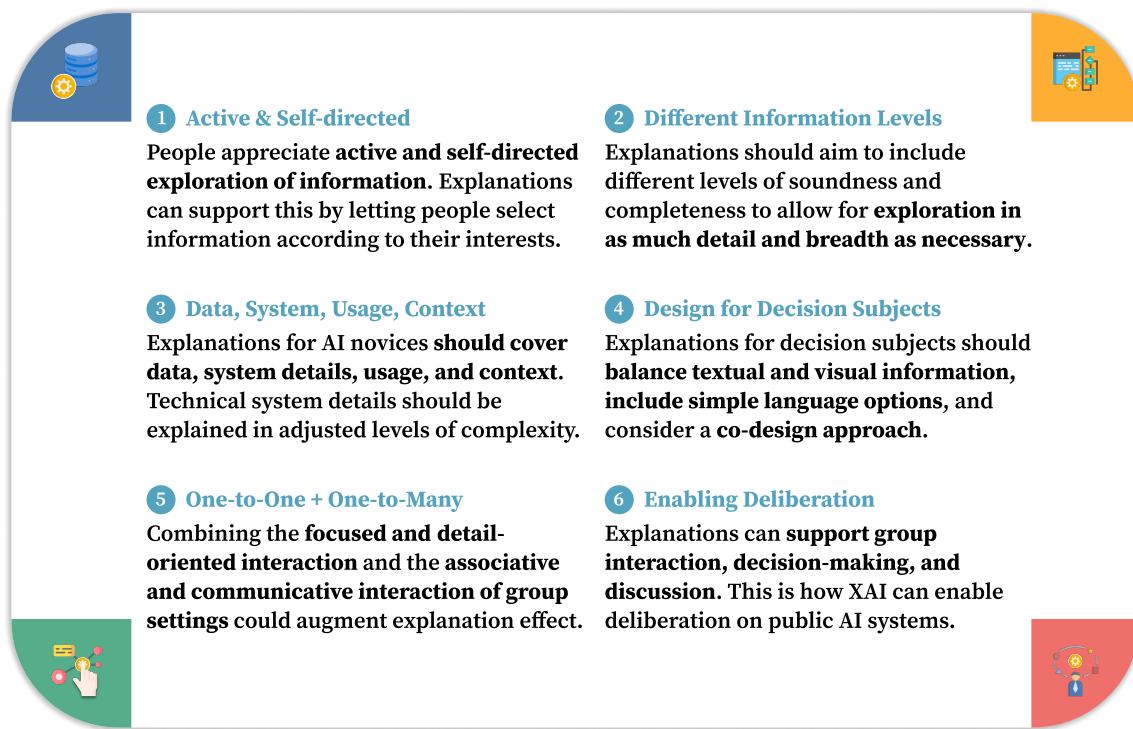


Fig. 8. Summary of implications regarding the design of explanations for individual and collaborative settings based on our findings.

6 LIMITATIONS

Like any research, this study had limitations. Due to the limited sample size, we did not analyze the impact of sex and/or gender on our results, limiting the results' generalizability regarding these aspects but not their overall validity. Further, our participants were recruited from organizations and networks in the same geographical region, perhaps resulting in regional or cultural biases. The presented use case is further embedded in a specific sociotechnical context [37]

that might affect participants' understanding and perceptions (e.g., perceptions might differ between employability prediction and credit approval), and thus, a change in the domain might also change the explanations' effect. However, this does not limit the transferability of the explanation design, which can be seen as a template that can be adjusted to other use cases. We further note that our participant sample is biased toward university education in the single interviews, which we addressed by comparing these participants mostly with university-educated participants in the focus groups.

We are further aware that the cooperation with civil society organizations in the recruitment of participants could have led to selection biases, especially in the form of convenience sampling (over-representation of readily available participants), self-selection bias (over-representation of strongly motivated participants), and interviewer bias (over-representation of agreeable or compatible participants) [25]. We aimed to counteract these biases by defining research goals and methods clearly before recruiting participants, by using multiple recruitment sources and methods, ensuring that group composition was diversified, and by reflecting on possible sampling influences in the analysis of results. Due to the qualitative approach, our recruitment strategy further did not aim for statistical generalizability but instead intended to cover a variety of "theoretically relevant cases" [25] and "careful contextualization" [26] to examine our research questions.

7 CONCLUSION

This paper tested a question-driven, modular explanation design with AI novices in groups and individual settings. We conducted an interview study involving 8 focus groups and 12 single interviews. We analyzed them to examine the effect of explanations on understanding, decisions, and decision confidence, participants' perceptions of key information, and the interaction processes in both settings. We found that explanations supported participants' understanding and decision-making differently, encouraging focused interaction in individual settings and shared understanding in group settings. Even though individuals could not exchange with others, the explanations still led to increased decision confidence and changes by supporting internal deliberation. In groups, the explanation design afforded a set of interactions that allowed participants to support each other's understanding, and further provided grounds for exchanging arguments about key aspects regarding the system's deployment. For groups that experienced collaborative failure, we suggest the modification of the explanation's design to highlight essential information and measures to create a more productive social dynamic. With this work, we aim to showcase the potential of combining explanations with group settings to enable AI novices to understand and deliberate about public 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] Kars Alfrink, Janus 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. <https://doi.org/10.1145/3544548.3580984>
- [2] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. *Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*. Technical Report. Österreichische Akademie der Wissenschaften. epub.oeaw.ac.at
- [3] 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. <https://doi.org/10.1177/1461444816676645>

- [4] Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. 2018. Improving refugee integration through data-driven algorithmic assignment. *Science* 359, 6373 (Jan. 2018), 325–329. <https://doi.org/10.1126/science.aoa4408>
- [5] Robert S. Baron. 2005. So Right It's Wrong: Groupthink and the Ubiquitous Nature of Polarized Group Decision Making. In *Advances in Experimental Social Psychology*. Vol. 37. Elsevier, 219–253. [https://doi.org/10.1016/S0065-2601\(05\)37004-3](https://doi.org/10.1016/S0065-2601(05)37004-3)
- [6] Christoph Baumberger, Claus Beisbart, and Georg Brun. 2017. What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science. In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, Stephen Grimm, Christoph Baumberger, and Sabine Ammon (Eds.). Routledge, 1–34.
- [7] Astrid Bertrand, James R. Eagan, and Winston Maxwell. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 943–958. <https://doi.org/10.1145/3593013.3594053>
- [8] Or Biran and Kathleen McKeown. 2017. Human-Centric Justification of Machine Learning Predictions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 1461–1467. <https://doi.org/10.24963/ijcai.2017/202>
- [9] Benjamin Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. (1984).
- [10] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (<conf-loc>, <city>Helsinki</city>, <country>Finland</country>, </conf-loc>)* (IUI '22). Association for Computing Machinery, New York, NY, USA, 807–819. <https://doi.org/10.1145/3490099.3511139>
- [11] 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>
- [12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [13] 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. <https://doi.org/10.1145/3290605.3300271>
- [14] Bundesagentur für Arbeit. 2021. Bearbeiten von Bewerberdaten durch Träger. arbeitsagentur.de/datei/dok_ba013193.pdf.
- [15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. <https://doi.org/10.1145/3449287>
- [16] Ruth M.J. Byrne. 2023. Good Explanations in Explainable Artificial Intelligence (XAI): Evidence from Human Explanatory Reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. Macau, SAR China, 6536–6544. <https://doi.org/10.24963/ijcai.2023/733>
- [17] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–23. <https://doi.org/10.1145/3544548.3580959>
- [18] Nancy Carter, Denise Bryant-Lukosius, Alba DiCenso, Jennifer Blythe, and Alan J. Neville. 2014. The Use of Triangulation in Qualitative Research. *Oncology Nursing Forum* 41, 5 (Sept. 2014), 545–547. <https://doi.org/10.1188/14.ONF.545-547>
- [19] Mohamed Amine Chatti, Mouadha Guesmi, Laura Vorergerd, Thao Ngo, Shoei Joarder, Qurat Ul Ain, and Arham Muslim. 2022. Is More Always Better? The Effects of Personal Characteristics and Level of Detail on the Perception of Explanations in a Recommender System. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Barcelona Spain, 254–264. <https://doi.org/10.1145/3503252.3531304>
- [20] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. 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, 1. <https://doi.org/10.1145/3593013.3593970>
- [21] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland UK, 1–12. <https://doi.org/10.1145/3290605.3300789>
- [22] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. <https://doi.org/10.1145/3544548.3581015>
- [23] 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. <https://doi.org/10.1145/3640543.3645199>
- [24] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. <https://doi.org/10.1089/big.2016.0047>
- [25] David Collier. 1995. Translating quantitative methods for qualitative researchers: The case of selection bias. *American Political Science Review* 89, 2 (1995), 461–466.
- [26] David Collier and James Mahoney. 1996. Insights and Pitfalls: Selection Bias in Qualitative Research. *World Politics* 49, 1 (Oct. 1996), 56–91. <https://doi.org/10.1353/wp.1996.0023>

- [27] Robert Patrick Collins, Johan Redström, and Marco Rozendaal. 2024. The right to contestation: Towards repairing our interactions with algorithmic decision systems. (2024). <https://doi.org/10.57698/V18I1.06> Publisher: International Journal of Design.
- [28] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (Sept. 2021), 103503. <https://doi.org/10.1016/j.artint.2021.103503>
- [29] Gregorio Convertino, Dorrit Billman, Peter Pirolli, J. P. Massar, and Jeff Shrager. 2008. The CACHE Study: Group Effects in Computer-supported Collaborative Analysis. *Computer Supported Cooperative Work (CSCW)* 17, 4 (Aug. 2008), 353–393. <https://doi.org/10.1007/s10606-008-9080-9>
- [30] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1571–1583. <https://doi.org/10.1145/3531146.3533213>
- [31] Clara Crivellaro, Rob Anderson, Daniel Lambton-Howard, Tom Nappey, Patrick Olivier, Vasilis VLachokyriakos, Alexander Wilson, and Pete Wright. 2019. Infrastructuring Public Service Transformation: Creating Collaborative Spaces between Communities and Institutions through HCI Research. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 15 (May 2019), 29 pages. <https://doi.org/10.1145/3310284>
- [32] 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. <https://doi.org/10.1007/s00146-020-00960-w>
- [33] 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. <https://doi.org/10.1017/S0047279420000203>
- [34] Michael A. DeVito, Jeffrey T. Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The Algorithm and the User: How Can HCI Use Lay Understandings of Algorithmic Systems?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI EA ’18*). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3170427.3186320>
- [35] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. ACM, Virtual Event USA, 1591–1602. <https://doi.org/10.1145/3461778.3462131>
- [36] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. <https://doi.org/10.1145/3613904.3642474> arXiv:2107.13509 [cs].
- [37] 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. <https://doi.org/10.1145/3579467>
- [38] 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. <https://doi.org/10.1145/3544549.3573832>
- [39] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (*CHI ’16*). Association for Computing Machinery, New York, NY, USA, 2371–2382. <https://doi.org/10.1145/2858036.2858494>
- [40] European Commission. 2024. Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations. .
- [41] Casey Fiesler, Jed R. Brubaker, Andrea Forte, Shion Guha, Nora McDonald, and Michael Muller. 2019. Qualitative Methods for CSCW: Challenges and Opportunities. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Austin TX USA, 455–460. <https://doi.org/10.1145/3311957.3359428>
- [42] Asbjørn Ammitzbøll Flügge. 2021. Perspectives from Practice: Algorithmic Decision-Making in Public Employment Services. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Virtual Event USA, 253–255. <https://doi.org/10.1145/3462204.3481787>
- [43] Timo Freiesleben and Gunnar König. 2023. Dear XAI Community, We Need to Talk!. In *Explainable Artificial Intelligence*, Luca Longo (Ed.). Springer Nature Switzerland, Cham, 48–65. https://doi.org/10.1007/978-3-031-44064-9_3
- [44] Archon Fung. 2003. Survey Article: Recipes for Public Spheres: Eight Institutional Design Choices and Their Consequences. *Journal of Political Philosophy* 11, 3 (2003), 338–367. <https://doi.org/10.1111/1467-9760.00181> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9760.00181>
- [45] Jutta Gamper, Günter Kernbeiß, and Michael Wagner-Pinter. 2020. Das Assistenzsystem AMAS: Zweck, Grundlagen, Anwendung. https://www.ams-forschungsnetzwerk.at/downloadpub/2020_Assistenzsystem_AMAS-dokumentation.pdf
- [46] Stephen R. Grimm. 2019. Varieties of Understanding. In *Varieties of Understanding*. Oxford University Press, 1–14. <https://doi.org/10.1093/oso/9780190860974.003.0001>
- [47] Mouadh Guesmi, Mohamed Amine Chatti, Shoeb Joarder, Qurat Ul Ain, Rawaa Altrash, Clara Siepmann, and Tannaz Vahidi. 2023. Interactive Explanation with Varying Level of Details in an Explainable Scientific Literature Recommender System. *International Journal of Human-Computer Interaction* (Oct. 2023), 1–22. <https://doi.org/10.1080/10447318.2023.2262797>
- [48] Jurgen Habermas. 1991. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press.
- [49] Maarten A. Hager and H. Wagenaar (Eds.). 2003. *Deliberative policy analysis: understanding governance in the network society*. Cambridge University Press, Cambridge, UK ; New York, USA.
- [50] Clément Henin and Daniel Le Métayer. 2022. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* 37, 4 (Dec. 2022), 1397–1410. <https://doi.org/10.1007/s00146-021-01251-8>

- [51] 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. <https://doi.org/10.1177/1049732316665344>
- [52] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. 2023. Explainable AI: roles and stakeholders, desirments and challenges. *Frontiers in Computer Science* 5 (Aug. 2023), 1117848. <https://doi.org/10.3389/fcomp.2023.1117848>
- [53] Jürgen Holl, Günter Kernbeiß, and Michael Wagner-Pinter. 2018. Das AMS-Arbeitsmarktchancen-Modell.
- [54] Judith E. Innes and David E. Booher. 2003. Collaborative policymaking: governance through dialogue. In *Deliberative Policy Analysis* (1 ed.), Maarten A. Hajer and Hendrik Wagenaar (Eds.). Cambridge University Press, 33–59. <https://doi.org/10.1017/CBO9780511490934.003>
- [55] IL Janis. 1971. Groupthink/Janis. *IL//Psychology Today* 5 (1971), 6.
- [56] Irving L Janis. 1972. Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. , viii, 277–viii, 277 pages.
- [57] David W. Johnson and Roger T. Johnson. 1985. *The Internal Dynamics of Cooperative Learning Groups*. Springer US, Boston, MA, 103–124. https://doi.org/10.1007/978-1-4899-3650-9_4
- [58] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
- [59] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. DeliData: A Dataset for Deliberation in Multi-party Problem Solving. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–25. <https://doi.org/10.1145/3610056>
- [60] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 702–714. <https://doi.org/10.1145/3531146.3533135>
- [61] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. 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 749, 22 pages. <https://doi.org/10.1145/3613904.3642849>
- [62] Frank Keil. 2019. How Do Partial Understandings Work? In *Varieties of Understanding*. Oxford University Press, 191–208. <https://doi.org/10.1093/oso/9780190860974.003.0010>
- [63] Frank C. Keil. 2003. Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences* 7, 8 (Aug. 2003), 368–373. [https://doi.org/10.1016/S1364-6613\(03\)00158-X](https://doi.org/10.1016/S1364-6613(03)00158-X)
- [64] Frank C. Keil. 2006. Explanation and Understanding. *Annual Review of Psychology* 57, 1 (2006), 227–254. <https://doi.org/10.1146/annurev.psych.57.102904.190100>
- [65] Norbert L. Kerr and R. Scott Tindale. 2004. Group Performance and Decision Making. *Annual Review of Psychology* 55, 1 (Feb. 2004), 623–655. <https://doi.org/10.1146/annurev.psych.55.090902.142009>
- [66] Max F. Kramer, Jana Schaible Borg, Vincent Conitzer, and Walter Sinnott-Armstrong. 2018. When Do People Want AI to Make Decisions?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) (AIES '18). Association for Computing Machinery, New York, NY, USA, 204–209. <https://doi.org/10.1145/3278721.3278752>
- [67] Richard A Krueger. 2004. *Focus groups : a practical guide for applied research* (3. ed., 6. print. ed.). Sage, Thousand Oaks, Calif. [u.a.]
- [68] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, Atlanta Georgia USA, 126–137. <https://doi.org/10.1145/2678025.2701399>
- [69] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, San Jose, CA, USA, 3–10. <https://doi.org/10.1109/VLHCC.2013.6645235>
- [70] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.
- [71] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Selsing, 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. <https://doi.org/10.1016/j.artint.2021.103473>
- [72] 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.
- [73] 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. <https://doi.org/10.1145/3359283>
- [74] 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. <https://doi.org/10.1145/3313831.3376590>
- [75] Q. Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-Driven Design Process for Explainable AI User Experiences. arXiv:<https://arxiv.org/abs/2104.03483> [cs.HC]
- [76] Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (Orlando, Florida, USA) (UbiComp '09). Association for Computing Machinery, New York, NY, USA, 195–204.

- <https://doi.org/10.1145/1620545.1620576>
- [77] Gabriel Lima, Nina Grgic-Hlaca, Jin Keun Jeong, and Meeyoung Cha. 2023. Who Should Pay When Machines Cause Harm? Laypeople's Expectations of Legal Damages for Machine-Caused Harm. 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, 236–246. <https://doi.org/10.1145/3593013.3593992>
- [78] Duri Long, Aadarsh Padiyath, Anthony Teachey, and Brian Magerko. 2021. The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences. In *Creativity and Cognition*. ACM, Virtual Event Italy, 1–10. <https://doi.org/10.1145/3450741.3465264>
- [79] Paola Lopez. 2019. Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. In *Proceedings of the 18th Annual STS Conference*. Graz, 289–309. <https://doi.org/10.3217/978-3-85125-668-0-16>
- [80] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) (*NIPS'17*). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.
- [81] Arthur Lupia. 2023. By Design: How People Adapt to Cognitive Limitations in Politics. *Topics in Cognitive Science* (Sept. 2023), tops.12690. <https://doi.org/10.1111/tops.12690>
- [82] David Mair, Laura Smillie, Giovanni La Placa, Florian Schwendinger, Milena Raykovska, Zsuzsanna Pasztor, René van Bavel, and European Commission (Eds.). 2019. *Understanding our political nature: how to put knowledge and reason at the heart of political decision-making*. Publications Office, Luxembourg. <https://doi.org/10.2760/374191>
- [83] Munir Mandviwalla and Lorne Olfman. 1994. What do groups need? A proposed set of generic groupware requirements. *ACM Trans. Comput.-Hum. Interact.* 1, 3 (Sept. 1994), 245–268. <https://doi.org/10.1145/196699.196715>
- [84] Joseph A. Maxwell. 2010. Using Numbers in Qualitative Research. *Qualitative Inquiry* 16, 6 (2010), 475–482. <https://doi.org/10.1177/1077800410364740>
- [85] Hugo Mercier and Hélène Landemore. 2012. Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation. *Political Psychology* 33, 2 (April 2012), 243–258. <https://doi.org/10.1111/j.1467-9221.2012.00873.x>
- [86] Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34, 2 (April 2011), 57–74. <https://doi.org/10.1017/S0140525X10000968>
- [87] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [88] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv:1811.11839 [cs.HC]* <https://arxiv.org/abs/1811.11839>
- [89] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). <https://christophm.github.io/interpretable-ml-book>
- [90] David Moshman and Molly Geil. 1998. Collaborative Reasoning: Evidence for Collective Rationality. *Thinking & Reasoning* 4, 3 (July 1998), 231–248. <https://doi.org/10.1080/135467898394148>
- [91] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Explainable recommendation: when design meets trust calibration. *World Wide Web* 24, 5 (Sept. 2021), 1857–1884. <https://doi.org/10.1007/s11280-021-00916-0>
- [92] Mohammad Naiseh, Nan Jiang, Jianbing Ma, and Raian Ali. 2020. Personalising Explainable Recommendations: Literature and Conceptualisation. In *Trends and Innovations in Information Systems and Technologies*, Álvaro Rocha, Hojjat Adeli, Luís Paulo Reis, Sandra Costanzo, Irena Orovic, and Fernando Moreira (Eds.), Vol. 1160. Springer International Publishing, Cham, 518–533. https://doi.org/10.1007/978-3-030-45691-7_49 Series Title: *Advances in Intelligent Systems and Computing*.
- [93] Mohammad Naiseh, Catherine Webb, Tim Underwood, Gopal Ramchurn, Zoe Walters, Navamayooran Thavanesan, and Ganesh Vigneswaran. 2024. XAI for group-AI interaction: towards collaborative and inclusive explanation. In *World conference for explainable artificial intelligence (17/07/24 - 19/07/24)*. <https://eprints.soton.ac.uk/493227/>
- [94] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* 2, 2 (Jan. 2018), 126–132. <https://doi.org/10.1038/s41562-017-0273-4>
- [95] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 568–578. <https://doi.org/10.18653/v1/N16-1070>
- [96] Jedrzej Niklas, Karolina Sztandar-Sztanderska, and Katarzyna Szymielewicz. 2015. Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making. panoptikon.org/sites/default/files/leadimage-biblioteka/panoptikon_profiling_report_final.pdf
- [97] Timothy J. Nokes-Malach, J. Elizabeth Richey, and Soniya Gadgil. 2015. When Is It Better to Learn Together? Insights from Research on Collaborative Learning. *Educational Psychology Review* 27, 4 (Dec. 2015), 645–656. <https://doi.org/10.1007/s10648-015-9312-8>
- [98] Legal Aid Ontario. 2025. What is a legal clinic? - Legal Aid Ontario — legalaid.on.ca. <https://www.legalaid.on.ca/faq/what-is-a-legal-clinic/> [Accessed 03-04-2025].
- [99] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junkleowitz, 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. <https://doi.org/10.1145/3593013.3594069>
- [100] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 35 (March 2022), 33 pages. <https://doi.org/10.1145/3495013>

- [101] Michael Quinn Patton. 1990. *Qualitative evaluation and research methods*, 2nd ed. Sage Publications, Inc, Thousand Oaks, CA, US. 532–532 pages.
- [102] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 959–972. <https://doi.org/10.1145/3531146.3533158>
- [103] 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. <https://doi.org/10.1145/2939672.2939778>
- [104] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26, 5 (2002), 521–562. https://doi.org/10.1207/s15516709cog2605_1
- [105] Brian K. Sato, Cynthia F. C. Hill, and Stanley M. Lo. 2019. Testing the test: Are exams measuring understanding? *Biochemistry and Molecular Biology Education* 47, 3 (May 2019), 296–302. <https://doi.org/10.1002/bmb.21231>
- [106] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2023. On the Impact of Explanations on Understanding of Algorithmic Decision-Making. 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, 959–970. <https://doi.org/10.1145/3593013.3594054>
- [107] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2024. Information That Matters: Exploring Information Needs of People Affected by Algorithmic Decisions. arXiv:2401.13324 [cs.HC]
- [108] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztañdar-Sztañderska, 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. <https://doi.org/10.1145/3531146.3534631>
- [109] Hong Shen, Alicia DeVos, Motahare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–29. <https://doi.org/10.1145/3479577>
- [110] Don Donghee Shin. 2023. *Algorithms, humans, and interactions: How do algorithms interact with people? designing meaningful AI experiences* (first edition ed.). Routledge, Boca Raton, FL. <https://doi.org/10.1201/b23083>
- [111] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press, Oxford.
- [112] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Enhancing Fairness Perception – Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. *International Journal of Human-Computer Interaction* (July 2022), 1–28. <https://doi.org/10.1080/10447318.2022.2095705>
- [113] M. K. Smith, W. B. Wood, W. K. Adams, C. Wieman, J. K. Knight, N. Guild, and T. T. Su. 2009. Why Peer Discussion Improves Student Performance on In-Class Concept Questions. *Science* 323, 5910 (Jan. 2009), 122–124. <https://doi.org/10.1126/science.1165919>
- [114] 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. <https://doi.org/10.1145/3531146.3534639>
- [115] Timo Speith, Barnaby Crook, Sara Mann, Astrid Schomäcker, and Markus Langer. 2024. Conceptualizing understanding in explainable artificial intelligence (XAI): an abilities-based approach. *Ethics and Information Technology* 26, 2 (June 2024), 40. <https://doi.org/10.1007/s10676-024-09769-3>
- [116] Jennifer Stromer-Galley. 2007. Measuring Deliberation's Content: A Coding Scheme. *Journal of Deliberative Democracy* 3, 1 (July 2007). <https://doi.org/10.16997/jdd.50>
- [117] Alistair Sutcliffe. 2005. Applying small group theory to analysis and design of CSCW systems. In *Proceedings of the 2005 workshop on Human and social factors of software engineering - HSSE '05*. ACM Press, St. Louis, Missouri, 1–6. <https://doi.org/10.1145/1083106.1083119>
- [118] Swiss Confederation. 2025. The referendum. ch.ch/en/votes-and-elections/referendum.
- [119] András Szigetvári. 2018. AMS bewertet Arbeitslose künftig per Algorithmus. *Der Standard* (2018). <https://www.derstandard.at/story/2000089095393/ams-bewertet-arbeitslose-kuenftig-per-algorithmus>
- [120] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. <https://doi.org/10.1145/3397481.3450662>
- [121] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy Artificial Intelligence. *Electronic Markets* 31, 2 (June 2021), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- [122] Marcel V. J. Veemman, Bernadette H. A. M. Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning* 1, 1 (April 2006), 3–14. <https://doi.org/10.1007/s11409-006-6893-0>
- [123] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (IUI '21). Association for Computing Machinery, New York, NY, USA, 318–328. <https://doi.org/10.1145/3397481.3450650>
- [124] P. C. Wason. 1968. Reasoning about a Rule. *Quarterly Journal of Experimental Psychology* 20, 3 (1968), 273–281. <https://doi.org/10.1080/14640746808400161>
- [125] Robert Stuart Weiss. 1995. *Learning from strangers: The art and method of qualitative interview studies* (1. free press paperback ed.). Free Press, New York, NY.
- [126] 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. <https://doi.org/10.1145/3613904.3642563>
- [127] Georg Wenzelburger, Pascal D. König, Julia Felfeli, and Anja Achtziger. 2022. Algorithms in the public sector. Why context matters. *Public Administration* (Dec. 2022), padm.12901. <https://doi.org/10.1111/padm.12901>
- [128] Maranke Wieringa. 2023. “Hey SyRI, tell me about algorithmic accountability”: Lessons from a landmark case. *Data & Policy* 5 (2023), e2. <https://doi.org/10.1017/dap.2022.39>
- [129] Grant P. Wiggins and Jay McTighe. 2005. *Understanding by design* (expanded 2nd ed.). Association for Supervision and Curriculum Development, Alexandria, VA.
- [130] Wei Xu. 2019. Toward human-centered AI: A perspective from human-computer interaction. *Interactions* 26, 4 (jun 2019), 42–46. <https://doi.org/10.1145/3328485>
- [131] Linda Zagzebski. 2019. Toward a Theory of Understanding. In *Varieties of Understanding*. Oxford University Press, 123–136. <https://doi.org/10.1093/oso/9780190860974.003.0007>
- [132] 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. <https://doi.org/10.1007/s00146-022-01480-5>

SUPPLEMENTARY MATERIAL

BETTER TOGETHER? THE ROLE OF EXPLANATIONS IN SUPPORTING NOVICES IN INDIVIDUAL AND COLLECTIVE DELIBERATIONS ABOUT AI

The supplementary material provides additional information on the study and use case. In Section A, we provide the features used by the employment prediction algorithm, the mock newspaper article introducing the use case to participants, and the study task description. Section B lists the self-report and interview questions that were asked during the first and second self-reports. Section C depicts all explanations that participants could receive in the study. Lastly, Section D provides the complete code book, listing themes and codes generated from the qualitative analysis.

A ADDITIONAL INFORMATION ON THE EMPLOYMENT PREDICTION ALGORITHM

In this section, we give additional information on the features used by the employment prediction algorithm, the mock newspaper article introducing the system to participants, and the study task participants solved.

A.1 Features used by the employment prediction algorithm

Table 5. The employment prediction algorithm uses a small set of features to calculate employability scores, including features describing demographic attributes, education, and past occupation, with "prior occupational career" being constituted by four variables. The term "cases" describes the number of times a job-seeker registered at the employment agency, "intervals" refers to a pre-defined time range, and "measures" describe support measures such as qualification courses and subsidization.

Variable	<ul style="list-style-type: none"> • Nominal values
Gender	<ul style="list-style-type: none"> • Male/Female
Age group	<ul style="list-style-type: none"> • 0–29/30–49/50+
Citizenship	<ul style="list-style-type: none"> • [Deployment country]/EU/Non-EU
Highest level of education	<ul style="list-style-type: none"> • Grade school/apprenticeship, vocational school/high- or secondary school, university
Health impairment	<ul style="list-style-type: none"> • Yes/No
Obligations of care (only women)	<ul style="list-style-type: none"> • Yes/No
Occupational group	<ul style="list-style-type: none"> • Production sector/service sector
Regional labor market	<ul style="list-style-type: none"> • Five categories for employment prospects in assigned job center
Prior occupational career	<ul style="list-style-type: none"> • [Variables as described below]
Days of gainful employment within 4 years	<ul style="list-style-type: none"> • <75%/≥75%
Cases within four 1-year intervals	<ul style="list-style-type: none"> • 0 cases/1 case/min. 1 case in 2 intervals/min. 1 case in 3 or 4 intervals • 0 cases/min. 1 case
Cases with a duration longer than 180 days	<ul style="list-style-type: none"> • 0/min. 1 supportive/min. 1 educational/min. 1 subsidized employment
Measures claimed	<ul style="list-style-type: none"> • 0/min. 1 supportive/min. 1 educational/min. 1 subsidized employment

A.2 Mock newspaper article

Special issue

The Review

33

Unemployed to be rated by algorithm

The potential of unemployed individuals will soon be comprehensively screened by a computer program

After years of development, a new IT program will be introduced nationwide in January, aimed at assessing the employment prospects of all unemployed individuals. Using a wide range of data, the algorithm will sort job-seekers into three categories: high, medium, and low chances of returning to the workforce.

The system analyzes personal data, including employment history, frequency and length of unemployment, and the individual's professional background. Key factors like age, nationality, and education are also considered. In addition, the algorithm takes into account other variables such as local labor market conditions.

Explicit three-part division

Much of this data is processed automatically, with the system pulling information from national social insurance databases. The criteria used to classify job seekers are already a vital part of daily job counseling, helping advisors tailor support to each individual's situation.

A key change with the new program is the explicit division of the unemployed into categories based on their chances of finding work. The 4,500 employment advisors will soon have access to the system's recommendations for each job-seeker they assist.

Individuals with a high probability of finding employment are those with a 66% chance of securing a three-month position within seven months. Those classified as having low prospects are predicted to have less than a 25% likelihood of obtaining six months of employment within two years. All others fall into the medium-prospects group.

Opinions are divided

The labor office board has stated that, initially, the algorithm's assessment will not directly impact decision-making. Advisors will continue to manage the allocation of support measures. However, new objectives tied to the three-part classification are in development, which could eventually influence how labor office funding is distributed.

According to the labor office board, the aim of the initiative is to make labor market resources more efficient in the long run. However, opinions on what exactly this means are divided.

Employers back the plan. "Anything that increases the chances of job placement is good from our perspective," said a representative of the Chamber of Commerce. The top priority, he added, must be to "use resources efficiently."

While employees are not fundamentally opposed to the program, they remain more cautious. A representative of the workers' union noted that they had set two conditions for the new system. First, advisors must retain the ability to override the computer-generated classifications, which he believes has been ensured. "It was also important to us that the group with poor prospects continues to receive support to help them re-enter the job market." R



Fig. 9. Mock newspaper article. Participants received initial information about the employment prediction algorithm in the form of a mock newspaper article. The article provided key information and featured the perspectives of employers and employee associations.

A.3 Task description

Task

Mr. Harald G., 49, has spent his life working as a waiter. Due to a knee surgery, he has recently experienced extended periods of unemployment. Additionally, he had to care for his mother for an extended time. Now that his caregiving responsibilities have ended, he comes to the initial meeting highly motivated. He is eager to undergo retraining and make a fresh start in his career, now that he is once again flexible with his time.



Harald G.

Age:	49
Gender:	Male
Education:	Apprenticeship
Citizenship:	[Deployment country]
Obligations of care: No	
Occupational group: Service	
Employment history: Less than 75% in last four years	
Health impairment: Knee problems	

All characteristics not specified have the value 0!

1. Can Harald change the data stored about him (e.g. to correct it)?
 yes no

2. Which group is Harald assigned to by the system?
 High (>66%)
 Medium (<66% & >25%)
 Low (<25%)

3. What support measures will Harald receive?
 Qualifying, such as courses and further training
 Stabilizing and increased support
 None

4. Can Harald appeal against this decision?
 yes no

Fig. 10. **Study task.** After the first exploration phase with the explanations, participants received a fictional job-seeker case example describing Mr. Harald G.¹⁰: A fictional job-seeker with a brief backstory and a list of features that would be used to calculate his employment chances. Participants solved four tasks formulated as questions as depicted. The correct answers are here marked with checked boxes. Whereas tasks 1, 3, and 4 required mostly information retrieval, task 2 could be solved in two ways: by either giving an estimate based on the rough weightings in the *system details* base explanations or by calculating the precise employment score. Participants had 15 (focus groups) or 20 (single interviews) minutes to solve the tasks. During that time, they could access and request all explanations and discuss possible solutions.

B SELF-REPORTS AND INTERVIEW GUIDE

B.1 Self-reports

Participants gave self-reports twice in the study, before and after the explanation phase (described in Section 3). In the following, we list each self-report question and the available answers.

- Understanding I + II
“I think that I understand the system...”
(1 = very little; 2 = little; 3 = neither/nor; 4 = well; 5 = very well)
- Individual decision I + II
“In your opinion, should the system be introduced?” (Yes / No)
- Decision confidence I + II
“In making this decision, I am...”
(1 = very uncertain; 2 = uncertain; 3 = neither/nor; 4 = certain; 5 = very certain)
- Explanation helpfulness
“Which explanations did you find most helpful for your understanding?”
(choose any from: *data, system details, usage, context*)
- Explanation influence on decision
“Which explanations were most influential to your decision?”
(choose any from: *data, system details, usage, context*)
- Contributing your voice (focus groups only)
“I was able to contribute my voice in the group discussion...”
(1 = very little; 2 = little; 3 = neither/nor; 4 = well; 5 = very well)
- Influence of discussion (focus groups only)
“The group discussion influenced my decision...”
(1 = very little; 2 = little; 3 = neither/nor; 4 = strongly; 5 = very strongly)

B.2 Interview guide

During the second self-report of participants, the investigator asked interview questions about participants' interaction with the explanations, their understanding processes, the most relevant information, and any additional situational questions. In the following, we list the questions composing the interview guide. The questions about inclusion and voice in the group were omitted in single interviews.

- Understanding II
How did the explanations help you to understand the system?
What did you find difficult to understand?
And how did the collaboration help you?
Was something missing? An explanation or a question?
- Individual decision II
How do you feel about this decision?
- Decision confidence II
How have the explanations and the collaboration influenced your decision confidence?
- Explanation helpfulness

Which of the explanations made you realize: Ah, I've understood something, that's good to know. And why? What effect did that have?

How did you communicate this to the group?

- Explanation influence on discussion

Which explanation made you think: Oh, that's important. It changes how I think about it. And why?

- Contributing your voice

How did you feel about the discussion process? Was everyone able to say everything?

- Influence of discussion

How do you feel about the decision the group made?

C COLLECTION OF EXPLANATIONS

This section depicts all explanations of the four categories *data* (Figure 12), *system details* (Figure 13), *usage* (Figure 14), and *context* (Figure 15). All explanation categories are split into three levels of detail; the background is colored differently for each level to facilitate distinction. Explanations of the base level were provided automatically; all others could be requested during the explanations phase. A detailed description of the explanations is given in Section 3. We insert the explanation overview again as Figure 11 for orientation.

Overview of explanations		
<i>You will receive the base level automatically. All further explanations are available on request at any time.</i>		
Dossier 1: Data		
A Format and structure B Data content C Limitations and risks		
Base	<input checked="" type="checkbox"/> How large is the data set and how was it collected? <input type="checkbox"/> What is the data's content? <input type="checkbox"/> What are populations?	<input type="checkbox"/> Is the data secure / complete / sound?
Level 2	<input type="checkbox"/> What does the data set look like? <input type="checkbox"/> Does the data represent the population?	<input type="checkbox"/> Can the data be misused?
Level 3	<input type="checkbox"/> Could the data set change over time? <input type="checkbox"/> What is not represented by the data?	<input type="checkbox"/> Can people obtain information about their own data and correct it?
Dossier 2: System details		
	A Features and weights B System process C Examples	
Base	<input checked="" type="checkbox"/> Which features does the system process and how are they weighted? <input type="checkbox"/> What is the employment chance and how is it calculated?	<input type="checkbox"/> What chance would the system calculate for Martin and Schifteh?
Level 2	<input type="checkbox"/> What are the exact weights of the features? <input type="checkbox"/> How accurate are the system's predictions?	<input type="checkbox"/> Why do people with similar characteristics have different employment chances?
Level 3	<input type="checkbox"/> How are the weights of the features calculated? <input type="checkbox"/> What errors can the system make?	<input type="checkbox"/> What chance would the system calculate for me?
Dossier 3: Usage		
	A Operation B Integration C Interaction with people	
Base	<input checked="" type="checkbox"/> Who operates the system? <input type="checkbox"/> How is the system integrated into the counseling?	<input type="checkbox"/> What does the group assignment mean for the people?
Level 2	<input type="checkbox"/> How can the system be checked and corrected? <input type="checkbox"/> How would the system make the work of counselors easier or more difficult?	<input type="checkbox"/> What interaction is there between system and jobseekers?
Level 3	<input type="checkbox"/> Why does automation bias play a role? <input type="checkbox"/> Why can't the system take over all job counseling?	<input type="checkbox"/> Do misallocations cause harm to those affected?
Dossier 4: Context		
	A Purpose and intention B Target group C Responsibility	
Base	<input checked="" type="checkbox"/> What is the official purpose and intention of the system?	<input type="checkbox"/> Who is the target group of the system? <input type="checkbox"/> Who is responsible for the system?
Level 2	<input type="checkbox"/> Why were these features selected? <input type="checkbox"/> How was the target group involved in the development process?	<input type="checkbox"/> How are the disadvantages of minorities compensated?
Level 3	<input type="checkbox"/> What is the political background to the introduction of the system? <input type="checkbox"/> Can people who are affected appeal against decisions?	<input type="checkbox"/> What ethical standards were used to guide the development of the system?

Fig. 11. Explanation overview.

<p>Dossier 1: Data</p> <p>A Form and structure</p> <p>How large is the data set and how was it collected?</p> <p>Scope : 860,277 entries for business cases . A business case refers to the period in which a person is unemployed and registered with the employment agency. This means that several business cases can exist for one person.</p> <p>Period : The data describes a period over the last four years.</p> <p>Storage location : Data warehouse on the employment agency server.</p> <p>Collection: The data was taken from the inventory of the Association of Social Insurance Funds. Consultants in the agency can supplement or correct the data if necessary.</p>	<p>Dossier 1: Data</p> <p>B Content of the data</p> <p>What is the content of the data?</p> <p>Each entry in a business case contains the characteristics of the job seeker as well as information on whether and how often they have found short-term or long-term work over the course of four years.</p> <p>What are populations?</p> <p>To calculate the employment opportunity, the job seekers in the data are grouped into four "populations". These populations differ in the completeness of their data and in the characteristics used to calculate the employment opportunity:</p> <ul style="list-style-type: none"> • "Full valid" population: Complete information over a 4-year period • "Partially valid" populations: People for whom information over a 4-year period is incomplete, roughly divided into <ul style="list-style-type: none"> ▪ Young people under 25 ▪ Recently immigrated ▪ People with an interrupted employment history 																																																
<p>Dossier 1: Data</p> <p>B Content of the data</p> <p>C Limitations and risks</p> <p>Is the data secure/complete/error-free?</p> <p>Secure : Yes.</p> <p>The data is stored encrypted on the employment agency's servers and is subject to the General Data Protection Regulation.</p> <p>Complete : Partial.</p> <p>Since data is not available on all people, the data set is divided into different "populations". For example, newly arrived people who do not yet have four years of data, as well as young people (<25 years) and people with an interrupted employment history belong to separate populations. Different characteristics are sometimes taken into account between the populations and the weighting of the characteristics differs.</p> <p>Error-free : Most likely.</p> <p>Since the data is obtained from the social insurance and the employment agency's internal databases, errors in the sense of incorrect information are unlikely. There are also quality controls.</p>	<p>Dossier 1: Data</p> <p>A Form and structure</p> <p>What does this data set look like?</p> <p>Here are five example entries:</p> <table border="1"> <thead> <tr> <th>Name</th> <th>Gender</th> <th>Age</th> <th>Education</th> <th>Health impairment?</th> <th>Obligations of care?</th> <th>Occupation</th> <th>Occupation history</th> </tr> </thead> <tbody> <tr> <td>Martina</td> <td>m</td> <td>27</td> <td>Compulsory education</td> <td>No</td> <td>No</td> <td>Cashier</td> <td>>75% in 4 years</td> </tr> <tr> <td>Sophie</td> <td>w</td> <td>29</td> <td>Master</td> <td>No</td> <td>No</td> <td>Study Assistant</td> <td>>75% in 4 years</td> </tr> <tr> <td>Harald</td> <td>m</td> <td>49</td> <td>Apprenticeship</td> <td>Yes (knee surgery)</td> <td>No</td> <td>Waiter</td> <td>>75% in 4 years</td> </tr> <tr> <td>Michael</td> <td>m</td> <td>34</td> <td>Apprenticeship</td> <td>No</td> <td>No</td> <td>Locksmith</td> <td>>75% in 4 years</td> </tr> <tr> <td>Sabine</td> <td>w</td> <td>34</td> <td>Apprenticeship</td> <td>No</td> <td>Yes</td> <td>Clerk</td> <td>>75% in 4 years</td> </tr> </tbody> </table> <p><i>Important: These entries are for illustrative purposes only and do not reflect the actual form of the data set. In addition, the information is reduced for the sake of simplicity. All people are of course fictitious.</i></p>	Name	Gender	Age	Education	Health impairment?	Obligations of care?	Occupation	Occupation history	Martina	m	27	Compulsory education	No	No	Cashier	>75% in 4 years	Sophie	w	29	Master	No	No	Study Assistant	>75% in 4 years	Harald	m	49	Apprenticeship	Yes (knee surgery)	No	Waiter	>75% in 4 years	Michael	m	34	Apprenticeship	No	No	Locksmith	>75% in 4 years	Sabine	w	34	Apprenticeship	No	Yes	Clerk	>75% in 4 years
Name	Gender	Age	Education	Health impairment?	Obligations of care?	Occupation	Occupation history																																										
Martina	m	27	Compulsory education	No	No	Cashier	>75% in 4 years																																										
Sophie	w	29	Master	No	No	Study Assistant	>75% in 4 years																																										
Harald	m	49	Apprenticeship	Yes (knee surgery)	No	Waiter	>75% in 4 years																																										
Michael	m	34	Apprenticeship	No	No	Locksmith	>75% in 4 years																																										
Sabine	w	34	Apprenticeship	No	Yes	Clerk	>75% in 4 years																																										
<p>Dossier 1: Data</p> <p>B Content of the data</p> <p>Do the data represent the population?</p> <p>The dataset contains records of people who were registered as unemployed with the employment agency. The distribution of the characteristics does not correspond to the entire population, but rather to this subgroup.</p> <p>Theoretically, a broad representation of people is possible through the combination of characteristics ("constellations"). There are around 81,000 of these combinations. In practice, data actually exists for around 31,000 of them.</p>	<p>Dossier 1: Data</p> <p>C Limitations and risks</p> <p>Can the data be used for other purposes?</p> <p>In principle, no.</p> <p>The General Data Protection Regulation ensures that data may only be used with consent and only for the defined purpose .</p> <p>In this case, that means: The data is used by the employment agency's consultants to calculate the individual's chance of employment.</p> <p>Use for other purposes, by other authorities or by people outside the employment agency is not legally permitted , unless there is specific consent from the job seeker.</p>																																																
<p>Dossier 1: Data</p> <p>A Form and structure</p> <p>Could the data set change over time?</p> <p>Yes.</p> <p>The data is updated annually to update the characteristics and employment relationships of the people. Information from further back in time is also deleted. If aspects of the employment agency's advisory system are changed, such as the definition of care responsibilities or employment relationships, these changes must also be implemented in the data.</p>	<p>Dossier 1: Data</p> <p>B Content of the data</p> <p>What is not represented by the data?</p> <p>Non-measurable factors and difficult living conditions</p> <p>Basically, the data can only capture values that can be represented with numbers or categories. Aspects such as personal motivation, housing situation or interpersonal skills are therefore not shown. Information on addiction, debt and prison sentences are also not shown.</p> <p>External factors such as crises and major events</p> <p>Exceptional events such as economic crises or epidemics are also not directly represented in the data. This means that data collected in a "normal" year (e.g. 2020) are not directly comparable with data from a crisis year (e.g. 2022, i.e. the start of the pandemic).</p>																																																
<p>Dossier 1: Data</p> <p>C Limitations and risks</p> <p>Can people obtain information about their own data?</p> <p>Yes.</p> <p>People have a right to information about data that is stored about them by the employment agency. Corrections and deletions are also possible.</p> <p>Job seekers are also informed about the calculation of the employment opportunity and the allocation to one of the three groups (high, medium, low).</p>																																																	

Fig. 12. Data.

Dossier 2: System details

A Features and values

What characteristics does the system process and how are they weighted?

To calculate the employment chance, the system uses 12 characteristics, including demographic information and work history.

These characteristics are positive (+), negative (-) or have no effect (0) on the employment chance. Characteristics with no effect (0) correspond to the "base group" with which other groups are compared. Here is an overview of the characteristics and their weightings:

Feature	Value	Weight
Gender	M	0
Age group	<30	0
	30-49	-0.13
	50+	-0.05
Origin	[Deployment country]	0
Training	Compulsory education	0
	Apprenticeship	+
	Hobby or higher	-
Health impairment	Yes	-0.05
	No	0

Feature	Value	Weight
Obligation of care	Yes	0
	No	0
Occupational group	Production	0
	Service	-0.13
Regional labor market	1/2/3/4/5	0.12 / -0.13 / -0.05
Employment history	> 75% employed in a year	0
	How often losing job in last year	0
	Less than 6 months	-0.05
Participation in support measures	More	0
	1 times supportive	-0.05
	1 times qualifying	-0.05
	1 time employment prospecting	-0.05

B System process

What is the employment chance and how is it calculated?

The employment chance is intended to express how likely someone is to find work in a certain time frame. It is calculated by comparing them with people who have similar characteristics.

To calculate this chance, the system goes through five steps:

1. The characteristics of a person are used as input.
2. The previously calculated weights are assigned to the characteristics.
3. These weights are summed and the number 0.1 is added to the result.
4. The result is converted to a percentage: the employment chance.
5. The chance is used to decide which group the person falls into:

Chance of employment	more than 65% → High group
	less than 65% but more than 25% → Medium group
	less than 25% → Low group

C Examples

What is the employment chance would the system calculate for Martin and Schifteh? These examples, and the photos too, are of course fictitious!

Martin belongs to the "base group", which means that his characteristics are defined as standard and are weighted with "0". The chance of this basic group is a flat rate of 52%. Schifteh is compared with the base group. The weightings show whether people with these characteristics have had better or worse chances than the base group in the past.

Example 1: Martin	Characteristic	Weighting	Characteristic	Weighting
Gender: Male	0	Gender: Female	0	
Age: <30	0	Age: 30-49	-0.13	
Education: Compulsory school	0	Education: Apprenticeship	+	
Origin: [Deployment country]	0	Origin: Iran	0	
Labor market: Type 1	0	Labor market: Type 5	-0.05	
Sector: Production	0	Sector: IT industry	-0.05	
Employment opportunity: 52%	52	Employment opportunity: 30 %	30	
Group: Medium		Group: Medium		

Example 2: Schifteh	Characteristic	Weighting
Gender: Female	0	
Age: 30-49	-0.13	
Education: Apprenticeship	+	
Origin: Iran	0	
Labor market: Type 5	-0.05	
Sector: IT industry	-0.05	
Employment opportunity: 30 %	30	
Group: Medium		

For the sake of simplicity, not all 12 features of the profiles are shown here.

Dossier 2: System details

A Features and values

What are the exact weights of the features?

The exact weights depend on the population. Here are examples of the weights for the population with complete information:

Feature	Value	Weight
Gender	W	-0.14
Age group	30-49	-0.13
	50+	-0.05
Origin	[Deployment country]	0
Training	Apprenticeship	+
	Hobby or higher	-
Health impairment	Yes	-0.05
	No	0

Feature	Value	Weight
Obligation of care	Yes	0
	No	0
Occupational group	Production	0
	Service	-0.13
Regional labor market	1/2/3/4/5	0.12 / -0.13 / -0.05
Employment history	> 75% employed in a year	0
	How often losing job in last year	0
	Less than 6 months	-0.05
Participation in support measures	More	0
	1 times supportive	-0.05
	1 times qualifying	-0.05
	1 time employment prospecting	-0.05

B System process

How accurate are the system's predictions?

To evaluate the accuracy of the system, its predictions are tested on unknown data (172,054 entries). Then it is compared how many predictions match reality and how many do not. Example: Below, the system predicted that 25,042 people would work, but in fact they did not find work.

Accuracy: 74.9%	(proportion of correctly predicted cases)	Real case
Sensitivity: 79.3%	(proportion of positively predicted positive cases)	Did not find a job
Specificity: 70.2%	(proportion of negatively predicted negative cases)	Has found work

Prediction	Did not find a job	Has found work
Did not find a job	58,853	18,220
Has found work	25,041	69,940

For the sake of simplicity, not all 12 features of the profiles are shown here.

Dossier 2: System details

A Features and values

Why do results with similar characteristics have different employment opportunities?

These examples, and the photos, are of course fictitious!

Michael and Sabine are both parents. However, the characteristic "care responsibilities" can only be applied to women, so there is another difference in their employment opportunities (in addition to the characteristic "female").

Example 3: Michael	Characteristic	Weighting	Example 4: Sabine	Characteristic	Weighting
Gender: Male	0	Gender: Female	0		
Age: <30	0	Age: 30-49	-0.13		
Education: Apprenticeship	+	Education: Apprenticeship	+		
Origin: [Deployment country]	0	Origin: [Deployment country]	0		
Sector: Production	+	Sector: Production	+		
Employment opportunity: 66 %	66	Employment opportunity: 59 %	59		
Group: High		Group: Medium			

For the sake of simplicity, not all 12 features of the profiles are shown here.

Dossier 2: System details

A Features and values

How are the weightings of the characteristics calculated?

The weightings are calculated by a statistical comparison between groups of people. First, a **base group** is defined, i.e. a group of people whose characteristics serve as a standard.

In this system, these are men under 30, with a maximum of compulsory school education and [deployment country]'s citizenship. Easy to remember: The base group has *all characteristics with a weighting of 0*.

Once this base group is defined, a model (called "logistic regression") is trained for each population to predict the relationship between individual characteristics and the chance of employment.

This model therefore provides the answer to the question, for example: If the chance of the base group is 52%, what is the chance of a person who has the same characteristics but is a woman?

The answer: Since, according to the data, women find slightly less work than the base group, the model calculates a minus for the characteristic "female".

B System process

What mistakes can the system make?

One type of error is **miscalculation**: The system predicts a person's employment chance inaccurately (i.e. too low or too high). This means that they are assigned to an unsuitable group. With a hit accuracy of around 75% (**System Details B - Level 2**), the prediction is wrong for around 25% of people. Another error for certain groups of people can further reduce the accuracy.

Another error can arise from the data situation (**Data B - Level 3**). Since events such as economic crises or the pandemic are not represented in the data, the employment opportunities would not apply to the changed labor market during a crisis.

The assignment of the **regional labor market type** (**System Details B - Extra**) is another source of error. Since the place of residence determines this assignment and the boundaries between advantageous and disadvantageous labor market types can be geographically close together, a slightly changed residential address can lead to a significantly changed employment chance.

Dossier 2: System details

A Features and values

What chance would the system calculate for me?

To calculate your own employment chance, the steps listed in **System Details B - Base** must be carried out. (In the explanations, only feature weights of the "population with complete information" are given, so the results correspond to this population.)

For example:

Feature	Value	Weight
Gender	M	0
Age group	30-49	-0.13
	50+	-0.05
Origin	[Deployment country]	0
Training	Study	+0.13
	Apprenticeship	-0.05
Obligation of care	No	0
Occupational group	Production	0
Regional labor market	1/2/3/4/5	0.12 / -0.13 / -0.05
Employment history	> 75% employed in a year	0
	How often losing job in last year	0
	Less than 6 months	-0.05
Participation in support measures	1 time	-0.05
	Less than 6 months	-0.05
	1 times supportive	-0.05
	1 time employment prospecting	-0.05

Explanation of conversion to %: **System Details B - Extra**

Fig. 13. System details.

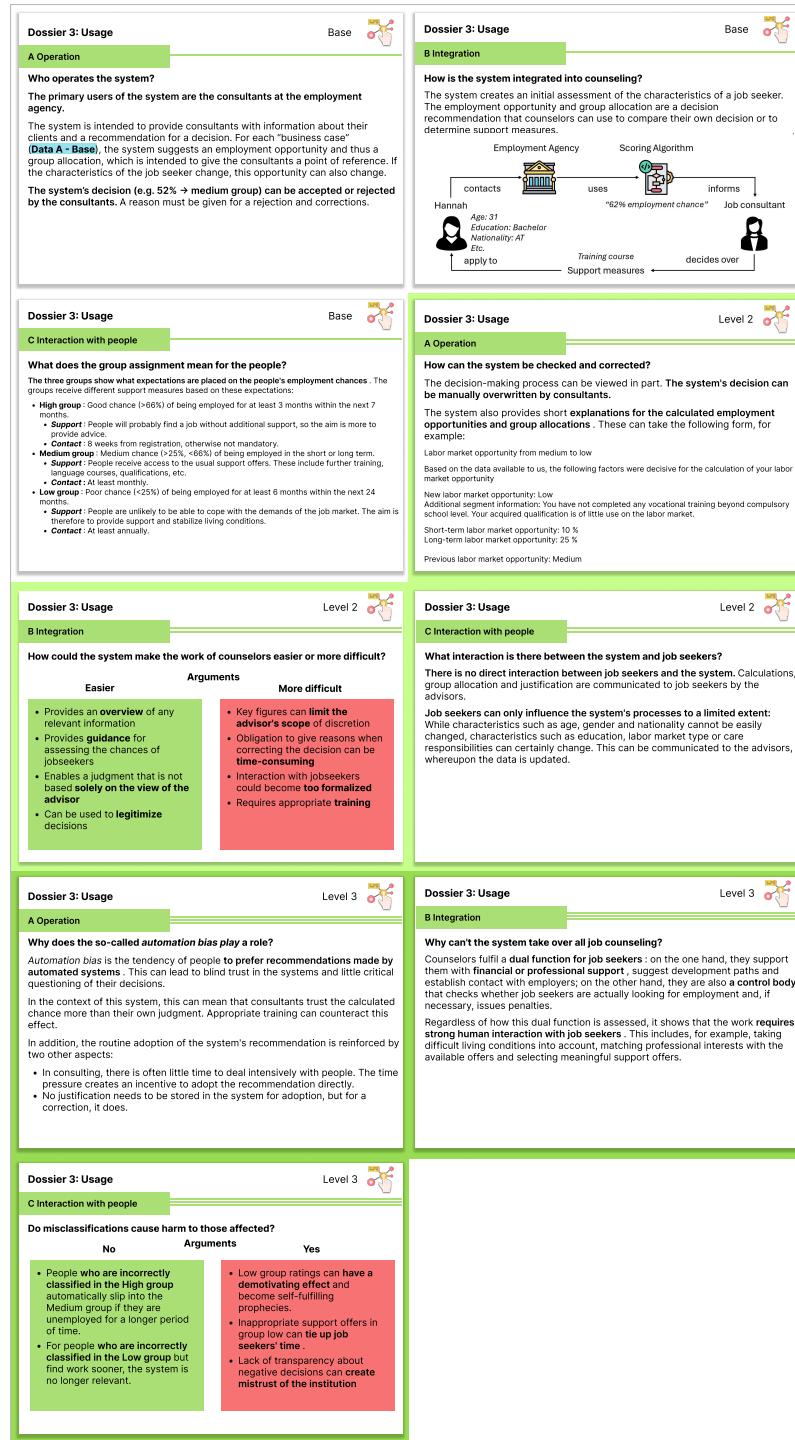


Fig. 14. Usage.

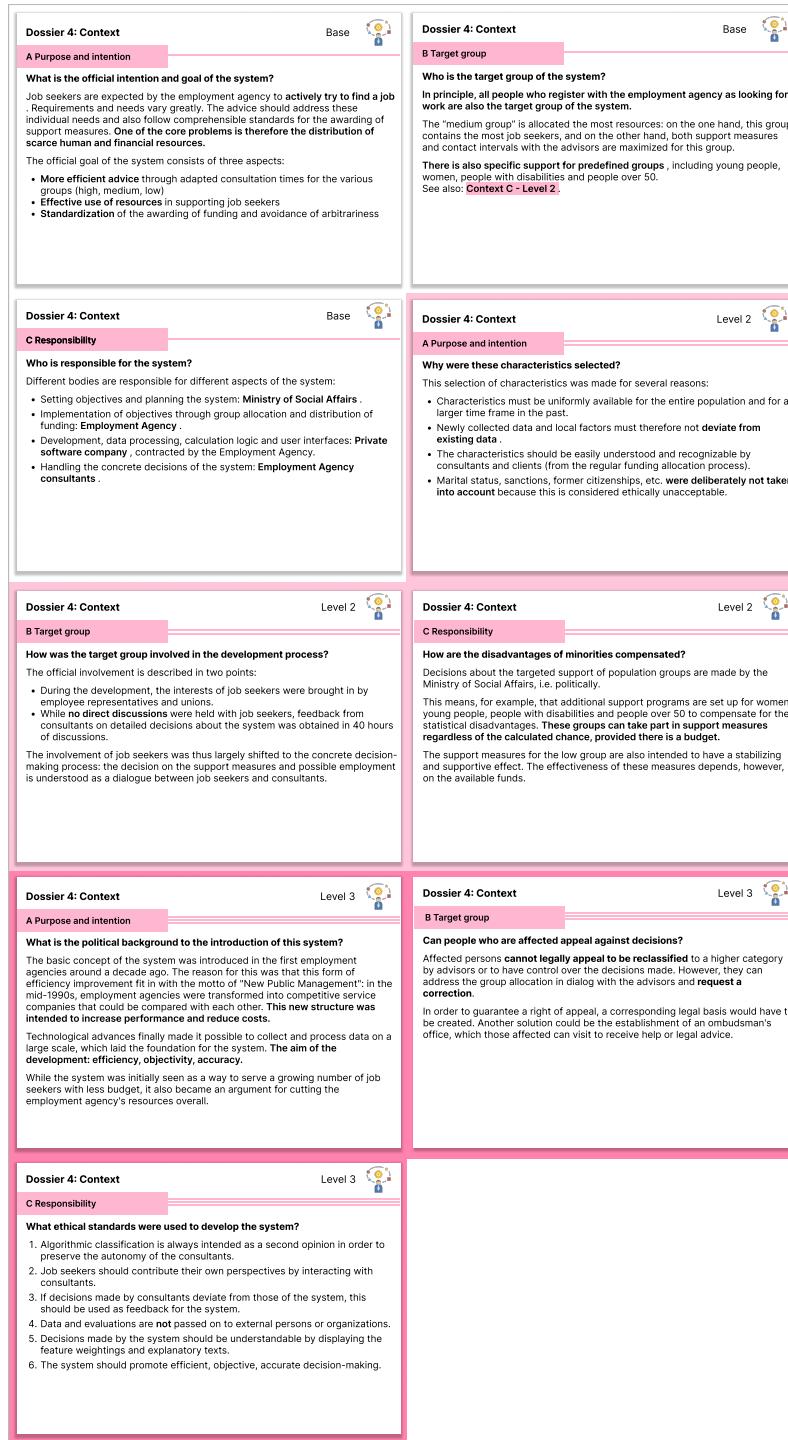


Fig. 15. Context.

D CODE BOOK

Table 6 lists the themes and codes developed from the interview data in the qualitative analysis. The table is split into three main sections: Deliberation, understanding, and experiences and opinions. The left column lists theme groups and the right columns list single themes as coded in the data.

Deliberation		
Deliberation - groups	<ul style="list-style-type: none"> • Appreciation of group setting • Deployment decision changed • Difficult to make deployment decision • Discussing diverging views • Discussion triggers • Does not know what to say • Following decisions of others 	<ul style="list-style-type: none"> • Forming opinions on deployment • Little discussion • No influence from group discussion • Strategic decision-making • Unanimous deployment decision • Voting against own interests • Weighing pros and cons
Arguments - groups	<ul style="list-style-type: none"> • Adverse cognitive effects • Conditions for deployment • Gaming the system • Influence of Human Factors • Integrating system into working processes • Scrutinizing the system 	<ul style="list-style-type: none"> • System can be misused • What's the benefit? • What's the intention? • Who can I trust? • Who's in control? • Who's responsible?
Deliberation - individuals	<ul style="list-style-type: none"> • AI harms trust in institution • AI will harm society • Attitude remained unchanged • Decision more uncertain than before 	<ul style="list-style-type: none"> • Deployment despite criticism • Difficult to make deployment decision • Weighing pros and cons • Would have liked a group setting
Arguments - individuals	<ul style="list-style-type: none"> • Conditions for deployment • Decision-makers are not the right people • Is it contestable? • Influence of human factors • Integration is the deciding factor • Judgment deviates from algorithm • Lack of transparency • Necessary to inform individuals • Projecting hopes on the AI 	<ul style="list-style-type: none"> • Scrutinizing the system • Supporting decision subjects • System inherits institutional dysfunctions • Unions should take a role • What are the consequences? • What's the benefit? • Who's in control? • Who's responsible?
Understanding		
Explanations - groups	<ul style="list-style-type: none"> • Cumbersome information uptake • Differing information needs • Exchanging explanation sheets • Explanation design flaws • Explanation structure does not make sense • Explanation structure works • Explanations help understanding 	<ul style="list-style-type: none"> • Incomplete coverage of information • Interest beyond time limit • Locating information • Order of processing information • Overwhelmed by information • Piecing together information • Relying on intuition over information

	<ul style="list-style-type: none"> • Explanations matched with participants • Gap between explanation and application 	<ul style="list-style-type: none"> • Requesting explanations in bulk • Suggestions for explanation design
Understanding - groups	<ul style="list-style-type: none"> • Abandoning understanding • Debating task solutions • Discussing interpretations • Dividing understanding burden • Impeded understanding • Is system already in use? 	<ul style="list-style-type: none"> • Outsourcing understanding • Participants work individually • Querying and explaining • Reaching working understanding • Sharing information with group • Suggestions for explanations in groups
Explanations - individuals	<ul style="list-style-type: none"> • Cumbersome information uptake • Difficult to locate information • Every category is important • Explanation design flaws • Explanation design suggestions • Explanations adjusted mental model • Explanations help understanding • Explanations influenced decision • Explanations matched participant • Explanations need to relate personally • Explanations not suited to decision subjects 	<ul style="list-style-type: none"> • Explanations require previous knowledge • Focus on context • Focus on system details • Focus on usage • Going into detail • Interest beyond time limit • Order of processing information • Overwhelmed by information • Relying on intuition over information • Requesting explanations in bulk • Skips category
Understanding - individuals	<ul style="list-style-type: none"> • Calculates employment chance • Faults in algorithmic design • Impeded understanding 	<ul style="list-style-type: none"> • Reaching a working understanding • Understanding vs information gain • Understanding requires example
Experiences and opinions		
Experiences - all	<ul style="list-style-type: none"> • AI is part of digitization • Comparing lived experiences • Decision subjects have no voice • Deficiencies in institution • Digital Humanism as institutional practice • Is not affected by system 	<ul style="list-style-type: none"> • No idea of AI • Overburdened Human-in-the-Loop • Using AI at work • Using AI without knowing it • Workplace wants to integrate AI
Opinions - all	<ul style="list-style-type: none"> • AI aversion • AI can assist in decisions • AI cannot replace humans • AI decisions must be revisable • AI increases objectivity • AI is inevitable • AI just appeared • AI misrepresents reality • AI openness • AI replaces humans • AI will not improve work 	<ul style="list-style-type: none"> • Deploying institution has bad reputation • Disagrees with policy choices • Discrimination with and without AI • Fears algorithmic imprint • Formalization is inevitable • Good intentions, badly executed • Need for AI-Human collaboration • No opinion • Peaked interest • Public narratives