

Deliberative XAI: How Explanations Impact Understanding and Decision-Making of AI Novices in Collective and Individual Settings

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

LAURA KOESTEN, University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis, Austria

TORSTEN MÖLLER, University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis, Research Network Data Science, Austria

SEBASTIAN TSCHIATSCHEK, University of Vienna, Faculty of Computer Science, Research Network Data Science, Research Group Data Mining and Machine Learning, Austria

XAI research often focuses on settings where people learn about and assess algorithmic systems individually. However, as more public AI systems are deployed, it becomes essential for XAI to facilitate collective understanding and deliberation. We conducted a task-based interview study involving 8 focus groups and 12 individual interviews to explore how explanations can support AI novices in understanding and forming opinions about AI systems. Participants received a collection of explanations organized into four information categories to solve tasks and decide about a system's deployment. These explanations improved or calibrated participants' self-reported understanding and decision confidence and facilitated group discussions. Participants valued both technical and contextual information and the self-directed and modular explanation structure. Our contributions include an explanation approach that facilitates both individual and collaborative interaction and explanation design recommendations, including active and controllable exploration, different levels of information detail and breadth, and adaptations to the needs of decision subjects.

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

1 Introduction

A growing number of algorithmic decision-making systems (ADM systems) are deployed and used in the public sector to decide about critical issues, such as employment, migration, and criminal justice [4, 19, 74, 95]. These systems can result in discrimination and damages for decision subjects [11, 68], especially when decision-making is opaque and uncontestable [1, 3, 22]. The field of explainable AI (XAI) aims to reduce these risks by developing methods to make ADM systems more understandable and facilitate their assessment and control [47]. But while much of XAI research is focused on individuals, research has shown that learning about and understanding complex topics benefits from collaboration in group settings [59, 63, 66, 79]. Work on public AI further argues that multi-stakeholder deliberation is an essential step in ensuring public participation in deployment decisions of these systems [11, 39, 62], but how explanations can fulfill these demands is an open question in XAI.

Crucially, many people affected by decisions of ADM systems have no technical background in developing or analyzing these systems. That is, a majority of decision subjects, domain experts, and other stakeholders of the public are lay people [24, 51, 78] or AI novices¹. Explanations for AI novices naturally need to fulfill different requirements than for AI practitioners, as they have to be adapted to their background [26], interests [47], and prior knowledge [18, 73, 82], which is why personalization [20, 61, 78] and interactivity [6, 18, 34] are central concepts in explanation design. Further, while explanation formats (e.g., interactive, visual, textual) are known to impact AI novices' understanding of algorithmic

¹We use the term in reference to "visualization novices" [13] and synonymously with "lay people" [24, 75, 82].

systems [8, 18, 82], the precise effect is inconsistent between studies [8] due to contextual factors such as participants' perceptions of the use case domain. Previous qualitative research identified the information needs of AI novices [73], emphasized user-preference for social and conversational formats [6], and employed participatory methods to design explanations in collaboration with end-users [48, 87]. This work connects to these insights and draws from disciplines such as CSCW [21, 30], cognitive science [41, 42], and educational psychology [66, 79] to examine how explanations can support understanding and decision-making of AI novices in individual and collaborative settings [62, 87].

Developing this design approach to improve understanding and facilitate deliberation meets multiple challenges. Group composition and dynamics place special demands on explanation design [60], as explanations must cover a diverse set of information needs [9, 73] and format preferences [6]. In addition, they must support collaborative interaction [52], such as sharing and simultaneous processing of information. That is, explanations need to provide comprehensive information but remain clear and navigable, and should foster group interaction but still be suitable for individual settings. We address these challenges by proposing a modular explanation design that spans four information categories from which users can select information according to their needs. Additionally, when validating explanation approaches qualitatively, it is crucial to contact the relevant stakeholder groups, which can be especially challenging with decision subjects in minority groups or difficult situations, who arguably have the highest risk of incurring negative consequences from ADM. To address this, we conducted two focus groups with decision subjects to include their perspectives and voices in developing explanations for AI systems in the public sector [11].

This paper explores how explanations can be used in individual and collaborative settings to support AI novices' learning, interaction, and decision-making regarding the deployment of ADM systems. We present the findings of a task-based interview study with 43 AI novices involving 8 focus groups and 12 single interviews. We propose an explanation design that addresses AI novices' information needs and applies current best practices in XAI [14]. The design consists of a collection of 36 explanations in the form of question-answer pairs organized into the four categories *data*, *system details*, *usage*, and *context*, which are further subdivided into subtopics and levels of detail (Figure 1). Participants used these explanations to solve the study tasks and decide about deploying an ADM system (Figure 3). The study's use case is an algorithm for the prediction of job-seekers' employability, connecting to previous work on ADM systems in employment [53, 65, 74]. We analyze participants' self-reported understanding, decision confidence, and perceptions of key information, and conduct a thematic analysis of group interactions with explanations. We discuss explanation design suggestions and the benefits and challenges of using explanations in collaborative settings. In the analysis of our findings, we put particular focus on the following research questions:

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

[RQ2] *Deliberation*: How do the explanations support group interaction with information and deliberation processes?

The contributions of our work include i) the validation of an explanation design building on current XAI principles, including question-driven and modular design, different levels of completeness and soundness, and suitability for both individual and collective settings; ii) an analysis of the kind and amount of information that participants requested most often and perceived as most important; iii) an analysis of how our explanations support group collaboration in solving tasks and group deliberation in forming opinions about ADM systems; and iv) recommendations for the design and use of these explanations and implications for the incorporation of group settings in XAI.

2 Background and Related Work

In this section, we briefly describe how our work connects to human-centered AI, research on explanation design for AI novices, and literature on understanding, interaction, and decision-making in groups.

2.1 Human-centered explainable AI

Explainability is often described as a cornerstone of responsible AI systems² [83], as explanations can enable people to understand [47] and contest AI decisions [1]. Similar concepts are embedded in the notion of human-centered AI [15], which proposes to 1) build systems based on user observation and stakeholder engagement and 2) empower rather than replace people by making these systems controllable and autonomy-preserving [76, 77, 91]. These measures become especially important in high-risk settings [29], such as employment [31, 74], immigration [4], or criminal justice [19], where erroneous and non-transparent algorithmic decisions can cause severe harm to those affected [68]. To better understand how explanations can contribute to these principles, human-centered explainable AI places emphasis on the sociotechnical dimensions of explaining AI systems [28] in that 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 [88], its lifecycle [25] and its different stakeholder groups [28]. In the context of this work, human-centered explainability is realized by testing and validating a design approach that has been developed based on qualitative exploration of AI novices' information needs [73]. Further, our work aims to examine suitable formats for explanations that can accompany the deployment of AI systems in public institutions [22] and inform approaches of practicing responsible AI in high-risk application settings [28].

2.2 Explanation design for AI novices

Explanations for non-technical audiences have different requirements than those for AI practitioners. Explanation methods like LIME [69], SHAP [54], and surrogate models [58] are better suited to experts due to their reliance on previous knowledge and technical literacy, while effective methods for AI novices are still being developed. As transparency does not equal understanding [3], simply making all information about a system available is no valid explanation strategy. Instead, explanations must be selected and designed to be suitable to the recipient's stakeholder role [47], prior knowledge [73], beliefs [57], and explanatory stance [14, 42]. Various studies examined how such explanations for AI novices might be realized: Cheng et al. [18] showed that "white-box" explanations (transparent models) can improve "objective"³ understanding but may overwhelm users, reducing perceived understanding. However, Bove et al. [8] used similar explanations and found that they had the opposite effect on participant understanding, which they attribute to a difference in the studies' use case domains (student admission vs. finance). Szymanski et al. [82] 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 underscore this variability, indicating that comprehension varies with personal and demographic factors and domain knowledge [26, 78, 84] and that personalized explanations could be effective in addressing this [20, 78]. Further aspects to be considered include the explanation's purpose [32] and the recipient's familiarity with AI [43]. Regarding the information needs, previous qualitative research outlined that AI novices value information about the context and intention of a system's deployment [39, 73] as well about the responsible institution [11]. Notably, Lee et al. [48] and Weitz et al. [87] conducted participatory workshops to design explanations with end-users in the public sector, finding that participants valued the collaborative settings and their

²While not every ADM system includes AI, we use both terms as we focus on decision-making systems with machine learning components.

³We use the term in reference to Cheng et al. [18] and Bove et al. [8], it means to describe factual or testable understanding.

improved understanding of algorithmic systems. Drawing from these approaches, this work focuses on the design and qualitative validation of explanations such that they address the information needs of AI novices [73] and support their understanding [47] and decision-making [6] in individual and collective settings.

Our explanation design draws from various approaches in previous work. The information base used for the explanations was compiled from different sources documenting the development and planned deployment of employment prediction algorithms [2, 74], producing an extensive collection of "scavenged" [89] material. To structure this collection, we drew from Lim and Dey [50]'s work on intelligibility types, Liao et al. [49]'s concept of question-driven explanation design, and Schmude et al. [73]'s separation of information categories into *data*, *system details*, *usage*, and *context*. We further apply Kulesza et al. [46]'s principles on explanation soundness (fidelity, complexity) and completeness (coverage, density) and draw from works that apply these principles in explanation design [16, 34, 45], by introducing a structure of sub-topics and a hierarchy of explanation levels. This combination of question-driven explanations, different levels of information detail and breadth, and user-controlled selection of information aims to introduce modularity and interactivity [18, 34, 72] while allowing for adjusting of explanations to users' needs [20, 78].

2.3 Understanding and decision-making in collaborative settings

Improving understanding is a central goal of explanations [47] and enables people to pursue their goals, such as assessing or contesting a system's decisions [36]. However, understanding can be defined in numerous ways [5, 33, 42, 93]. This work draws from literature in the learning sciences, cognitive sciences, and XAI to define understanding as i) connecting and applying information, whereas "knowledge" only stores information [5, 33], ii) being the attempt to grasp the underlying structure of a phenomenon by way of simplification [93], iii) consisting of several "facets" that include both the analytical and the emotional connection to information [90], and iv) rarely being exhaustive, but rather a "working" mental model that is attained by recognizing and filling gaps until the learner deems it sufficient [42]. Due to the challenge of defining and measuring understanding [71], recent research has proposed an "abilities-based" approach [80] which defines understanding through the skills the learner possesses, connecting to comparable operationalizations by the learning sciences [90].

While individual understanding has been the subject of many studies in XAI [17, 18, 72, 84], understanding in group settings has been less explored. However, different disciplines have investigated collaborative understanding in detail: In the cognitive sciences, distributed cognition [41] and outsourced understanding [42] are components for the analysis of explanatory understanding, whereas educational psychology produced insight on the effects of group tutoring [59, 63, 66, 79] and comparisons to one-on-one tutoring [7]. Further, research in the field of computer-supported collaborative work provides approaches for the analysis of collaborative settings [30], group composition [21], and group interactions [81]. Recent research on public AI [95] has emphasized the importance of including stakeholders by introducing deliberation early in the design process [39, 92, 94]. Further, recent work in XAI has 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 enormously from "one-to-one" interactions due to "complexities in group dynamics, cognitive bias amplification, trust issues within the group, and group-centric evaluation" [62]. Drawing from these conceptualizations, this work aims to explore empirically how explanations can support group deliberation by improving understanding and informing decision-making processes.

3 Methods

In this section, we describe our methods and study procedure. We conducted a task-based semi-structured interview study with 43 participants recruited from civil society organizations, a job agency, and the author's extended network (Section 3.5), structured into 8 focus groups (3–5 participants) and 12 single interviews. Participants were presented with the study's employment prediction use case (Section 3.1) and received a collection of explanations about this system (Section 3.2) before solving four tasks and making decisions about the system's deployment (Section 3.3). The study closed with an interview, lasting 90–120 minutes for focus groups and 60 minutes for single interviews. We analyzed the explanations' effect, participants' prioritization of information, and the individual and collective interaction with the explanations (Section 3.4). The university's research ethics committee approved this study.

3.1 Use case: The employment prediction algorithm

The employment prediction algorithm⁴ is a system meant to support counselors of a public employment agency in assessing job-seekers' employability. It uses a logistic regression model trained on historical data to assess how demographic features (e.g., age, education, nationality) influence employment chances, producing short-term and long-term employment scores as recommendations. Counselors use these recommendations to decide about adequate support measures for the job-seeker (courses, stabilizing measures, etc.) but can override the system's recommendations with justification. The system is comparable to algorithmic profiling tools that are used in various countries to assist in job-seeker assessment and resource allocation [53, 65, 74] and which repeatedly highlight the limits of automation in interpersonal interaction [12, 23, 65, 74]. Due to its documentation and exemplification of typical ADM systems' challenges in public employment, the employment prediction algorithm was thus chosen as the use case for this study.

3.2 Explanation design

Our explanation design was composed of 36 question-answer pairs about the employment prediction 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), which were informed by work about the information needs of AI novices [73]. Each category was further divided into subtopics with three levels of increasing detail (base level, level 2, level 3). Every explanation contained a question (e.g., "Who operates the system?") answered with a brief text or image printed on an A5 paper sheet (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 structure of the explanations was intended to i) facilitate uptake by providing information in a modular format and ii) provide different levels of both soundness (information levels) and completeness (subtopics) of information [16, 45, 46]. The selection of subtopics and information levels was guided by work on ADM systems in public employment [2, 65, 74] and research on the prioritization of information by AI novices [73, 87]. The self-directed provision of information allowed participants to select explanations according to their interests [20] and information needs. The question-answer style was motivated by work on question-driven explanations [49] and was intended to encourage engagement. The explanations made use of different techniques that include feature importance, local and global explanations, examples, counterfactuals, and argumentative approaches (Figure 2). Lastly, the explanations were presented in an analog paper format to encourage interactions such as handing around paper sheets, pointing, and reading aloud.

⁴More detailed information on the use case has been omitted to adhere to the anonymization policy but will be re-inserted for the final version.

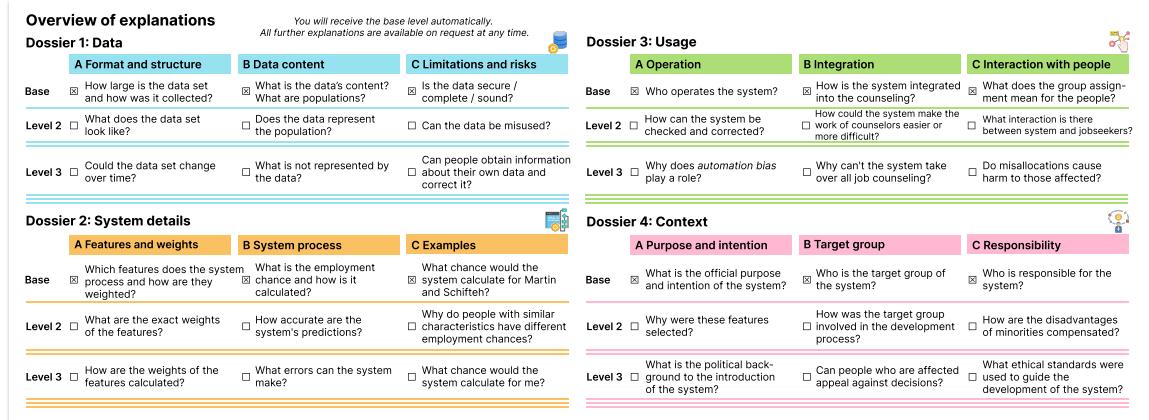


Fig. 1. Overview of explanations. Explanations were designed as a collection of 36 question-answer pairs. The questions were assigned to 4 categories – *data*, *system details*, *usage*, and *context* – each containing 9 questions. Participants received the base explanations at the beginning of the explanation 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 und 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?

Arguments	
Easier	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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
Gender	M	0
	F	-0,14
Age	<30	0
	30-49	-0,13
	50+	-0,7
Nationality	Deployment country	0
	EU	+0,16
	Others	
Education	Compulsory school	0
	Apprenticeship	+0,28
	High School or higher	+0,01
Impaired health	Yes	-0,67
	No	0

Feature	Value	Weight
Duty of care for children or family	Yes	-0,15
	No	0
Professional group	Production	+0,17
	Service	0
Regional labour market	1/2/3/4/5	0/-0,34/-0,18/-0,03/-0,82
Employment history	<75% employed in 4 years	-0,74
	>75% employed in 4 years	0
	>75% employed in 2 years	0
How often job-seeking in 4 years?	0	0/-0,05/+1,19/+1,00
Job-seeking for how long?	Less than 6 months	0
	At least 1 time more than 6 months	-0,8
Participation in support measures	None	0
	1 time supporting	-0,57
	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 single question was printed on a sheet of A5 paper together 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 setup and procedure

The study procedure consisted of several steps that are depicted in Figure 3. This section describes the introduction of the use case, the explanation phase comprising exploration, tasks, and the group decision, and the mechanisms for participants' self-reports on understanding, confidence, and other aspects.

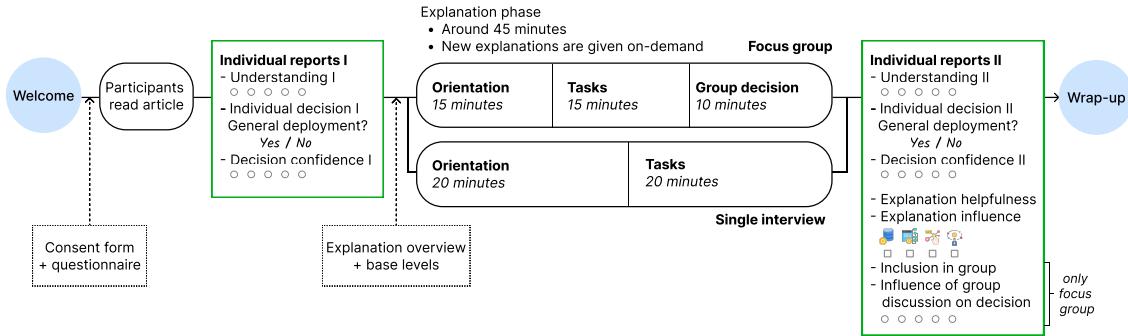


Fig. 3. Overview of the study procedure. Participants filled out consent forms and demographic questionnaires and received a (mock) newspaper article with key information about the ADM system. At this point, participants had to give a first report of their initial understanding, deployment decision, and decision confidence. They then received the explanations, completed four tasks, and groups decided collectively about the system's deployment. Participants then again reported their understanding, individual decision, and decision confidence. Further, they reported the most helpful and influential explanation categories and their perceived inclusion and voice in the group discussion. Interview questions were asked during the second individual report. Finally, participants were invited to share additional comments, were compensated, and thanked for their time. The process for single interviews differed from focus groups in that the collective decision and questions about inclusion and voice in the group were omitted.

Use case introduction. 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.

Exploration phase. After receiving the explanation overview and base level explanations of each category, participants had 15 minutes (focus groups) or 20 minutes (interviews) to become familiar with the structure and mark relevant explanations. Participants could freely decide which explanations to read and request and were encouraged to discuss them.

Tasks. Participants then received a fictional job-seeker case example describing Mr. Harald G.⁵, 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 covering i) the job-seeker's ability to change information stored about him (such as education, physical impairment, etc.), ii) his employability scoring by the employment prediction algorithm, iii) the support measures he would receive, and iv) his ability to contest the decision. All questions had multiple answers with one correct choice and could be answered with information present in the explanations in different categories and detail levels. 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.

⁵Details on the inspiration for this case example where omitted to adhere to anonymization policy, but will be re-inserted for the final version.

Decision task. Lastly, participants in focus groups had 10 minutes to discuss the system's deployment and make a collective yes / no decision. If no consensual decision was reached in time, participants either decided themselves how to resolve the situation (e.g., majority vote) or received the option of stating conditions for the system's deployment.

Self-reports. Participants reported their understanding, decision confidence, and inclusion and voice in the group with 5-point Likert scales, their decision on the system's deployment with yes / no voting ballots, and their choice of the explanation categories most "helpful for their understanding" and "influential to their decisions" by selecting the categories that applied. Self-reports were conducted using a color-coded system to ensure anonymity during focus groups. Each participant was assigned a color and given matching reporting material. During the second self-report, the investigator also asked interview questions about participants' interaction with the explanations, their understanding processes, the most relevant information, and any additional situational questions. In single interviews, the collective decision and questions about inclusion and voice in the group were omitted.

3.4 Analysis

All focus groups and interviews were audio-recorded and transcribed. These transcripts, participants' self-reports during the studies, and the investigators' field notes, provided the data basis for analysis.

For RQ1-Explanations, we focus on how the explanations helped and impacted participants' understanding of the ADM system and examine the differences between individual and group settings. For this, we compare participants' self-reported understanding before and after the explanation phases, their performance in the study tasks, and which and how many explanations were requested. Further, we conduct an inductive thematic analysis [10] of participants' responses to the investigator's interview questions, capturing articulations about the strengths and weaknesses of the explanations' design, the impact of the time limit, understanding barriers, and information perceived as most relevant. While the quantitative items thus serve to characterize the diversity of participants' perceptions [55] and facilitate qualitative exploration [86], they are not intended to invoke "inference [...] of greater generality" [55] nor impose a mental model based on variance theory [67].

For RQ2-Deliberation, we first focus on how participants shared and combined information to solve the study tasks and, second, on how explanations enabled group deliberation. For the first part, we conduct an inductive and deductive thematic analysis of the groups' interaction processes with explanations, connecting our descriptions of these interaction processes to mechanisms for collaborative success and failures as described by Nokes-Malach et al. [66]. For the second part, we compare self-reported decision confidence before and after participants received and discussed the explanations, summarize their feedback about the group setting, and conduct an inductive thematic analysis of focus group conversations in which explanations drove group deliberation and convergence, facilitating the opinion-forming of participants.

3.5 Participants

Recruitment. Table 1+2 provides an overview of the study participants. Participants were recruited through cooperation with civil society organizations, an employment agency, and the authors' extended network. Recruitment criteria included full legal age and no technical knowledge about AI systems, i.e., AI novices. 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 [44]. 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 [35].

Choice of participants. Participants were 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 ADM system is used in, such as job counselors or advisors (groups 2, 3, 7, and 8). 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 5 and 6). All remaining participants are considered as members of the general public and were included to test changes in explanation effects and participants' perceptions (groups 1 and 4). 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 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	DSGVO 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 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, structured according to our research questions: how question-driven, modular explanations support understanding in individual and group settings (RQ1, Section 4.1) and how groups used explanations⁶ to solve the study tasks and form opinions about the system's deployment (RQ2, Section 4.2). Participant labels denote the study setting (focus group: A–H / single interviews: S) and the individual participant, as listed in Tables 1+2.

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

Participants received a collection of 36 explanations to solve the study tasks and decide about the system's deployment individually and (in focus groups) collectively (Section 3). The following subsections describe the explanations' effect on participants' understanding in both settings (4.1.1), design aspects that impeded understanding (4.1.2), the time limit's effect on prioritization of information (4.1.3), the interaction differences between single and group settings (4.1.4), and participants' ratings of explanations' helpfulness and influence (4.1.5).

Table 3. Change in self-reported understanding and decision confidence. Participants indicated their perceived understanding of the system and confidence in their decision on a 5-point Likert scale before and after receiving explanations and discussing them (Section 3.3). This table summarizes the change between their two self-reports. Changes between reports are encoded with symbols and colors, + (increase), - (decrease), and = (no change). Multiple + or - indicate a corresponding change on the 5-point scale.

Focus groups	F2	C3	F1	B4	A2	H1	H2	E1	H4
Increased understanding									
Δ Self-reported understanding	+++	++	++	+	+	+	+	+	+
Δ Decision confidence	-	++	=	+++	+	+	+	=	=
Unchanged understanding									
Δ Self-reported understanding	=	=	=	=	=	=	=	=	=
Δ Decision confidence	+++	++	+	+	+	=	=	=	=
Decreased understanding									
Δ Self-reported understanding	-	-	-	-	-	-	-	-	-
Δ Decision confidence	+	+	=	=	+++	=	+	=	=
Single studies									
S6	S12	S8	S7	S4	S9	S1	S3	S11	S5
Δ Self-reported understanding	+++	+	+	=	=	=	=	=	-
Δ Decision confidence	++	+	=	+++	+	=	-	++	=

4.1.1 Explanations had mixed effects on participants' understanding. Participants reported their understanding before and after receiving and discussing the explanations. Table 3 shows that **self-reported understanding decreased and increased for comparable numbers of participants**, while also often remaining unchanged. Participants gave divergent feedback on how they perceived the explanations, describing their structure and design as "nicely presented" (A2, C2), "good to get an overview" (C3, H4), "well-selected" regarding the questions (G2), "active and controllable" (S8), and as a good primer (F4). Critical comments described them as "too much" (D1, S4), "at first overwhelming" (H5), "confusing" (B1, D1), "theoretical" (D2), and "quite complex" (C2). However, this was not always seen as an issue: "*I believe that it is complex. But I also believe that it needs to be.*" (D2).

Participants gave **nuanced reasons why their perceived understanding decreased**. Several focus group participants stated they did not fully understand the technical explanations, including numbers and a formula (B3, C3,

⁶We describe every question and answer pair as one "explanation" and their entirety as a "collection of explanations" (Section 3).

Table 4. Individual and collective decisions about deployment of the ADM system. Participants were asked for their decision about the deployment of the employment prediction system before and after receiving explanations and discussing them (as described in Section 3.3). Focus groups further made a collective decision about the deployment. Instances in which participants changed their votes from their first to their second report are coloured blue.

Focus groups				Single interviews						
ID	1st decision	Collective decision	2nd decision	ID	1st decision	Collective decision	2nd decision	ID	1st decision	2nd decision
A1	No		Yes	F1	Yes		Yes	S1	Yes	Yes
A2	No		Yes	F2	Yes		Yes	S2	No	No
A3	Yes	Yes		F3	Yes	No	No	S3	No	No
A4	No		Yes	F4	No		No	S4	No	No
B1	No		No	F5	Yes		Yes	S5	Yes	No
B2	Yes		No	G1	Yes		Yes	S6	No	No
B3	Yes	No		G2	No		No	S7	No	No
B4	Yes		No	G3	Yes		No	S8	No	No
C1	No		No	G4	No		No	S9	No	No
C2	No	No		H1	No		No	S10	Yes	Yes
C3	No		Yes	H2	No		No	S11	Yes	No
D1	No		No	H3	No		No	S12	No	No
D2	No	No		H4	No		No			
D3	No		Yes	H5	No		No			
E1	No		No							
E2	No	No								
E3	No		Yes							

G4). In contrast, participants in single interviews with comparable backgrounds met fewer challenges with technical explanations, often stating that they appreciated applying this knowledge when solving the tasks (S2, S8, S11). In these cases, the setting likely mediated the ease of interaction, as participants in single interviews could concentrate without distraction (further described in Section 4.1.4+ 5.2). However, irrespective of the setting, confrontation with mathematical details seemed to re-adjust participants' benchmarks for "high understanding" and **calibrate their perceptions**. Likewise, several participants stated that the different subtopics and levels of detail in the explanations helped them identify the information they were still missing to judge the whole system (B1, F3, G3). Both observations indicate that the explanations revealed participants' blind spots and helped them to overcome the "illusion of explanatory depth" [70] – an overestimation of one's understanding due to limited knowledge and misled intuition – which is exemplified by S2's comment: *"The more I look at the explanations now, the more I realize that I simply assumed a lot of things but didn't actually know how it worked."* In short, **participants gauged their understanding relative to the amount of information available**. That is, since participants only had time to read through parts of all available explanations and were aware that they had not seen all of them, they reported a decrease in perceived understanding when it, in fact, improved. The implications for measuring explanations' effect on understanding are discussed in Section 5.

4.1.2 The explanation design was not equally suited to every participant group. In contrast to participants whose perceived understanding was improved and calibrated by the explanations, we observed that the explanations did not fully support understanding or interaction for some groups. In these cases, the **explanations' language and textual format were listed as impeding factors**. The language style was described discrepantly as "well understandable" (P3, E3, H1) but also "very difficult" (E2, F2) and "too technical" (B3, B4) depending on participants' background. Crucially, in Groups E and F (decision subjects), participants stated that the language posed challenges. E2 compared the explanations to official letters they receive: *"I get letters that I don't understand from the court. I understand every single word, but I don't understand the context."* In line with this, job counselors in Groups C, G, and H stated that explanations in simple language would be required if they were supposed to be useful for their job-seeking clients. Similar comments were

made about the explanations' heavy reliance on textual information, described as demanding and requiring plenty of concentration. While previous work found that textual explanations can be effective despite users' dispreference [82] or are even preferred [87], a more interactive and visual design would likely facilitate interaction: "*If there's a program where, when you click on it, it opens as a window and you can read about it, that would be great, and you can deal with it at your own pace.*" (H5) Both, simpler language options and an interactive version, are thus useful suggestions for improving the explanation design to ensure that explanations are helpful for a broader range of stakeholder groups. Further, Section 4.2.1 describes how these understanding challenges impacted group interaction.

4.1.3 Time limits led to forced prioritization of information, with different effects on participants. All participants had 40 minutes to become familiar with the explanation structure, request new explanations, solve the given tasks, and, in focus groups, make a collective decision (Table 4). The time limit was meant to incentivize participants to prioritize information and, where possible, work together to process larger amounts. This succeeded for most focus group's participants, who stated that "*when you have little time, the teamwork is extremely good*" (B2) and that "*even a brief exchange is enough, otherwise it becomes too much*" (C3). However, the time limit also led to induced stress, especially in single interviews (S2, S6, S7, S8, S9, S11). S9 commented that this reminded them of their work as a social center manager: "*Basically, the time to read about something new does not exist.*" (S9). This points to the need for brief and compact synopses complementing the detailed explanations. Regarding prioritization, many **participants liked choosing explanations according to their own interest**, which gave the feeling of being in control (S8), provided a direction for exploration (H1), and justified skipping explanation categories in favor of others (S6, S9). Allowing for the self-directed selection of relevant information by users could thus be beneficial to encourage interaction and engagement.

4.1.4 Some single participants worked through surprising amounts of explanations. Compared to focus groups, participants in single interviews often requested and engaged with equal or even higher amounts of explanations (as depicted in Figure 4). Often, participants in single interviews requested "the whole context" (S3) or "all of level 3" (S4) of one category, appearing to deal with the amount of information easily. A likely reason for this is that **single interviews supported more focused and in-depth interaction with explanations compared to group settings**. Participants in single interviews also performed better in the study tasks overall and, as mentioned above, often calculated the job-seeker's precise employment chance in task 2, which required following a multi-step process. This is noteworthy, as *none* of the groups, aside from the pilot groups, attempted to calculate the precise chance. This cannot be explained by a difference in educational or professional background, as even focus groups with university graduates in domain expert professions did not perform these calculations, again indicating that the reason might be the difference in supported interaction. For example, the parallel processing of information might have impeded understanding of the system's entirety in groups, as described by G3: "*You only pick out the most important things. And then you haven't gathered information about the thing as a whole.*" This highlights the benefits and trade-offs of both settings, suggesting that individual engagement with explanations might be helpful *before* discussing them in a group setting. The implications of this are discussed further in Section 5

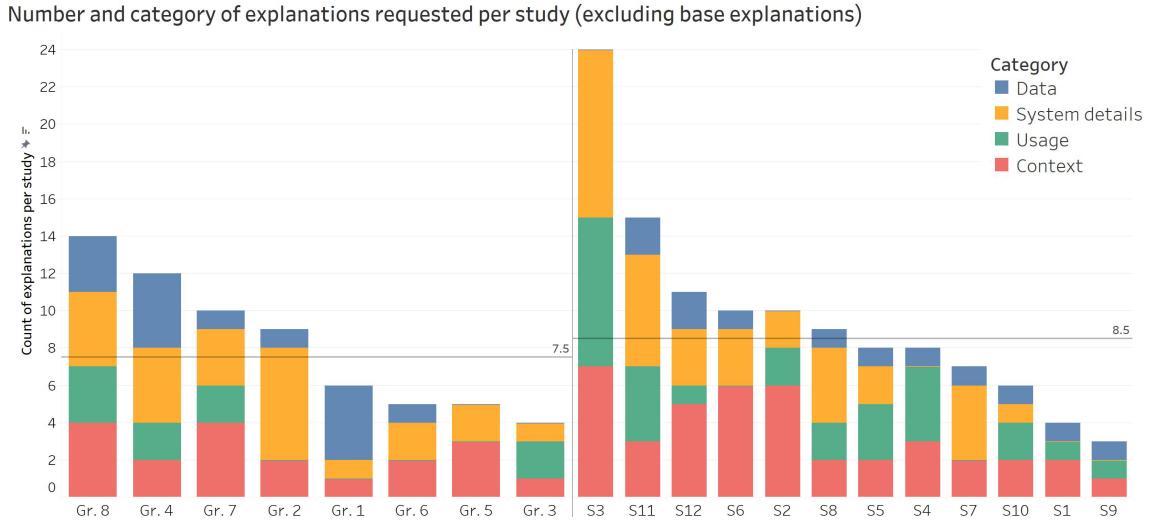


Fig. 4. Explanations requested per study and setting. Participants could request up to 36 explanations from the study investigator during the explanation phase. The left side shows explanations requested by focus groups, the right side by participants in single interviews. The horizontal lines indicate 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.

4.1.5 All explanation categories were deemed helpful and influential, but system details and context were most often mentioned. In the second self-reports, participants were asked which explanation category was most helpful for their understanding, most influential for their decisions, and whether specific explanations stood out. Regarding understanding, participants in focus groups responded with a rather even distribution over all categories, as depicted in Figure 5, and frequently mentioned that "*all of them [are relevant]... I don't think you can leave anything out, really*" (D3). Both participants in focus groups and single interviews further stated that *system details* were crucial, as they were perceived as "tangible" (S6), "concrete" (S8), and to be "the devil in the detail" (B4). Specifically, participants found explanations about the features and weighting most important (D1, G1, H3, S5), which G4 commented with: "*That is the central point, the basis of the whole system.*" Even though *context* was less often reported as an influential explanation category, participants requested these explanations most often (Figure 6) and frequently mentioned affected persons' inability to contest decisions (C1, D1, E1) and the system's political background (A1, S3) as relevant – both belonging to *context*. Notably, participants in single interviews found *data* much less helpful and less influential, stating, e.g., that GDPR would likely cover the important aspects of it (S8) or that they prioritized another category in the time available: "*I didn't deal with it in such detail [...]. You read it [the base level] and already feel informed. As is often the case with data protection, 'yeah, close enough.'*" (S6). Other participants stated that the technical details were not too important for them but that they instead focused on how the system was used practically and which values it represented (S3, S10), corresponding to the categories *usage* and *context*.

Most helpful and most influential explanations in each setting

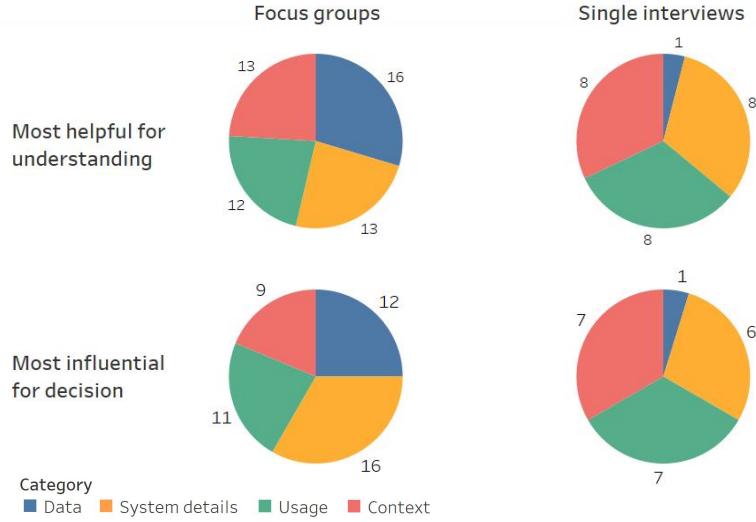


Fig. 5. Most helpful explanation categories for understanding and most influential categories for participants' decisions. Participants could select any number of explanation categories for both questions, including none and all four.

Number and category of individual explanations requested (excluding base explanations)

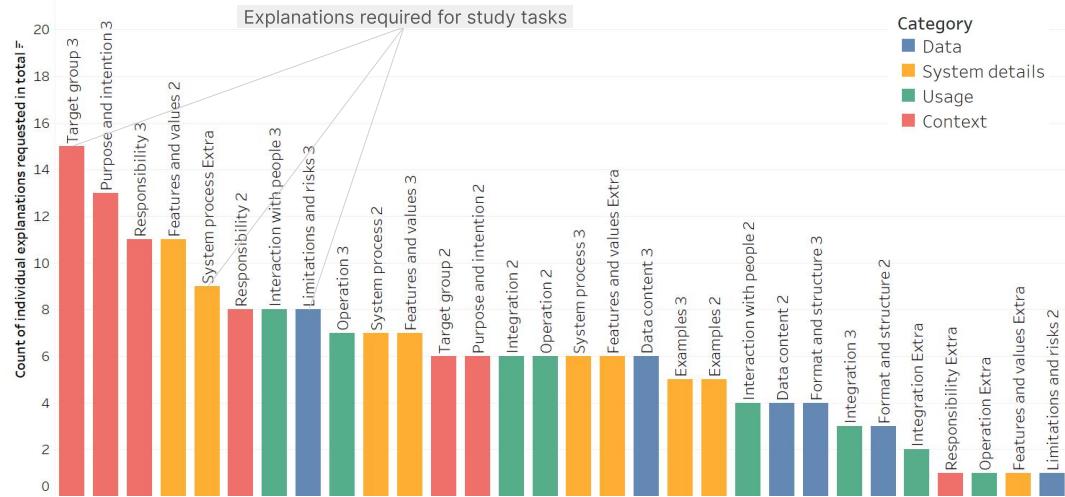


Fig. 6. Individual explanations requested overall. Bars show how often specific explanations were requested overall, labeled with subtopic and level. The four most requested explanations, aside from those requested for study tasks (Section 3.3), are of the category *context* (left to right: political background, ethical standards, support of minorities, involvement of target group), followed by explanations about *usage* (LTR: harm of misallocations, automation bias, checking and correcting, influence on work of counselors). Explanations about *data* were less requested in comparison.

4.2 RQ2-Deliberation: How do the explanations support group interaction with information and deliberation processes?

In this section, we describe how our explanations supported group collaboration in exchanging information (Section 4.2.1) and in discussing and forming opinions about the system's deployment (Section 4.2.2).

4.2.1 Explanations support collaborative success if group dynamic is constructive. We identify several processes that groups used to interact with information collectively. Our descriptions of these interaction processes include splitting (by category, order, priority), sharing (swapping, reading out loud, proposing solutions), analyzing (discussing, explaining), and combining (merging, locating, correcting). The degree to which groups used these processes appeared to depend on the familiarity between group members, their domain expertise, and their understanding. In the following, we analyze three exemplary conversation excerpts that showcase different forms of interaction in focus groups while also demonstrating the diversity in collective interaction. For each excerpt, we describe the interaction processes that emerged and then connect them to mechanisms that drive collaborative failure or success (marked by quotation marks) as outlined by Nokes-Malach et al. [66].

Case 1: Information sharing leads to increased engagement and construction of common ground. Group G was composed of participants who hold executive positions in a civil society organization and are responsible for the management and training of employees. Participants in this group showed high engagement with information, frequently sharing explanations and exchanging their perspectives by tapping on their professional experience. The excerpt begins several minutes after participants received the explanations just as G4 examines an explanation comparing the algorithmic score for two job-seekers:

G4: *That's bad, the two of them. Look, I've got this [system details-C-base], "What chances would the system calculate for Martin and Schifteh?"* (sharing: reading out loud)

G2: *Schifteh is probably worse off, isn't she?* (analyzing: discussing)

G4: *Schifteh has 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.* (sharing: reading out loud; analyzing: explaining) [...]

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...* (analyzing: discussing)

G4: *She gets two minuses for living in [a specific city district⁷]. People who live in [this district], or in a district that is not well regarded, have totally negative...* (analyzing: explaining)

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

Framed in the context of mechanisms that drive collaborative success, G4 first draws the group's attention to an explanation they deem relevant, describing the algorithm's treatment of two different cases of job-seekers. G4 reads the information to the group, which prompts G2 and G3 to contribute their own perspectives, "increasing engagement" while also "cueing their prior knowledge". As a result, G1 is also engaged by the information and articulates interest in looking at it. In combination, the interaction has established "common ground" by sharing information the group can use for further collaboration. Interactions like these were common in Groups G and H (job counselors and trainers), indicating that the explanations' format and content suited these compositions.

⁷Information was omitted to adhere to anonymization policy and will be re-inserted for the final version.

Case 2: Discussion of solutions leads to correction. Group B consisted of both staff members and volunteers of a civil society organization. The group interaction was mostly driven by one participant who requested and shared most of the information and thus acted as a guide and informer, potentially driven by her role as deputy manager. While this led to a more one-sided dynamic, the other participants later appreciated outsourcing part of the understanding process to B3. In the excerpt below, B3 proposes a solution for task 2 (chance of employment that the system assigns Harald G.; group low, medium, or high) that is then discussed:

- B3: *I chose "group low" for task 2. (sharing: proposing solutions)*
- B4: *I put him [Harald G.] in "group medium". (sharing: proposing solutions)*
- B1: *Why not medium, B3? (analyzing: discussing) [...]*
- B3: *With 49 years he has one minus. And the knee problems could be a "health impairment". (analyzing: explaining)*
- B2: *But if he has a job where he sits, that doesn't matter. (analyzing: discussing)*
- B3: *But I don't see that this is considered anywhere. And I think the fact that he hasn't been employed much in the last 4 years has a very negative effect. What information do you have that speaks against that? (combining: correcting; analyzing: discussing)*
- B1: *Because with 50, he would still be in "group medium". [...] Then he's male and from [this country⁸]. (analyzing: explaining)*
- B3: *Male is only 0, the citizenship is also only 0. That doesn't make up for anything. (combining: correcting) [B1 agrees] [...]*
- B4: *Then I'll correct it to low again. (combining: merging)*

The group members in this situation "negotiate multiple perspectives" on the correct solution, notice discrepancies in their "common ground," and "re-expose" themselves to information to "correct errors" in their knowledge. B3 led the interaction and corrected other participants repeatedly, which could have caused adverse reactions from them (e.g., due to being "evaluated"). However, B2 and B4 in fact later articulated appreciation for this interaction, stating that "*B3 filtered that out well*" (B2) and that "*it was helpful that you [B3] always read it out and corrected it straight away*" (B4). Correction was therefore seen as constructive to the discussion, possibly because of B3's leading role in the organization and because participants already knew each other.

Case 3: Attempted error-correction fails to create common ground. Group E was composed of people who were job-seeking at the time of the interview or in the past. Members of this group were not familiar with each other. Interaction was mostly fragmented and participants rarely shared or discussed information. In the excerpt below, participants were asked about their responses to task 4 (whether Harald G. is able to contest the system's decision):

- E2: *I say you can always object. Even if you're locked up, you can go to court somewhere. (sharing: proposing solutions)*
- E3: *Yes, I think so too. (sharing: proposing solutions)*
- E1: *Well, I've already read that [the explanation], and it says that you can't appeal against it. But it says that you could if a legal basis were first created, which doesn't exist at the moment. (combining: correcting; analyzing: explaining)*
- Investigator: *Okay. Is there anything you [E2, E3] would like to change now, based on this information? [E2 & E3 indicate that they do not]*

⁸Information was omitted to adhere to anonymization policy and will be re-inserted for the final version.

Investigator: *You would say that he can contest the decision?*
 [E2 & E3 agree]

Despite E1's reference to an explanation and their attempt to "correct errors" in what E2 and E3 believe to be the solution, the interaction fails to establish "common ground" or "negotiate multiple perspectives", as the different viewpoints are neither elaborated nor discussed collectively. Possibly, E2 and E3 did not believe that E1 or the explanation were correct, or they felt embarrassed for being corrected. Further, E1 had no particular motivation to persuade the other participants of the correct solution, as E1 later self-described as "introverted" and stated that they usually preferred individual over group settings. Comparable mechanisms unfolded only in Group F, where participants tended to read explanations themselves or disengage rather than interact with other participants. This might have had multiple reasons: First, both Groups E and F indicated a low interest in the topic of labor; second, participants in both groups had trouble understanding the explanations due to language barriers and the heavy reliance on text; and third, participants in both groups were unknown to each other. In combination, this might have created a group dynamic that was not constructive, i.e., neither argumentative [56], solution-oriented [64], or familiar [38], which disrupted collaborative interaction with the explanations as well as discussion processes.

Summary. The conversation excerpts in cases 1 and 2 demonstrate that the explanation design supported various collaborative interaction processes. These processes led participants to create common ground and find a task solution through error correction, which are markers of "collaborative success" [66]. In contrast, case 3 demonstrates that the explanations did not always achieve their purpose of facilitating interaction, highlighting that the interplay of explanation format and group dynamics must be considered in the design of explanations in group settings.

4.2.2 Explanations in group settings allow for informed discussion of pivotal topics and opinion formation about system deployment. For most participants, deciding about the employment prediction algorithm's deployment was a clear choice: 7 out of 8 groups voted "No", and a majority of participants also voted "No" in their individual decisions (Table 4). Confidence in these decisions increased strongly after the explanation phase (Table 3), with most participants indicating that they now felt better informed. Comparing single interviews with focus groups, two effects were notable: First, a third of focus group participants changed their vote in the second self-report, and second, participants in single interviews stated repeatedly that they would have found a discussion helpful.

This is further underscored by participants' feedback on the focus groups: Most felt included in the discussions (26 of 31) and were influenced in their decisions by them (18 of 31). Specifically, participants appreciated the group size, which allowed for balanced contributions and productive discussion (C1, D1, E2), and valued the exchange of perspectives: "*Everyone understands a little differently and sees a different focus, I think it's really important to talk about that.*" (H5). We see this as evidence that the explanations were suitable for the collaborative setting and supported group deliberation. This section thus aims to describe the synergy effects that emerge when combining explanations with group discussion and how this affected participants' perceptions. In the following, we describe three conversation excerpts in which explanations lead groups to discuss their core concerns and beliefs.

Case 1: Group discusses changes of perspectives on the system's deployment. 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. In the excerpt below, B2 gives their perspective on the system's deployment, and B3 responds with theirs, referring to explanations about the system's features and weightings (*System details, Base, B + Level 2, C*), which fuel the discussion that follows:

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, the apprenticeship of our Harald [the example job-seeker] was rated positively. He can't use the apprenticeship at all for retraining. These things are given far too little consideration. [...] 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. [...]*

B1: *What I'm wondering is, what would be the real benefit of introducing the system? What could be found out with much less effort? After all, employees should be able to see which degree you have quite quickly.*

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 provided information changed B3's attitude toward the system and how this change, in turn, affected the group discussion. The combination of explanations and group discussion here had a major effect on participants' perceptions, as three participants in this focus group changed their vote from "yes" in the first self-report to "no" in the second. Interactions like B2's appreciation of B3's arguments and B1's questions about the system's benefits depict processes of constructive group deliberation directly supported by the explanations.

Case 2: Group debates the meaning of "unalterable" facts. Group D was composed of participants who were neither job-seekers nor domain experts and thus would be affected only indirectly by the system's deployment. Even so, when D1 referred to an explanation describing how two different job-seekers would be judged by the system (*System details*, Base, C), it led to a discussion about the objectivity and adequacy of using data to represent a person:

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 the "basis", I think there are cracks in this basis. And I'm afraid of these cracks, 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.*

In this excerpt, the discussion surfaces discrepancies in how participants perceived the factuality and value of data. While the discussion in this focus group did not lead to a consensual decision in the time given, we still see this exchange

as an important process in the deliberation on public AI systems. The merit of this discussion was later acknowledged by D1, who found the explanations confusing but stated that these exchanges were the "centerpiece" and the most intriguing part. This highlights the role of explanations in the surfacing and debating of divergent opinions.

Case 3: Group discusses consequences of the system's deployment for affected people. Group H was composed of staff members of a civil society organization specializing in helping young adults to take up apprenticeships. While all participants were clearly against the system's deployment, the group decision task still led to a discussion about why specifically they opposed the system. The excerpt begins with H2's comment about an explanation describing that job-seekers do not have a right of contestation (*Context, Level 3, B*):

H2: *Just the fact that you can't contest the decision, that's a no-go.*

H1: *I think that having such systems introduced by institutions that themselves have very little idea of the real work – they may even want to help and support work systems, but that completely misses the mark because they are making decisions that are again driven by economic interests.*

H4: *And the life of the people who are then affected by the system is not represented at all, the psychosocial situation of the people is not taken into account.*

H3: *It is if they are depressed...*

H4: *Then it's minus, minus, exactly. It simply deprives people of the opportunities they deserve. And that's probably also my personal experience, I've worked with long-term unemployed adults who also had physical or mental impairments. The problem is that the longer I'm out of the job market, the fewer opportunities I have and people are already in a bad enough situation. And then you're totally stuck in this [algorithmic] system. I just think that's unfair.*

H2: *In terms of labor market policy, it may be justified, but in terms of social policy, it's totally wrong.*

The excerpt shows that even participants with unanimous attitudes toward the system found value in discussing and sharing their reasons for disapproving its deployment. H4 in particular appeared to be incensed by the system's representation of soft factors. Their account of working with long-term job-seekers was later appreciated by the other participants, who stated that this perspective was valuable to them. The explanations in this case served to facilitate group conversation about the meaning and consequences of the system's deployment and encouraged participants to share their views, stances, and experiences.

Summary. While Section 4.2.1 describes that explanations were effective in supporting group collaboration, this Section shows that their role in providing information for discussions between participants is equally important. Explanations helped participants in debating topics that are pivotal to the acceptance, trust, and application of ADM systems, in effect supporting their collaborative reasoning [59]. In this way, the explanations support what Kawakami et al. [39] describe as "deliberations around *whether* and *under what conditions* to move forward with developing or deploying" a public sector AI tool. This deliberation is important, as it allows participants to express and reflect on their opinions, feel heard, and consider other perspectives when making decisions.

5 Discussion

In this section, we discuss the implications of our findings for the design of explanations for AI novices (Section 5.1) as well as suggestions for their applications in collaborative settings (Section 5.2), which are summarized in Figure 7.

5.1 Reflections on the explanation design

Explanations improved or calibrated participants' understanding, but measurements should be reconsidered. In Section 4.1.1, we described that while participants gave discrepant responses in their understanding self-reports (Table 3), they rarely felt that their understanding truly decreased or remained the same. Instead, we argue that the explanations changed their frame of reference for "understanding" by revealing blind spots and making participants realize that their initial understanding was worse than assumed. Drawing from definitions in the cognitive sciences, the explanations thus supported participants in developing understanding by i) making participants realize that their initial understanding was incomplete [70] and ii) supporting them in locating and closing relevant understanding gaps [42]. The effect of participants reporting decreased perceived understanding following the explanations can be attributed to the fact that these explanations led them to attain a "partial understanding" [40], but also showed them the difference to a complete, comprehensive understanding. The difficulty of self-assessing one's understanding is a known issue [70] and previous studies in XAI have shown similar effects caused by white-box explanations [18], which, due to their high information density, led to increases in objective understanding but decreases in self-perceived understanding. This is not necessarily an issue regarding participants' understanding, and together with the substantial increases in decision confidence, it can be seen as evidence that our explanations effectively improved and calibrated participants' understanding. However, it also demonstrates that the relativity of internal cognitive states should be accounted for when measuring understanding. For this reason, recent work in XAI has proposed measurement based on abilities [80]. This approach appears promising, considering that our single interview participants often found applying their knowledge in the study tasks helpful. Still, it likely requires modification when used in collective settings, as, for example, participants in focus groups never calculated the precise employment chance in task 2. The operationalization of the abilities-based approach for both individual and group settings is thus a direction for future research.

Participants considered all four explanation categories relevant and appreciated technical and contextual information. Our explanation design covered the categories *data*, *system details*, *usage*, and *context*. In the second self-reports, participants indicated which of these categories were "helpful for their understanding" and "influential to their decisions" (Section 3.3). In Section 4.1.5, we describe that focus groups reported that all categories helped their understanding but that *system details* more often than other categories influenced their decisions. In contrast, participants in single interviews reported all categories equally often as helpful and influential except *data*. Possible explanations for this effect include the forced prioritization due to the time limit, which led participants in single interviews to skip information about data as it was believed to be less relevant. In comparison, some focus groups proceeded chronologically through the explanations and thus spent more time on *data* as the first category. Further, the findings highlight that participants welcomed non-technical information about the system's intention, ethical values, and practical usage (Figure 6) but also assigned high importance to the model's exact feature weights. We see this as confirmation that participants' information needs cover the entirety of an ADM system and that explanations for AI novices should aim to provide both technical (*data*, *system details*) and contextual (*usage*, *context*) information.

Explanations for decision subjects require a co-design approach. In Section 4.1.2, we describe that our explanations were most effective for participants who did not face challenges with either the language, large amounts of

text, or the explanations' structure. While these attributes are not entirely congruent with participants' education levels, we observe that university-educated participants faced comparatively little difficulty interacting with the explanations, as the conversation excerpts in Sections 4.2.1+4.2.2 show. The differences were most apparent in Groups E and F, where participants stated that the explanations did not help them fully grasp the system as they required a high language and reading proficiency. We see this as evidence that a co-design approach similar to Weitz et al. [87]'s work is required to build explanations for groups composed like Groups E and F, as it might be difficult to identify which content and format of explanations are required by decision subjects. On the other hand, the findings demonstrate the explanation design's suitability for groups of domain experts, who also stated that this format of conveying information would be useful as a training measure.

Suggestions for improving the interaction design and combining individual and collective settings. In Sections 4.1.1 and 4.1.2, we described that according to participants' feedback, the explanation design offers comprehensive and flexible information selection and encourages self-directed and active exploration, but that it also has a high access threshold and requires adjustment to the subdivided structure. For focus group participants, this design resulted in participants splitting and swapping explanations, sharing relevant parts, and discussing surprising information, which overall resembled an associative approach that supported group interaction but also led to a more fragmented picture of the system. In single interviews, participants could select and interact with information at their own pace and in a focused manner but, on the other hand, missed exchange and external perspectives. These effects can be described in the context of collaborative success and failure [66], which states that groups require extra cognitive effort to coordinate and manage interactions, which can lead to memory limitations, information retrieval disruption, and blocking of thought processes if these coordination costs are too high. In contrast, individuals do not incur these cognitive costs, but of course can also not benefit from collaborative success mechanisms such as increased working memory or complementary knowledge. A future iteration of the explanation design could thus follow an individual with a collective interaction phase to combine the advantages and offset limitations. Further, a digital version would facilitate overview (summaries) and navigation (menus) while allowing for simple language options and cross-references. Lastly, the essential information of each explanation category, which in our design was contained in the base levels (Section 3.2), could use a balance of textual and visual design to ensure that essential information does not rely only on textual understanding.

On a side note, we 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 [46], 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 background rather than essential information gradually. 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 ADM systems, such that this structure allows for exploration while not obscuring essential or complex information.

5.2 Benefits and challenges of using explanations in collaborative settings

In Sections 4.2.1+4.2.2, we described that focus groups' interactions with information led to "collaborative success" [66] when the group dynamic was constructive, i.e., argumentative [56], familiar [38], and solution-oriented [64], and that explanations can inform conversations about central topics in public AI deployment [39]. In the following, we discuss these findings by outlining the resulting benefits and challenges for collaborative XAI.

Benefits: Explanations support collaborative reasoning, problem-solving, and deliberation. In the best cases, focus groups in our study had an open, communicative, and investigative atmosphere that facilitated the up-taking, sharing, and discussion of information. In these settings, the explanations showed their strengths by allowing for the distribution of reading 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 collaboration and informed decision-making around deploying a public AI system [39] by creating common ground that participants used as a basis for deliberation. This deliberation process is the differentiating factor compared to "one-to-one" [60] explanation settings, which might allow for more concentrated learning [7], but miss the exchange of knowledge and perspectives that could become central for societal discourse about public AI [95]. While this collaborative atmosphere has several preconditions, such as containing cognitive biases (groupthink [37], equality bias [62]) and requiring trust between group members [38], our findings suggest that explanations can be an effective enabler in supporting these crucial group interactions. Regarding learning and understanding, XAI could further benefit from connecting to research showing that small groups can exhibit collective rationality [59], outperform the wisdom of the crowds [63], and solve tasks correctly even if no group member knows the solution beforehand [79].

Challenges: Explanations must adapt to different groups and require constructive dynamics. While our explanations were effective in supporting interaction and opinion formation in most focus groups, we also observed several challenges resulting from **group compositions and trust between participants**. While the explanations typically worked well in groups of domain experts, especially G (executive staff) and H (job trainers and counselors), special interaction situations occurred in Groups B and H due to two participants' leading roles in the organizations. This could have resulted in disrupted interaction due to fear of being evaluated by someone perceived as superordinate, but, in fact, had no negative impact on collaboration. In Groups E and F (decision subjects), however, the unfamiliarity of participants and the explanation's high access threshold led to a lack of interactions and discussion, indicating that the explanation did not fully support these compositions. In these groups, **a lack of trust between participants** likely disrupted interactions when participants were unknown to each other, amplifying effects such as the fear of being evaluated [66]. This underscores the importance of creating trust between group members in collaborative XAI settings [38]. 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 [85], or creating an atmosphere that allows participants to interact in roles (e.g., proponents and opposition) to facilitate discussion without revealing personal opinions. Future work should examine how such measures can be incorporated into explanation design to support interaction in groups of comparable compositions. Lastly, and surprisingly, we did not observe many **cognitive biases** such as "groupthink" [37] (aiming for consensus at the cost of negotiation and alternatives) or the "equality bias" [62] (downplaying one's expertise to weigh everyone's opinion) in our study. In one example of groupthink, participants in Group A changed their vote to "Yes" to reach a group decision, perhaps driven by their (stated) lack of being personally affected 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 also not being affected personally. 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" [56], and to explain the system in a way that makes it more personally relevant to participants [90], e.g., by emphasizing connections to their own experiences.

5.3 Explanation design suggestions

We summarize the implications of our findings in the form of suggestions for the design of explanations that are suited to AI novices in individual and group settings, as depicted in Figure 7.

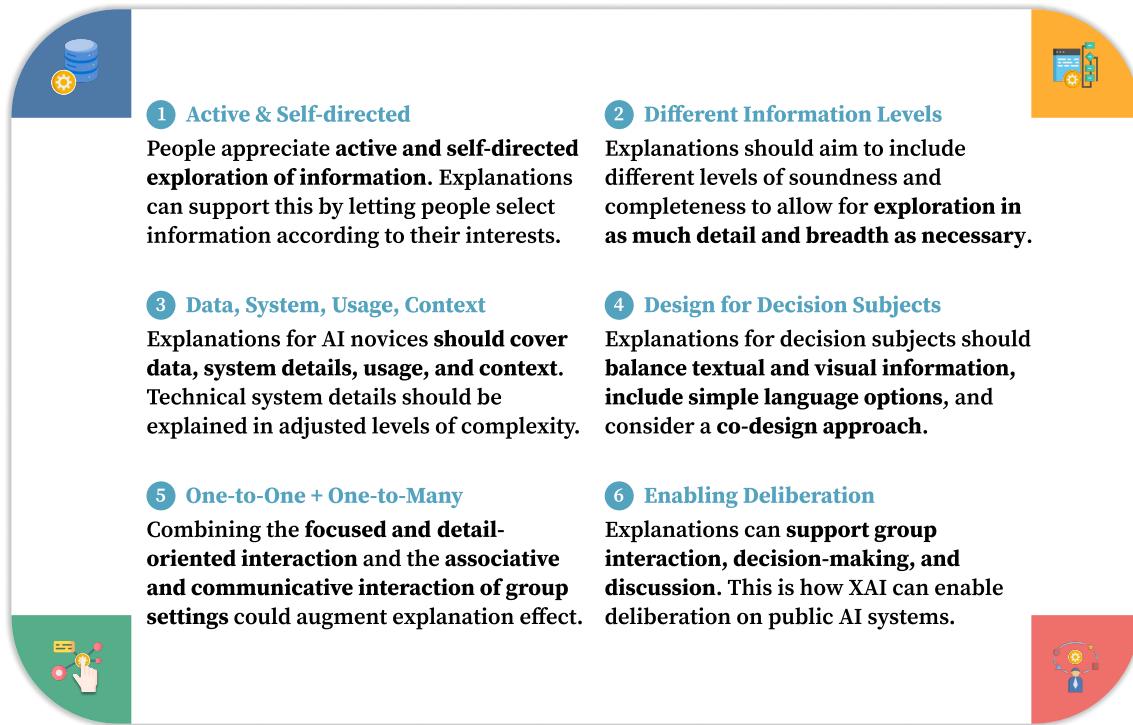


Fig. 7. An overview of how our findings inform the design of explanations for individual and collaborative settings.

6 Limitations

In this section, we briefly describe limitations of our study that the reader should consider. 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 embedded in a specific sociotechnical context [27] 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 has a bias toward university education in the single interviews, which we addressed by comparing these participants mostly with university-educated participants in the focus groups.

7 Conclusion

In this paper, we proposed and empirically validated a question-driven, modular explanation design for individual and collaborative settings. We conducted an interview study with AI novices in 8 focus groups and 12 single interviews, analyzing the explanations' effect on understanding and decision confidence, participants' perceptions of key information, and the interaction processes in both settings. We found that explanations were effective at supporting and calibrating participants' understanding and at informing individual and collective decision-making. Our findings further showed that our explanations support focused and detail-oriented interaction in individual settings while also being suited to be shared and discussed in group settings. In groups with constructive dynamics, interactions with the explanations led to "collaborative success" [66] and informed group discussions about pivotal topics regarding the system's deployment. For groups in which the explanations were less effective, we suggest adjusting the group dynamic, modifying the explanation's format, and using co-design approaches to better adapt explanations to various needs. Based on these findings, we present suggestions for the design of explanations for AI novices and outline how their application can support collaborative reasoning and discussion. With this work, we aim to contribute an explanation design approach to the field of collaborative XAI and present implications for how explanations can support stakeholder deliberation 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.oaw.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] 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.
- [6] 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>
- [7] Benjamin Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. (1984).
- [8] 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>
- [9] 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>
- [10] 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>
- [11] 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>

- [12] Bundesagentur für Arbeit. 2021. Bearbeiten von Bewerberdaten durch Träger. arbeitsagentur.de/datei/dok_ba013193.pdf.
- [13] Alyxander Burns, Christiana Lee, Ria Chawla, Evan Peck, and Narges Mahyar. 2023. Who Do We Mean When We Talk About Visualization Novices?. 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 819, 16 pages. <https://doi.org/10.1145/3544548.3581524>
- [14] 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>
- [15] 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>
- [16] Mohamed Amine Chatti, Mouadh Guesmi, Laura Vorgerd, Thao Ngo, Shoeb 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>
- [17] 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>
- [18] 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>
- [19] 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>
- [20] 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>
- [21] 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>
- [22] 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>
- [23] 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>
- [24] 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>
- [25] 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>
- [26] 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].
- [27] Upol Ehsan, Koustuv Saha, Mummun 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>
- [28] 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>
- [29] European Commission. 2021. Laying Down Harmonised Rules on Artificial Intelligence. digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.
- [30] 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>
- [31] 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>
- [32] 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
- [33] 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>
- [34] Mouadh Guesmi, Mohamed Amine Chatti, Shoeb Joarder, Qurat Ul Ain, Rawaa Alat rash, 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>

- [35] 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>
- [36] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. 2023. Explainable AI: roles and stakeholders, desires and challenges. *Frontiers in Computer Science* 5 (Aug. 2023), 1117848. <https://doi.org/10.3389/fcomp.2023.1117848>
- [37] Irving L Janis. 1972. Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. , viii, 277–viii, 277 pages.
- [38] 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
- [39] 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>
- [40] 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>
- [41] 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)
- [42] 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>
- [43] Max F. Kramer, Jana Schaich 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>
- [44] Richard A Krueger. 2004. *Focus groups : a practical guide for applied research* (3. ed., 6. print. ed.). Sage, Thousand Oaks, Calif. [u.a.].
- [45] 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>
- [46] 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>
- [47] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (July 2021), 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- [48] 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>
- [49] 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>
- [50] 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>
- [51] 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>
- [52] 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>
- [53] 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>
- [54] 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.
- [55] Joseph A. Maxwell. 2010. Using Numbers in Qualitative Research. *Qualitative Inquiry* 16, 6 (2010), 475–482. <https://doi.org/10.1177/1077800410364740>
- [56] 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>
- [57] 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>
- [58] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). <https://christophm.github.io/interpretable-ml-book>
- [59] 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>

- [60] 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>
- [61] 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.
- [62] 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/>
- [63] 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>
- [64] 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>
- [65] Jędrzej 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.
- [66] 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>
- [67] Michael Quinn Patton. 1990. *Qualitative evaluation and research methods*, 2nd ed. Sage Publications, Inc, Thousand Oaks, CA, US. 532–532 pages.
- [68] 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>
- [69] 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>
- [70] 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
- [71] 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>
- [72] 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>
- [73] 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 \[cs.HC\]](https://arxiv.org/abs/2401.13324)
- [74] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobel, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). ACM, Seoul Republic of Korea, 2138–2148. <https://doi.org/10.1145/3531146.3534631>
- [75] Hong Shen, Haojian Jin, Angel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. <https://doi.org/10.1145/3415224>
- [76] 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>
- [77] Ben Schneiderman. 2022. *Human-centered AI*. Oxford University Press, Oxford.
- [78] 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>
- [79] 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>
- [80] 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>
- [81] 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>
- [82] 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>
- [83] 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>

- [84] 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>
- [85] P. C. Wason. 1968. Reasoning about a Rule. *Quarterly Journal of Experimental Psychology* 20, 3 (1968), 273–281. <https://doi.org/10.1080/14640746808400161>
- [86] 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.
- [87] 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>
- [88] 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>
- [89] 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>
- [90] Grant P. Wiggins and Jay McTighe. 2005. *Understanding by design* (expanded 2nd ed.). Association for Supervision and Curriculum Development, Alexandria, VA.
- [91] 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>
- [92] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAccT '22). Association for Computing Machinery, New York, NY, USA, 535–563. <https://doi.org/10.1145/3531146.3533118>
- [93] 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>
- [94] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–32. <https://doi.org/10.1145/3579601>
- [95] 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>

Appendix

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.

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.⁹

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	Nominal values
Gender	Male/Female
Age group	0–29/30–49/50+
Citizenship	[Deployment country]/EU/Non-EU
Highest level of education	Grade school/apprenticeship, vocational school/high- or secondary school, university
Health impairment	Yes/No
Obligations of care (only women)	Yes/No
Occupational group	Production sector/service sector
Regional labor market	Five categories for employment prospects in assigned job center
Prior occupational career	[Variables as described below]
Days of gainful employment within 4 years	<75%/≥75%
Cases within four 1-year intervals	0 cases/1 case/min. 1 case in 2 intervals/min. 1 case in 3 or 4 intervals
Cases with a duration longer than 180 days	0 cases/min. 1 case
Measures claimed	0/min. 1 supportive/min. 1 educational/min. 1 subsidized employment

⁹More detailed information on the deployment context of the system has been omitted to adhere to the anonymization policy but will be re-inserted for the final version.

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. 8. 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																	
<p>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.</p> <p> Harald G.</p> <table border="1" style="margin-left: 20px;"> <tr> <td>Age:</td> <td>49</td> </tr> <tr> <td>Gender:</td> <td>Male</td> </tr> <tr> <td>Education:</td> <td>Apprenticeship</td> </tr> <tr> <td>Citizenship:</td> <td>[Deployment country]</td> </tr> <tr> <td>Obligations of care:</td> <td>No</td> </tr> <tr> <td>Occupational group:</td> <td>Service</td> </tr> <tr> <td>Employment history:</td> <td>Less than 75% in last four years</td> </tr> <tr> <td>Health impairment:</td> <td>Knee problems</td> </tr> </table> <p style="text-align: center;"><i>All characteristics not specified have the value 0!</i></p>		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
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																
<ol style="list-style-type: none"> Can Harald change the data stored about him (e.g. to correct it)? <input checked="" type="checkbox"/> yes <input type="checkbox"/> no Which group is Harald assigned to by the system? <input type="checkbox"/> High (>66%) <input type="checkbox"/> Medium (<66% & >25%) <input checked="" type="checkbox"/> Low (<25%) What support measures will Harald receive? <input type="checkbox"/> Qualifying, such as courses and further training <input checked="" type="checkbox"/> Stabilizing and increased support <input type="checkbox"/> None Can Harald appeal against this decision? <input type="checkbox"/> yes <input checked="" type="checkbox"/> no 																	

Fig. 9. **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?