International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Better Together? On the Design and Use of Explanations to Support Novices in Individual and Collective Deliberations About AI

Timothée Schmude, Laura Koesten, Torsten Möller & Sebastian Tschiatschek

To cite this article: Timothée Schmude, Laura Koesten, Torsten Möller & Sebastian Tschiatschek (07 May 2026): Better Together? On the Design and Use of Explanations to Support Novices in Individual and Collective Deliberations About AI, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2650562

To link to this article: https://doi.org/10.1080/10447318.2026.2650562

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2650562

Better Together? On the Design and Use of Explanations to Support Novices in Individual and Collective Deliberations About AI

Timothée Schmude[a], Laura Koesten[b,c,d], Torsten Möller[e] and Sebastian Tschiatschek[f]

[a]Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science, University of Vienna, Vienna, Austria; [b]Department of Human-Computer Interaction, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE; [c]Faculty of Computer Science, Research Group Visualization and Data Analysis, University of Vienna, Vienna, Austria; [d]Center for Technology Experience, AIT Austrian Institute of Technology GmbH, Vienna, Austria; [e]Faculty of Computer Science, Research Network Data Science, Research Group Visualization and Data Analysis, University of Vienna, Vienna, Austria; [f]Faculty of Computer Science, Research Network Data Science, Research Group Data Mining and Machine Learning, University of Vienna, Vienna, Austria

ABSTRACT

Everyone affected by high-risk AI systems should have the chance to understand them and consider the merits and harms of their deployment. Explanations of AI systems can support this goal, yet it is rarely explored how lay audiences use them for deliberation. In this paper, we examine how explanations support groups and individuals of AI novices in learning and deciding about an AI system. We use question-driven explanations spanning four information categories and evaluate them in a task-based interview study with 8 focus groups and 12 individuals. We find that groups allow participants to use team cognition to make sense of explanations, but rely heavily on social dynamics. In contrast, single settings support focused understanding but lack exchange and discussion. We contribute suggestions to make explanation designs suitable for AI novices and discuss their use in individual and collective settings to support understanding and deliberation of AI systems.

KEYWORDS

Explainable AI; understanding; deliberation; qualitative methods; focus groups

1. Introduction

A growing number of AI systems[1] are deployed in the public sector to decide about critical issues, such as employment, migration, and criminal justice (Bansak et al., 2018; Chouldechova, 2017; Scott et al., 2022; Züger & Asghari, 2023). These systems can have consequences for all stakeholders but tend to have significant impact on their decision subjects (i.e., the people the system decides over), such as discrimination or misclassification (Brown et al., 2019; Raji et al., 2022). These harms intensify when decision-making is opaque and uncontestable (Alfrink et al., 2023; Ananny & Crawford, 2018; de Fine Licht & de Fine Licht, 2020). For these reasons, public AI systems should be considered as "matters of public interest" (Züger & Asghari, 2023), meaning that they need to be explainable, justifiable, and open for public deliberation (Brown et al., 2019; Kawakami et al., 2024; Naiseh et al., 2024). Explanations can make AI systems more understandable and easier to assess and control (Langer et al., 2021). While explainable AI (XAI) research is often focused on individuals, research has shown that group settings can facilitate the understanding of complex topics (Moshman & Geil, 1998; Navajas et al., 2018; Nokes-Malach et al., 2015). Further, group settings encourage the exchange of views and arguments (Smith et al., 2009; Stromer-Galley, 2007), which are vital when engaging in deliberation (collectively finding a solution to a problem (Habermas, 1991)). Yet, XAI research has not explored in

CONTACT Timothée Schmude timothee.schmude@univie.ac.at Faculty of Computer Science, Research Network Data Science, Doctoral School Computer Science, University of Vienna, Währinger Str. 29, 1090 Vienna, Austria

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2026.2650562.

detail how AI novices interact with and use explanations collectively in comparison to individual settings, a gap that this paper aims to address.

Individuals affected by decisions of AI systems often do not have technical knowledge that would facilitate understanding and scrutinizing the decision-making systems. This work understands this user audience as "lay people" (DeVito et al., 2018; Lima et al., 2023; Shulner-Tal et al., 2023) or "*AI novices*" (Mohseni et al., 2020): users who may use AI products in their daily life but have no technical background nor knowledge about the underlying logic and construction of these systems. Explanations for AI novices naturally have different requirements than explanations for AI practitioners, such as developers and creators of AI systems (Liao et al., 2020), as they have different expertise (Ehsan et al., 2024), interests (Langer et al., 2021), and prior knowledge (Cheng et al., 2019; Schmude et al., 2025; Szymanski et al., 2021). Explanation formats (e.g., visual, textual, dialogue) are known to impact AI novices" understanding (Bove et al., 2022; Cheng et al., 2019; Szymanski et al., 2021) but show inconsistent effects (Bove et al., 2022) due to contextual factors such as participants' perceptions of the use case domain. Personalizing explanations can help bridge the gap between explanation affordances and user needs (Conati et al., 2021; Naiseh et al., 2020; Shulner-Tal et al., 2023), but no single explanation spans all needs and contexts (Freiesleben & König, 2023), so analysis of the ADM domain and audience is still necessary. Requirements of AI novices are only recently addressed in XAI, and deliberation scenarios have rarely been explored (Naiseh et al., 2024). Prior work in XAI has elicited novices' information needs (Schmude et al., 2025), co-designed explanations with public-sector users (Lee et al., 2019b; Weitz et al., 2024), and examined deliberation in human–AI teams (Chiang et al., 2023). To our knowledge, however, no study has examined how question-driven explanations support domain experts and decision subjects to understand AI in the employment domain, and how AI novices use these explanations to deliberate AI deployment in group versus individual settings.

Designing explanations that support understanding and deliberation for AI novices in both group and individual settings meets multiple challenges. Group composition and dynamics place special demands on explanation design (Naiseh et al., 2021), as explanations need to cater to a diverse set of information and format preferences (Bertrand et al., 2023; Bove et al., 2024; Schmude et al., 2025). They must further support a joint understanding process and collaborative interactions (Long et al., 2021), such as sharing and combining information, while providing comprehensive details and maintaining clarity and navigability. We address these challenges by proposing a modular explanation design that spans four information categories (*data, system details, usage, and context*) from which users can select. Another challenge consists of validating explanation approaches qualitatively with the relevant stakeholder groups. Specifically, XAI research does not always include people from marginalized population groups, who are most likely to be affected negatively as decision subjects (Brown et al., 2019). To address this, we conducted two focus groups with decision subjects to include their perspectives and voices on AI systems in the public sector.

To examine the role of explanations in supporting AI novices' understanding and deliberation, we present the findings of a task-based interview study with 43 participants, involving 8 focus groups and 12 single interviews. For this study, we used an explanation design comprising 36 single explanations in question-answer pairs. These explanations are organized into the four categories: *data*, *system details*, *usage*, and *context*, and further assigned to subtopics and levels of detail (Figure 1). Participants used these explanations to solve the study tasks and decide whether to deploy a public AI system (Figure 3). We used an employment scoring algorithm that is connected to previous work on AI systems in employment (Lopez, 2019; Niklas et al., 2015; Scott et al., 2022). Our analysis examines participants' self-reported understanding, decision confidence, and perceptions of key information. We further conducted a thematic analysis of how participants interacted with explanations in both settings. The study's procedure, methods, and analysis approach are guided by the following research questions:

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

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

Our findings suggest that explanations support both individual and group settings but develop different understanding facets: individuals better applied their knowledge in the study tasks, while groups built contextual understanding by exchanging information, perspectives, and anecdotes. Group success further depended on social dynamics: productive dynamics enabled knowledge sharing, correction, and constructive disagreement, while negative dynamics led to discouragement and abandoned understanding. We thus find that to support AI novices in deliberation, explanations should surface relevant information early, include value and norm considerations, encourage scrutiny, and, when possible, combine individual and group settings to leverage the benefits of both. Our contributions include i) an in-depth account of how AI novices use explanations to understand an AI system and deliberate its deployment, identifying interaction benefits and drawbacks of individual and group settings; ii) an analysis of which types of explanations participants requested most often and perceived as most important; iii) recommendations for facilitating understanding and opinion formation in group settings; and iv) a question-driven, modular explanation approach that accommodates different levels of detail and is suitable for both individual and group settings. We envision that this work can provide valuable starting points for future XAI research that examines how explanation design and social setting interact to support AI novices in understanding and forming opinions of public AI systems.

2. Background and related work

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

2.1. Human-centered explainable AI

Explainability is often described as a cornerstone of responsible AI systems (Thiebes et al., 2021), as explanations can enable stakeholders such as domain experts and decision subjects to understand (Langer et al., 2021) and contest AI decisions (Alfrink et al., 2023). A similar focus is set by the domain of human-centered AI (Capel & Brereton, 2023), which proposes to build AI systems that 1) are based on user-experience design and stakeholder engagement, and 2) empower rather than replace people by being controllable and autonomy-preserving (Donghee Shin, 2023; Shneiderman, 2022; Xu, 2019). These principles become especially important in high-risk settings (European Commission, 2024), such as employment (Ammitzbøll Flügge, 2021; Scott et al., 2022), immigration (Bansak et al., 2018), or criminal justice (Chouldechova, 2017), where erroneous or nontransparent algorithmic decisions can cause severe harm to decision subjects (Raji et al., 2022). In response to these risks, the domain of *human-centered explainable AI* (HCXAI) examines how explanations can contribute to "equitable and ethical Human-AI interaction" (Ehsan et al., 2023b). It assumes that transparency alone is not enough to make AI systems explainable (Ananny & Crawford, 2018), but that explanations need to consider the system's social context (Wenzelburger et al., 2024), its lifecycle (Dhanorkar et al., 2021), and its different stakeholder groups (Ehsan et al., 2023b). In the context of this work, human-centered explainability is achieved by testing and validating a design approach intended to support AI novices in understanding AI systems and making informed decisions about their deployment in public institutions (Züger & Asghari, 2023).

2.2. Designing explanations for AI novices

Most people who interact with AI systems in public institutions are lay people or AI novices: "users who [might] use AI products in daily life but have no (or very little) expertise on machine learning systems" (Mohseni et al., 2020). In this study, the term denotes users who may be familiar with common AI applications, such as generative chat or translation systems, but have little to no understanding of their technical foundations, including machine learning or symbolic, rule-based techniques. In contrast, AI practitioners are stakeholders with experience developing, creating, or technically interacting with AI systems and their underlying models (Liao et al., 2020). Explanatory methods such as LIME

(Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), and surrogate models (Molnar, 2025) are typically tailored to AI practitioners and presuppose technical knowledge, and hence often fail to meet the needs of AI novices. Addressing novices' needs requires understanding how non-experts understand and perceive AI systems. Prior HCI research has analyzed lay understandings to examine user perception of algorithmic systems (DeVito et al., 2018; Eslami et al., 2016). Similarly, XAI research has begun exploring the information needs of AI novices to design explanations for broader audiences (Schmude et al., 2025). However, few explanation designs fully adopt the perspective of AI novices (Cheng et al., 2019; Szymanski et al., 2021). In the following, we summarize current approaches to AI novices' information needs and explanation design practices.

Regarding **information needs**, qualitative research shows that AI novices value information about a system's context and purpose (Kawakami et al., 2024; Schmude et al., 2025) and about the deploying institution (Brown et al., 2019). In addition to information about system mechanics and outputs, explanations for AI novices also require normative information, such as justifications for design choices (Biran & McKeown, 2017). Regarding **information coverage**, prior work argues that transparency does not equal understanding (Ananny & Crawford, 2018) and that providing all available information is not an effective explanation strategy. Empirical studies show that "white-box" explanations can improve objective understanding[2] but may overwhelm non-experts and reduce perceived understanding (Cheng et al., 2019). Later work (Bove et al., 2024) found the opposite effect, attributed to differences in application domains, suggesting that information quantity should be context-dependent. Regarding **explanation format**, Szymanski et al. (2021) found that while lay users preferred visual explanations, they performed better with textual ones. Other studies confirm this discrepancy and suggest that comprehension varies with demographics and domain knowledge (Ehsan et al., 2024; Shulner-Tal et al., 2023; Wang & Yin, 2021). Adapting explanations to the target audience may be addressed through **personalization** (Conati et al., 2021; Shulner-Tal et al., 2023), whereby explanations are suited to stakeholder role (Langer et al., 2021), prior knowledge (Schmude et al., 2025), beliefs (Miller, 2019), and explanatory stance (Byrne, 2023; Keil, 2006). Additional considerations include the explanation's purpose (Freiesleben & König, 2023) and the user's familiarity with AI (Kramer et al., 2018). The explanation design presented in this study was informed by the approaches described here by, e.g., covering both mechanistic and normative information and allowing for a degree of personalized selection regarding topics and detail levels.

2.3. Question-driven explanation design

The explanation design in this study follows a question-driven design process introduced by Liao et al. (2021), which grounds explanation design in questions posed by end users. The process is composed of four phases: question elicitation, question analysis, mapping questions to modeling solutions, and design and evaluation. Although many XAI toolboxes exist, designers still face challenges in selecting appropriate explanation methods for specific use cases, as available methods are often designed for developers, whose interpretability questions differ from those of other user groups (Schmude et al., 2025). User-centered approaches such as question-driven design seek to bridge this mismatch between existing explanation methods and user needs, known as the "sociotechnical gap" in XAI (Ehsan et al., 2023a).

For this study, we draw on question elicitation and analysis from prior work (Liao et al., 2020; Schmude et al., 2025) to construct a set of user questions that the explanations should address. We compiled information from multiple sources documenting the employment prediction algorithm (Allhutter et al., 2020; Scott et al., 2022), resulting in a collection of "scavenged" (Wieringa, 2023) material. To map this information to user questions, we followed two approaches. First, we adopted design guidelines by Lim and Dey (2009), which propose that context-aware applications should provide information about system mechanisms, situational context, and inputs and outputs. Second, we drew on information categories identified in an interview study with 24 AI novices (Schmude et al., 2025), which grouped information needs into four categories: *data*, *system details*, *usage*, and *context*. While *data* and *system details* address questions about training data, inputs, workings, and outputs, *usage* and *context* address what Lim and Dey (2009) describe as "situation," namely the system's

operation and deployment context. These four categories represent updated versions of those proposed by Lim and Dey (2009), adapted to the needs of AI novices.

We further applied principles of explanation soundness (fidelity, complexity) and completeness (coverage, density) (Chatti et al., 2022; Guesmi et al., 2024; Kulesza et al., 2015) by introducing subtopics and a hierarchy of explanation levels. Combining question-driven explanations, levels of detail, and user-controlled information selection supports modularity and interactivity (Cheng et al., 2019; Guesmi et al., 2024; Schmude et al., 2023) and enables adaptation of explanations to users' needs (Conati et al., 2021; Shulner-Tal et al., 2023). Section 3.3 describes how these principles were implemented in the final explanation design.

2.4. Analytical lenses: Understanding and deliberation

In the following sections, we introduce understanding and deliberation as the primary analytical lenses for this paper. Section 3.6 then operationalizes them for the evaluation of the explanation design and settings.

2.4.1. Individual and collaborative understanding of AI systems

Improving understanding of an AI system is the primary goal of explanations, as understanding enables assessment (e.g., fairness) (Langer et al., 2021) and action (e.g., contestation) (Henin & Le Métayer, 2022) for stakeholders. However, understanding has many definitions (Baumberger et al., 2017; Grimm, 2019; Keil, 2006; Zagzebski, 2019). We draw from learning sciences, cognitive sciences, and XAI to define understanding as i) connecting and applying information (Baumberger et al., 2017; Grimm, 2019), ii) attempting to grasp underlying structure through simplification (Zagzebski, 2019), iii) comprising multiple facets that include analytical and emotional connections to information (explain, interpret, apply, take perspective, empathize, self-reflect) (Wiggins & McTighe, 2005), and iv) forming a "working" understanding by recognizing and filling gaps until deemed sufficient (Keil, 2006). Because understanding is difficult to define and measure (Sato et al., 2019), recent work proposes an abilities-based approach (Speith et al., 2024), aligned with learning-science operationalizations (Wiggins & McTighe, 2005). We examine understanding by analyzing which facets participants use to grasp the algorithmic system, their performance in the study task, and their confidence in the deployment decisions (Section 3.5).

While individual understanding is well studied in XAI (Chen et al., 2023; Cheng et al., 2019; Schmude et al., 2023; Wang & Yin, 2021), group understanding has received less attention (Chiang et al., 2023). Cognitive science has examined distributed cognition (shared cognitive load) (Hollan et al., 2000) and outsourcing (Keil, 2006) (delegating understanding) as effective mechanisms in collaborative settings, such as navigation (Keil, 2003). These require "cognitive symbioses with mutually supporting roles" (Keil, 2006), i.e., a constructive working dynamic. Similarly, educational psychology shows that peer discussion (Smith et al., 2009), collaborative reasoning (Moshman & Geil, 1998), and aggregated knowledge (Navajas et al., 2018) let groups perform better than individuals. Group performance, however, depends on interaction dynamics (Nokes-Malach et al., 2015): failures may originate from increased memory load and disrupted retrieval (losing train of thought), while success involves shared ground and the combination of task-relevant information. Thus, while it is not clear from the outset whether groups are better for learning than one-on-one settings (Bloom, 1984), their advantages of sharing cognitive load and exchanging perspectives likely supports solving complex problems and evaluating AI system deployment.

CSCW research has long examined collaborative settings, group composition, and interaction (Convertino et al., 2008; Fiesler et al., 2019; Sutcliffe, 2005). XAI research has only recently begun to consider explanations for group interactions, noting that "many-to-one" interactions differ from "one-to-one" due to group dynamics, cognitive bias amplification, and trust issues (Naiseh et al., 2024). Recent cognitive science research further explores how including AI systems in teams affects cognition and interaction, described as *teamness* (Cooke et al., 2024). These works examine trust, performance (Myers et al., 2024), and shared mental models (Narayanan et al., 2025), but assume that an autonomous AI act as an equal team member. Although prior XAI work comparing individual and group

understanding in AI-assisted decision-making found little difference (Chiang et al., 2023), how groups use explanations in collective reasoning thus remains underexplored. This study addresses this gap by focusing on human-only teams tasked with deciding about AI deployment to empirically compare explanation effects on AI novices' understanding in group and individual settings.

2.4.2. Deliberating on AI systems

Deliberation, meaning informed reasoning and decision-making, builds on understanding (de Fine Licht & de Fine Licht, 2020) and enables citizens to debate public-sector AI proposals and their consequences (Kawakami et al., 2024; Züger & Asghari, 2023). Habermas (1991) defines deliberation as the exchange of rational-critical arguments aimed at finding solutions. Such arguments are grounded in truth or a *shared understanding* of reality, are open to judgment, and can be defended. This link between shared understanding and deliberation is central to our examination of explanation effects. Deliberation occurs across many domains shaping politics and social life (Lupia, 2024), including public referendums (e.g., Swiss federal and state laws (Swiss Confederation, 2025)), citizen forums on public issues (e.g., water supply in California (Innes & Booher, 2003)), and grassroots community initiatives (e.g., the right to repair movement (Collins et al., 2024)). These settings share that they involve social entities engaged with their environment and use deliberation and productive conflict to negotiate policy issues (Hajer & Wagenaar, 2003). Although such formats are imperfect and may incur cognitive biases, such as *groupthink*[3] (Baron, 2005; Janis, 1971; Naiseh et al., 2024), they provide spaces for the public to gather, discuss, form opinions, and decide on public interests.

We argue that AI systems in public institutions constitute such public interests, captured by the term *public interest AI*. Züger and Asghari (2023) use this term to emphasize that public-sector AI systems must demonstrate their benefits and meet obligations of justifiability, equality, openness to validation, technical security, and emergence from a *deliberative or co-design process*. Identifying formats that support deliberation on public AI remains an open research challenge. Prior work has examined the use of mini-publics (Fung, 2003) for co-designing algorithmic policy (Lee et al., 2019b) and for supporting procedural justice in algorithmic resource allocation (Lee et al., 2019a). HCI research further shows that participatory formats can connect communities and institutions in public service transformation (Crivellaro et al., 2019). In XAI, studies indicate that group discussions facilitate critical analysis of AI recommendations and that presenting both pros and cons leads to more frequent and productive deliberation (Chiang et al., 2023). However, it remains to be explored how in-person collaboration supports deliberation about deployment decisions for high-stakes public AI systems. We address this gap by implementing mini-publics as focus groups composed of three stakeholder groups: domain experts, decision subjects, and members of the general public. We compare deliberation processes across groups and against individuals' "internal deliberation" in single interviews (described in Section 3.6) to derive insights on how explanation designs and social settings support AI novices in deliberating public AI systems.

3. Methods

In this section, we describe our methods and study procedure. We conducted a task-based, semi-structured interview study with 43 participants (Section 3.1), divided into 8 focus groups of 3–5 participants each and 12 single interviews (Section 3.4). Participants were presented with the study's algorithmically-supported employment prediction use case (Section 3.2) and a collection of explanations about the system employed in this use case (Section 3.3). Participants then solved four tasks concerning the algorithmic system's handling of a fictional job-seeker and decided about the system's deployment (Section 3.5). The study closed with an interview, lasting 90–120 min for focus groups and 60 min for single interviews. We analyzed individual and collective interactions, including explanations, self-reports, and deliberation processes (Section 3.6). The University of Vienna's Research Ethics Committee approved this study under reference 01189.

3.1. Participants

3.1.1. Recruitment

Tables 1 and 2 provide an overview of the study participants. Recruitment partially focused on individuals with prior contact to the job agency, either professionally or personally. Participants were recruited through cooperation with civil society organizations, an employment agency, public calls for participation, and the authors' extended network. For focus groups, the authors contacted staff from these organizations, as identified in previous studies, or used general inquiry channels to describe the study and invite participation. Interested organizations all offered to support the recruitment process by coordinating with the authors to select and invite potential participants, and to arrange a suitable location and time for the studies. Groups A, B, C, E, F, G, and H were organized this way. Group D was recruited through the authors' network and consisted of individuals who had previously been job-seeking. For individual studies, participants were recruited using the same channels, and calls for participation were further advertised on screens and information boards throughout different city districts. All studies were conducted in person, either in an office or in public spaces. Participants were compensated with 30€ for participation in focus groups (90–120 min) and 20€ for single interviews (60 min). Our approach to organizing, composing, and moderating focus groups was informed by Krueger (2004). 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 (Hennink et al., 2017).

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

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

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

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

3.1.2. Recruitment criteria

Participants were required to be of full legal age and AI novices, i.e., to have no technical knowledge about machine learning systems as described in Section 2.2. These criteria were screened in a pre-questionnaire before invitation to the study using two questions: "How would you rate your knowledge of algorithms?" and "How would you rate your knowledge of artificial intelligence (AI)?" Each question could be answered on a scale corresponding to "no knowledge at all" to "professional and detailed knowledge." Here, the first question elicited technical knowledge, as participants familiar with AI tools might have rated their AI expertise as high but were unlikely to have knowledge of algorithms without a strong technical interest or background.

Participants were asked about their knowledge of and experience with AI systems through interview questions at the beginning of each study. Most participants were familiar with services based on large language models, including chatbots, machine translation services, and search tools, but were unfamiliar with algorithmic decision-making systems such as the study's use case. Participants who had worked as counselors or trainers, such as those in Groups B, C, and H, reported being aware of efforts by public institutions to automate services through the use of software. However, the systems in question were described as implementing symbolic, rule-based models by matching users to specific job profiles based on certain keywords, following a different logic than the system used in the study.

3.1.3. Group composition

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 individuals who are competent in the field for which the AI system is used, such as job counselors or advisors (groups B, C, G, and H). We define decision subjects as individuals who may be impacted by the system's decision, such as job seekers and those who have previously been job seeking (groups E and F). All remaining participants are considered members of the general public and were included to test changes in explanation effects and participants' perceptions (groups A and D). A similar classification of roles was applied to the recruitment individual participants. Here, S1, S2, S4, and S5 were recruited in their roles as past decision subjects, as they had been job-seeking the past and reported on the study's topics from this experience. S3, and S6 to S12 were recruited as domain experts who had professional experience in interacting the Public Employment agency.

3.2. Use case: The AMS employment prediction algorithm

3.2.1. Description

The *AMS algorithm*[4] is a system developed to calculate the employability of job-seekers in Austria. It was created by a private company for the Austrian Public Employment Agency between 2015 and 2021 but was never used as a live system and put on hold in 2021 due to privacy objections (Allhutter et al., 2020). The system uses a logistic regression model trained on historical data to predict job-seekers' employment chances based on demographic features (such as age, education, nationality, etc.) and work history. The outputs are a short-term and long-term employment score for each job-seeker (Gamper et al., 2020). These scores would serve as recommendations for the job-seekers' counselors at the employment agency to assist in deciding about suitable support measures. Counselors could overwrite the system's predicted scores of job-seekers but would need to give a reason for doing so (Allhutter et al., 2020; Holl et al., 2018). More information is provided in Supplementary Appendix A.

3.2.2. Choice of use case

Algorithmic tools that assist in assessing job-seekers and resource allocation have been deployed in various countries, including Germany (Bundesagentur für Arbeit, 2021), Poland (Niklas et al., 2015), and the Netherlands (Desiere & Struyven, 2021). However, the introductions of these applications also repeatedly led to sociotechnical conflicts (Scott et al., 2022). The deployment of the *AMS algorithm* was motivated by three overarching goals: a) increasing consultation efficiency, b) increasing support measure effectiveness, and c) reducing arbitrariness (Gamper et al., 2020). Detailed reports warned that counselors might over-rely on the algorithm or hesitate to overrule its suggestions (Allhutter et al., 2020). Further, the algorithm's model and underlying data structure were predicted to discriminate against marginalized groups, who would lack the option to contest the system itself (Lopez, 2019). Transparency and ongoing scrutiny of the algorithm were listed as necessary measures to prevent these risks (Allhutter et al., 2020). As the *AMS algorithm* represents a larger class of algorithmic decision-making systems that spark public debate around their deployment in public institutions (Raji et al., 2022), it exemplifies how AI systems become matters of public interest and presents a suitable use case for our study.

3.3. Explanation design

3.3.1. Description

The explanation design comprises 36 question-answer pairs about the *AMS algorithm*. Each question belongs to one of four categories, *data* (format, content, limitations), *system details* (features, model, examples), *usage* (operation by and interaction with users), and *context* (intention of deployment, target group, responsible actors). Each category is further divided into topics with three levels of increasing detail (base level, level 2, level 3). Every explanation takes the form of an A5 paper sheet and contains a question (e.g., "Who operates the system?") answered with a brief text or image (cf. Figure 2). The answers to the questions were manually compiled from the information that was gathered about the *AMS algorithm*, as described in Section 2.3. Information is presented in various formats, primarily relying on textual content and utilizing highlighting, colors, and illustrations to emphasize key points. During the study, participants first received an overview of the explanations (Figure 1) and the four base explanations, and could request levels 2 and 3 at any time during the explanation phase (as depicted in Figure 3). The explanations were provided on paper to allow participants to interact with them physically and to facilitate social interactions, such as exchanging, pointing, and reading to each other.

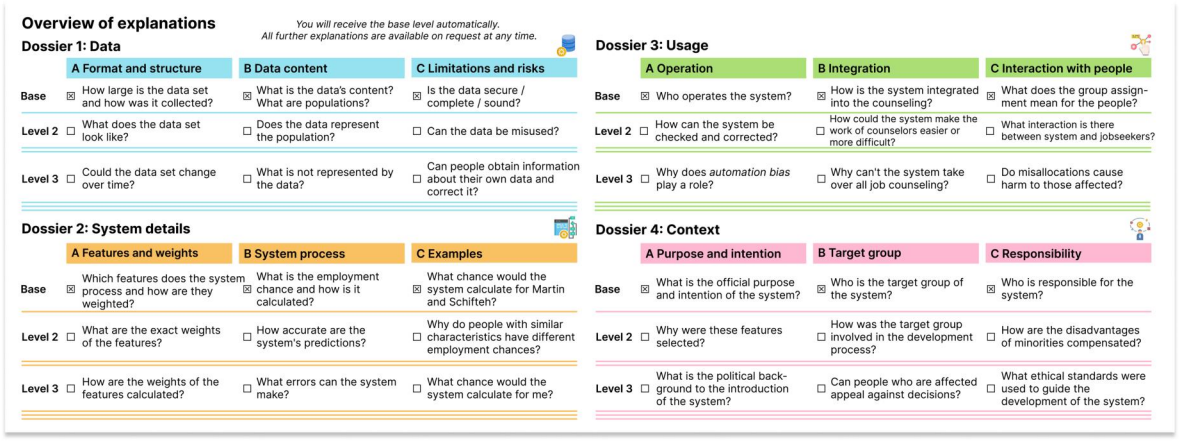
3.3.2. Design foundations

The explanation design is intended to allow the users to explore information in a flexible and self-directed manner. To this end, the design employs a *modular structure*, which divides the explanations into four information categories, each further subdivided into topics and three levels of detail (base level, level 1, and level 2). This structure provides information on every aspect of the AI system, enabling participants to select the most relevant content according to their information needs and preferences. It further aims to avoid limitations of "groupware" systems (Mandviwalla & Olfman, 1994) by supporting both individual and collaborative interaction and by providing multiple user perspectives (e.g., domain experts, decision subjects).

The four information categories, *data*, *system details*, *usage*, and *context* are based on research on AI novices' information needs and covers both technical and sociotechnical system aspects (Schmude et al., 2025). The subdivision of explanation categories into topics organizes the broad supply of information while accommodating different needs of information completeness and soundness (Kulesza et al., 2013) and introducing a degree of personalization (Chatti et al., 2022).

The division of explanations into levels of detail is inspired by design approaches of previous work (Guesmi et al., 2024) and is intended to signal to users that explanations have different levels of depth. Beginning with fundamental information at the base level, every subsequent detail level should reveal a new, yet increasingly peripheral or technical aspect of the topic. For instance, the base level of the topic "Examples" in category *system details* (Figure 1) simply introduces two different cases and their assigned employment probability. The second level presents a counterfactual through two cases that differ in only one aspect: their gender. The third level introduces a personalized local explanation by inviting participants to calculate their own employability chance using the formula provided. In this case, these detail levels require an increasing level of technical literacy, hence serving to signal an increase in difficulty, as key aspects of the calculation were already communicated in the first levels. For the information categories *usage* and *context*, higher detail levels are not intended to include more complex information, but rather to include information that develops side aspects, such as the role of *automation bias* and the system's political background. However, assigning information to levels of detail is challenging and subject to interpretation. In our study, this separation produced specific issues that conflicted with the design intentions, which are described in Sections 4.1.2, 4.1.4, and 5.4.

Lastly, the explanation design includes several established explanation methods such as feature importance, local and global explanations, examples, counterfactuals, and argumentative approaches (Figure 2). Depending on the guiding question, different explanation methods are suitable to provide sensible answers. For example, the system's feature weights are shown in the form of a global explanation, whereas a consideration of whether the system makes the work of counselors easier or harder is presented in an argumentative format. The explanation design thus organically implements various methods from explanation method taxonomies (Speith, 2022).



The four icons IST: 2026-04-15: 11:21:08 PM ** This track PDF was generated by the online proofing system (for reference only) ** Page 103 of 104 used in this figure and throughout the paper are provided by Freepik (data, usage), Flat Icons (system details), and noomtah (context) through Flaticon.com.

Figure 1. Overview of explanations. Depicted is an overview of all 36 questions, which acted as the explanations and were organized into four categories: *data*, *system details*, *usage*, and *context*. Each category includes three levels of questions, from basic to advanced. 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.[5]

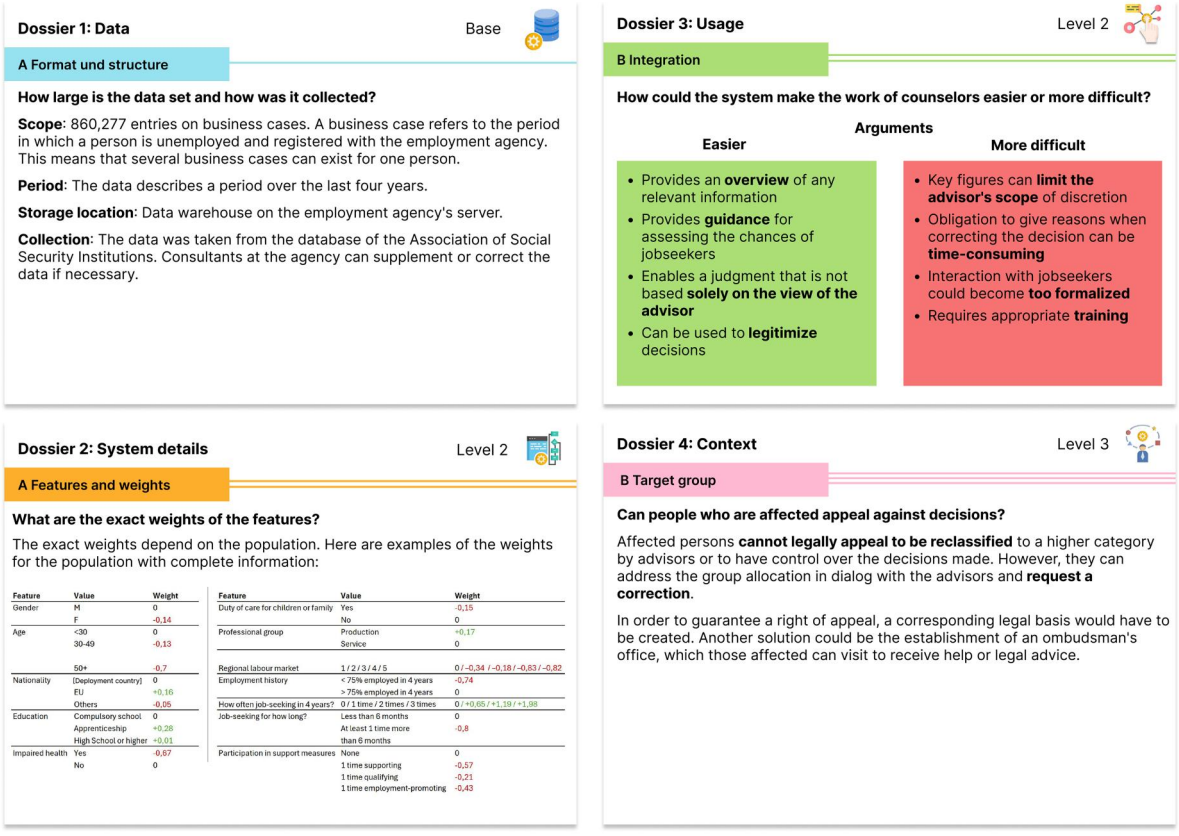


Figure 2. Four explanation examples. The figure shows examples of explanations in the four categories: *data*, *system details*, *usage*, and *context*, at different levels. The explanations provide information on data collection and structure, the system's feature weights, the system's impact on counselors, and the rights of affected individuals to contest decisions. Explanations were printed on sheets of A5 paper and could be either fully textual or contain visual elements, such as charts or colored shapes. Each category was given a different color and icon to facilitate navigation.

3.3.3. Pilot studies

The concrete design of the explanations was iteratively designed and tested with three pilot focus groups composed of participants from the authors' network of colleagues. The second and third focus groups included AI novices to improve comparability with the target participants. All focus groups were asked for their feedback on the selection of the four information categories, the formulation of the questions, the helpfulness of the answers, and whether any information was missing. Feedback from the first pilot group led to adaptations in the provision of explanations such that later detail levels would be available from the beginning instead of being revealed progressively; the inclusion of an explanation overview to guide navigation between categories and detail levels; the inclusion of a joint decision in the study process to encourage discussion within the group; format changes to the explanation design to introduce more visual information and reduce text volume. Feedback from the second and third pilot groups led to minor cosmetic changes in the explanation format as well as to small changes to the study procedure, such as an increase in the time provided for initial exploration of the explanations.

3.4. Study procedure

In the following, we describe the procedure in the focus group and single interview setting (depicted in Figure 3). Like previous work in XAI (Chiang et al., 2023), we conducted individual interviews and focus groups to compare how the social setting would affect participants' understanding and deliberation processes. Self-report and interview questions are listed in Supplementary Appendix B.

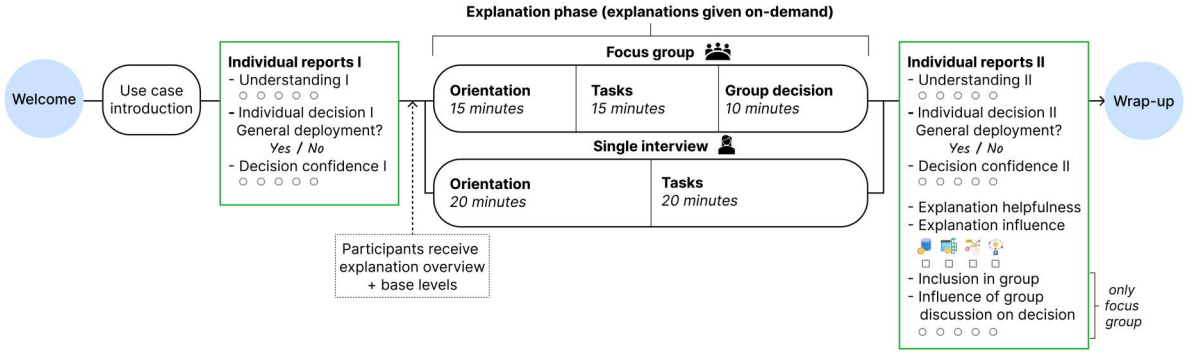


Figure 3. Overview of the study procedure. The figure outlines the study procedure, beginning with a welcome, use-case introduction, and initial individual reports. Participants then entered the explanation phase, freely exploring explanations before completing the tasks, with focus groups additionally making a joint decision. Afterwards, participants completed a second set of individual reports assessing understanding, decisions, decision confidence, and the influence and helpfulness of explanations (plus group dynamics for focus groups). The session concluded with a brief wrap-up. Focus groups and single interviews differed only in the explanation phase and the second individual report.

3.4.1. Focus group procedure

Throughout the study, participants sat together with the investigator and could freely interact with each other. They first completed consent forms and questionnaires about demographics and knowledge about employment (domain knowledge) and AI systems (technical literacy). A round of introductions followed, during which each group member stated their name and described their most recent interaction with AI to break the ice. The investigator then explained the study procedure and distributed a mock newspaper article introducing the AI use case (included in Supplementary Appendix A). Participants indicated their understanding, deployment decision, and confidence in their decision for the first time (3.5). The study's explanation phase followed (including orientation, task, and decision phase), throughout which participants received and kept access to all explanations. Initially, the group received an overview of the explanations and all base-level explanations. All other explanations could be requested at any time. After 15 min, participants received task sheets and had another 15 min to complete them, deciding independently whether they wanted to work together or individually. Finally,

the group had 10 min to make a joint decision on deployment (yes/no, with conditions allowed). A second round of individual reports followed, described in the next section, and then the investigator concluded the study. The focus group studies lasted approximately 90 min.

3.4.2. Single interview procedure

The majority of the study procedure remained the same in single interviews. However, the explanation phase did not include a group decision phase. Instead, the orientation and task phases were prolonged to 20 min each to provide individuals with the same total time as focus groups to interact with the explanations. Participants in single interviews were asked to think aloud while reading through the explanations, except during periods of focused attention, to provide insights into their thought processes. The study investigator further asked intermittent questions about participants' impressions and perceptions of information during the explanation phase. The individual study setting was included to analyze how understanding processes changed depending on whether participants worked in a group or alone, and whether the explanation design would support both settings. Single interviews took around 60 min.

3.5. Study elements

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

3.5.1. Use case introduction

Participants received initial information about the *AMS algorithm* in the form of a mock newspaper article inspired by an Austrian newspaper publication from 2019 (Szigetvari, 2018). The article provided key information about the system's basic workings, goals, and deployment context and featured the opinions of employers and employee associations about its merits and risks. The presentation format was chosen to provide an introductory summary of the AI system, utilizing a familiar layout and non-technical language, while highlighting both the pros and cons of the system's deployment. Thus, the article served as a basic introduction to the use case, which aimed to approximate the amount of information participants might receive through the media. This way, participants received a baseline of information with which to assess their initial understanding, the decision to deploy, and their confidence in that decision. Further, this introduction served to outline relevant aspects that could be explored using the explanations.

3.5.2. Explanation phase (orientation, tasks, and group decision)

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

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

Task 1: Can Harald change the data stored about him (e.g., to correct it)? (yes / no)
Task 2: In which group of employment chances does the system categorize Harald? (high (>66%) / medium (<66% & >25%) / low (<25%))

Task 3: Which support measures will Harald receive? (qualifying measures, such as courses and training / stabilization and increased supervision / none)
Task 4: Can Harald appeal against this decision? (yes / no)

For the group decision, focus group participants had 10 min to discuss the system's deployment and were asked to collectively decide whether to accept or reject it. This was intended to simulate a small referendum in which each participant's vote counted toward the final outcome. If no consensual decision was reached in time, participants were asked how they thought the situation should be resolved (e.g., by a majority vote). They were further informed about the option to specify conditions for the system's deployment.

3.5.3. Individual reports I and II

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

To prevent influence between participants' reports, individual reports in focus groups were conducted anonymously and re-assigned by the study examiner using a color-coded reporting system (Figure 4). Each participant was assigned a color and given the complete material for all individual reports. For each reporting question asked by the study investigator, participants took the corresponding paper slip from their materials, wrote their answer, and placed it in a gathering container that hid it from view. Due to this process, each participant held only the reporting material relevant to themselves, which avoided potential mix-ups and prevented them from seeing how other participants responded to the reports. The gathered reporting material was recorded by the investigator at the end of the study.

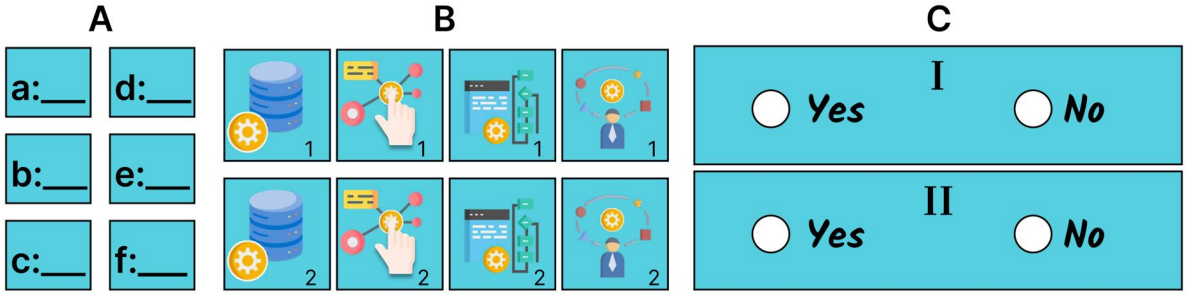


Figure 4. Material for individual reports of participants. The figure depicts the reporting materials used by participants to answer questions during the individual reports I and II. The material was provided as laminated paper slips in different colors. Slips that were numbered with letters a to f (A) were used to report on questions using 5-point likert scales for understanding, confidence, inclusion in group, and influence of group discussion, and were answered by writing a number between 1 and 5. Slips with icons (B) served as a selection of the most helpful and influential explanation categories, and participants were asked to select any number of icons. Slips with decisions (C) served as voting ballots for deployment decisions and were answered by ticking yes or no.

3.6. Analysis

All focus groups and interviews were audio-recorded and transcribed. These transcripts provided the data basis for the thematic analysis. Participants' individual reports, task solutions, decisions, and the investigators' field notes provided further data for within- and between-subject comparisons of understanding and decision-making. A summary of metrics and analysis methods used is provided in Table 3.

3.6.1. Thematic analysis

For both research questions, we conducted a thematic analysis (Braun & Clarke, 2006) of participants' articulations to develop a qualitative account of their understanding and decision-making processes. The first author conducted an initial inductive coding of the transcripts and derived preliminary themes. All authors then discussed and refined the codebook, after which the revised codes and themes, as well as the deductive frameworks, were reapplied to all transcripts in a second pass. The resulting inductive code base was structured along the overarching categories of understanding, deliberation (including decision-making processes and arguments), opinions (e.g., regarding AI and policy choices), and experiences (e.g., anecdotes and lived situations). The full codebook is provided in the Supplementary Appendix. We highlight that while the quantitative items in participants' self-reports characterize the diversity of participants' perceptions and facilitate qualitative exploration (Weiss, 1995), they are not intended to invoke "inference […] of greater generality" (Maxwell, 2010) nor impose a mental model based on variance theory (Patton, 1990).

3.6.2. RQ1-explanations

RQ1 focuses on the explanations' impact on participants' understanding of the study's use case and the differences between individual and group settings. We employ a triangulation approach (Carter et al., 2014) by using three methods to analyze the effect of explanations on participant understanding. Firstly, comparing participants' individual reports before and after the explanation phase was a subjective indicator of changes in their understanding and decision confidence. Secondly, participants' answers to the four study tasks indicated their factual or testable understanding. Lastly, participants' verbal reports during and after the explanation phase were used to analyze their understanding processes and barriers thematically. In a deductive analysis, we compared their interactions to mechanisms of "collaborative success and failure" (Nokes-Malach et al., 2015) and the "six facets of understanding" (Wiggins & McTighe, 2005). With this three-part combination, we examined participants' subjective understanding, their information gain, and the cognitive processes of their understanding. This choice was motivated by educational psychology research indicating that understanding cannot be solely elicited through test questionnaires (Sato et al., 2019), but involves emotional (Wiggins & McTighe, 2005) and meta-cognitive processes (Veenman et al., 2006) that are equally important. Our focus on understanding is motivated by previous XAI research, which highlights the importance of understanding in decision-making processes (Hoffman et al., 2023; Langer et al., 2021; Donghee Shin, 2023).

3.6.3. RQ2-deliberation

RQ2 focuses on how participants formed opinions about the *AMS algorithm*, weighed the pros and cons of its deployment, and settled on a deployment decision. To this end, we compare participants' decision confidence before and after the explanation phase. We further conduct an inductive and deductive thematic analysis of participants' interactions in both settings to connect them to the "elements of deliberation" (Stromer-Galley, 2007) – a set of characteristics of deliberation processes. Based on this, we analyze when participants used arguments (grounded, defensible positions), opinions (personal judgments on things, values, states), and personal experiences (Mercier & Landemore, 2012; Stromer-Galley, 2007) to consider the system's deployment. For single interviews, we examine participants' responses to interview questions during the study to examine their reasoning process and "internal deliberation" (Mercier & Landemore, 2012). Lastly, to account for one of the most prevalent cognitive biases in group settings, we examine focus groups for occurrences of groupthink (Baron, 2005; Janis, 1971) – an effect that sets in when concurrence-seeking in groups overrides realistic argumentation and discussion.

3.6.4. Summary of analyses and metrics

Table 3 summarizes the analysis methods and metrics used in the study, describes the corresponding analysis targets, and maps them to the study procedure.

Table 3. This table summarizes the targeted concepts, metrics, and analysis methods used in the study.

Concept to be measured	Metric	Study part
Self-reported understanding	Likert-Scale	Individual reports I + II (two times)
Deployment decision	Yes / no decision	Individual reports I + II (two times)
Decision confidence	Likert-Scale	Individual reports I + II (two times)
Participant understanding	Task performance	Explanation phase: tasks (Harald G's case)
Explanation category helpfulness	Multiple choice	Individual reports II
Explanation category influence on decision	Multiple choice	Individual reports II
Felt inclusion in group	Likert-Scale	Individual reports II
Influence of group discussion on decision	Likert-Scale	Individual reports II
Concept to be measured	**Analysis method**	**Study part**
Participant understanding	Inductive thematic analysis	Explanation phase, individual reports II
Participant understanding	Deductive thematic analysis	Explanation phase, individual reports II
Deductive framework: facets of understanding (Wiggins & McTighe, 2005)		
Participant interaction experience	Inductive thematic analysis	Explanation phase, individual reports II
Participant interaction experience	Deductive thematic analysis	Explanation phase, individual reports II
Deductive framework: Collaborative success and failure (Nokes-Malach et al., 2015)		
Participant deliberation	Inductive thematic analysis	Explanation phase, individual reports II
Participant deliberation	Deductive thematic analysis	Explanation phase, individual reports II
Deductive framework: Elements of deliberation (Stromer-Galley, 2007), internal deliberation (Mercier & Landemore, 2012), groupthink (Baron, 2005)		

The study parts correspond to the study steps depicted in Figure 3 and described in Section 3.5. For every mention of deductive thematic analysis, the table also describes the theoretical frameworks employed.

4. Results

This section presents our results, answering our research questions: How question-driven, modular explanations[7] support understanding in individual and group settings (RQ1, Section 4.1) and how AI novices used explanations to form opinions and decide about the system's deployment (RQ2, Section 4.2). The participant labels denote the study setting (focus group: A–H / single interviews: S) and the participant ID, as listed in Tables 1 and 2. To distinguish themes in the analysis, *inductive themes* are italicized, while "deductive themes" are put in quotation marks. Block quotes from participants are italicized, and block quotes from focus groups are color-coded to visually represent speakers. All interviews were conducted in German, and quotes were translated into English.

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

To examine how AI novices used the explanations to understand the study's use case, we analyzed their self-reports, articulations, and interactions in both settings. We found that each setting supported different aspects of understanding, suggesting a tradeoff between them. We first describe how the explanations contributed to *shared understanding* and "collaborative success" in groups (4.1.1) and continue with the explanations' role in instances of "collaborative failure" (Nokes-Malach et al., 2015) (4.1.2), summarized in Figure 5. We then describe individuals' interactions with the explanations (4.1.3), participants' feedback on the explanation design (4.1.4), and summarize the benefits and drawbacks of both settings for XAI (4.1.5).

4.1.1. Groups' benefits: Shared understanding and increased engagement

In the best cases, groups leveraged the modular explanation structure to utilize distributed cognition (Keil, 2003), whereby participants process information in parallel and then combine it, thereby supporting interactive team cognition, which involves creating knowledge through team interactions. Our analysis focuses on the exchanges and interactions between team members, following the premise that "team cognition is an activity, not a property or a product" (Cooke et al., 2013). We use the term *shared understanding* to capture interactions that realize these forms of cognition. Examples of such interactions included *locating information together*, *sharing information with others*, *discussing interpretations*, *debating task solutions*, and *querying and explaining* (a question by a group participant invites other participants to contribute their thoughts). The explanations were only afforded to this set of interactions within groups, as they required social interaction with other participants. For example, in Group C, participant C1 read the first study task aloud and asked for input (*querying*), after which the group discussed solutions (*explaining*):

C1 *Can Harald change the data stored about him? Yes, he can certainly change it, can't he?* […]

C3 *Which stored data, the one down there?* [points at Harald's demographic features]

C1 *Yes, just that.*

C3 *49 – no, male – no. The apprenticeship – no, Austria – he can still change that. Duty of care – he could get married or have children. He could change his service sector. He could change his career. Impairment …*

C1 *Well, what is meant by 'change'? When he enters the data, he can change the data. He doesn't have to specify the knee problem.* […]

C3 *So he can change it.*

C2 *Yes.*

Cognitive mechanisms of collaborative success. Interactions such as *querying and explaining* and *discussing interpretations* rely on cognitive mechanisms of collaboration between participants, such as "sharing working memory resources," "complementing others' knowledge', "re-exposing information," and "correcting errors." These cognitive mechanisms are aspects of "collaborative success" (Nokes-Malach et al., 2015) and provide groups with multiple ways to tackle explanations. For example, participants tended to work through information about *usage* and *context* alone or in pairs but raised explanations with the group when they were difficult to understand or piqued their interest. This was often the case with explanations of *system details*, which included numerical information crucial to understanding the system's calculations. We describe the process of using others' understanding to close gaps in one's own understanding as *outsourcing* (Keil, 2006). Since understanding AI systems involves interacting with a variety of different information categories (e.g., technical, political, social), outsourcing provides a way to hand information to the team member most competent in this category. For example, in Group B, B2 expressed their appreciation for B3′s help in solving study task 2: "*It was a math problem. You [B3] filtered it out well. It was very analytical. With your help, we were able to recognize these weak points.*" In contrast to *querying and explaining*, which participants used to invite input or spark conversation, *outsourcing* was thus used for the active delegation of an impeded understanding process into the group.

Social mechanisms of collaborative success. The explanations further served to support mechanisms of collaboration by encouraging the group to share their personal experiences and opinions about certain aspects of the AI system. We present an excerpt from Group G as an illustration, which was composed of participants in leadership roles of a social institution. Here, participant G4 *shared an explanation* that documented the algorithm's impact on two job-seekers ("joint management of attention"), which prompted G2 and G3 to *discuss interpretations* ("increased engagement"). This interaction established "common ground" that the group later used for deliberation:

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

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

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

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

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

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

Role of explanations in supporting collaborative success. Addressing both sides of collaboration, cognitive and social, is an important aspect in supporting the interactions that create interactive team cognition (Cooke et al., 2013). When explanations support both sides of collaboration, they can thus be a key driver in developing *shared understanding*. We argue that the set of interactions termed *shared understanding*, which is enabled through the combination of our explanation design and the team cognition of groups,

Table 4. Reported understanding and task performance.

Focus groups

ID	Und. I	Und. II	Change	Task performance	ID	Und. I	Und. II	Change	Task performance
A1	2	2	0	●●○○	F1	2	4	+2	●●○○
A2	1	2	+1	●○○○	F2	1	4	+3	●●○○
A3	4	4	0	●○○○	F3	5	2	−3	●○○○
A4	4	4	0	●○○○	F4	3	3	0	●○○○
B1	4	2	−2	●●●○	F5	4	4	0	●●○○
B2	4	2	−2	●○○○	G1	5	5	0	●●○○
B3	5	5	0	●●●○	G2	3	2	−1	●●●○
B4	2	3	+1	●●○○	G3	5	2	−3	●●○○
C1	4	4	0	●○○○	G4	5	4	−1	●●○○
C2	2	2	0	●●○○	H1	4	5	+1	●●●●
C3	1	3	+2	●○○○	H2	4	5	+1	●●●●
D1	4	3	−1	●●○○	H3	4	4	0	●●●●
D2	4	3	−1	●●●●	H4	4	5	+1	●●●●
D3	4	4	0	●●○○	H5	4	4	0	●●●●
E1	3	4	+1	●●○○					
E2	3	3	0	●○○○					
E3	4	4	0	●○○○					

Single interviews

ID	Und. I	Und. II	Change	Task Performance	ID	Und. I	Und. II	Change	Task Performance
S1	5	5	0	●●●○	S7	4	4	0	●●●●
S2	4	2	−2	●●●○	S8	3	4	+1	●●●●
S3	4	4	0	●●●●	S9	5	5	0	●●○○
S4	4	4	0	●●○○	S10	4	3	−1	●●●○
S5	5	4	−1	●●●●	S11	4	4	0	●●●●
S6	1	4	+3	●●●○	S12	2	4	+2	●●○○

This image displays a table summarizing changes in participants' understanding and task performance in focus groups and single interviews. It includes columns for each participant's ID, two reports of understanding (labeled "und. I" and "und. II"), the change between these reports, and task performance. The "change" column shows how participants' ratings shifted, color-coded to indicate whether the change was positive (green), neutral (orange), or negative (red). Task performance is represented with dot patterns, where a higher number of dots indicates better performance.

presents a valuable pathway to help AI novices understand algorithmic systems. These interactions can be especially useful when group members have different domain expertise and information needs, as the group can then use complementary knowledge and outside perspectives to make sense of information. At the same time, these interactions are partly dependent on the group dynamic. Positive interactions like those described were especially frequent in Groups G and H (job counselors and trainers), where participants knew and trusted each other. To provide counterexamples, the next subsection describes instances where groups encountered challenges in understanding, illustrating the importance of social mechanisms.

4.1.2. Groups' drawbacks: Process loss and susceptibility to social dynamics

In some of the focus groups, participants lost track of information, forgot their train of thought, or abandoned understanding altogether. We summarize these effects under the term *impeded understanding* and its final result as *abandoned understanding*.

Factors in impeded understanding. We found that *impeded understanding* occurred due to *explanation design flaws* and co-occurred with adverse social dynamics, resulting in "process loss" (groups falling short of their potential performance (Kerr & Tindale, 2004)). For some participants, the benefits of the explanations' modular structure turned into disadvantages when it hampered them in navigating and retrieving information. Such impeded interactions included *cumbersome information uptake*, being *overwhelmed by information*, and *relying on intuition over information*. Furthermore, for some participants, the group setting exacerbated these impediments. For example, participant B3 stated that "*For me, it doesn't make sense to […] split up [the explanations], and everyone reads a part, that's actually not enough.*"

Mechanisms of collaborative failure. This assertion aligns with the preposition that interactive team cognition is located in the exchanges between group members, "rather than the static properties of their shared knowledge structure" (Cooke et al., 2013). Therefore, impediments in group understanding can be connected to mechanisms of "collaborative failure" (Nokes-Malach et al., 2015). When groups encountered difficulties interacting with the explanations, they also incurred a "memory coordination

cost" (increased cognitive load) and a "retrieval strategy disruption" (losing their train of thought). We illustrate these mechanisms with an example from Group G. Although this group was composed of participants with university education, it did not succeed in calculating the employment chance for Study Task 2, in contrast to single-interview participants with the same education.

G4 *We can go through the features briefly. Where is the piece of paper with this terrible matrix?* […]

G2 *I still don't understand which value to put. To calculate it, I need an exact value for the weighting.*

G1 *You can calculate it with this. The apprenticeship has 52%. I believe that he [Harald] has over 25%.*

G3 *Yes, definitely, I mean, roughly speaking…*

G1 *She [Shifteh] has over 30%. And she also has 2 minuses […] and a plus.*

G3 *That's also how I estimated it.* […]

G2 *But how do you calculate it? […] And why are there differences between the general weighting and the exact calculation? That doesn't click for me right now.*

Explanations' role in impeded understanding. Here, the levels of detail in the explanations acted against participants' understanding by obscuring the actual feature weights, which were first accessible in level 2 of *system details*. In the explanations base level, feature weights were encoded with plus and minus signs in the intention to make them easier to read and avoid information overload as observed in previous work on "white-box" explanations (Cheng et al., 2019). While this was an advantage in other cases, this form of encoding reduced the clarity of the explanations and obscured information in the excerpt above. This issue may be resolved through improved navigation, but it highlights the challenge of simplifying information without omitting key aspects.

Adverse group dynamics are a key driver of abandoned understanding. Importantly, *impeded understanding* alone did not mean that collaboration failed altogether; rather, it depended on how groups dealt with understanding issues. Here, key aspects were group cohesion (Kerr & Tindale, 2004) and constructiveness (Niculae & Danescu-Niculescu-Mizil, 2016). In groups where members didn't know each other from before, and mutual trust was not yet established, the group dynamic gave room to negative social mechanisms, such as "social loafing" (group loses motivation) and "fear of evaluation" (being criticized by others), and participants began to *abandon understanding*.

Having problems with understanding something naturally causes a feeling of embarrassment (Rozenblit & Keil, 2002). So, to share understanding problems with someone else, one must be able to trust the other person. Therefore, while positive group dynamics allow participants to make understanding issues a group activity to be resolved collectively, if participants do not trust the group dynamics, they would rarely raise their understanding issues with others. Consequently, if an explanation text appeared unintelligible to a member of a group, but this member did not trust the dynamic enough to engage in collaborative mechanisms, they abandoned the effort to understand this specific explanation, and perhaps any explanation at all. This means that when group members are too embarrassed or uninterested to share their lack of understanding, and they begin to fight for themselves, this signifies the breakdown of collaboration and interactive team cognition (Cooke et al., 2013). Limiting these negative dynamics and promoting mutual trust is therefore an essential goal of both the explanation design and social setting.

We found examples of these adverse social dynamics most frequently in Groups E and F, which were composed of job-seekers. Participants had trouble engaging with the explanations and abandoned interactions and understanding by saying: "*Probably [you can solve it] with that, but I don't know, I'm too stupid for that.*" (E3) or "*I don't know what I should say. Everything has already been said.*" (F2). Here, two things failed: The explanations failed to make crucial information accessible, and the group failed to uplift members who were discouraged. Interactions that offset this discouragement, such as *locating information together* and *outsourcing*, were not realized in Groups E and F. Co-design approaches could make explanations more suitable for decision subjects and lower the understanding barriers, as demonstrated with public servants (Weitz et al., 2024). Further, XAI should employ methods that foster a productive social dynamic, which we identify as the second key aspect to support "collaborative success" and *shared understanding* (Figure 5).

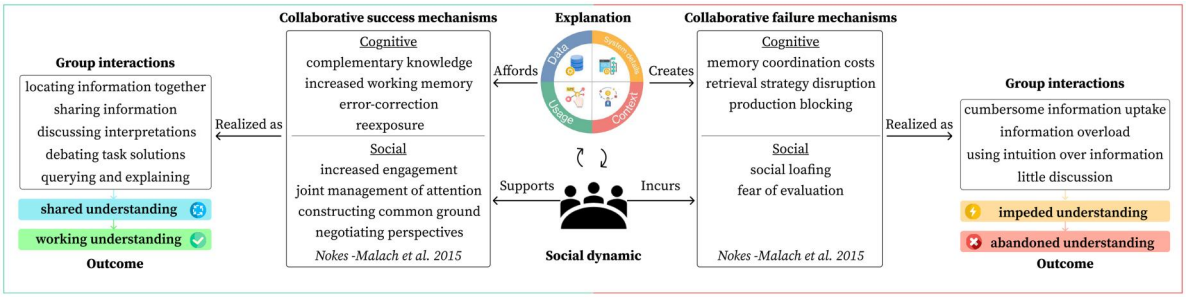


Figure 5. Both explanations and social dynamics are key factors for collaborative performance. This figure depicts a conceptual model of how collaborative success and failure mechanisms shape group interactions and their outcomes. At the center, explanations and the social dynamic act as forces that can either contribute to collaborative success or failure. On the left, productive interactions, such as jointly locating information and discussing interpretations, lead to *shared* and *working understanding* through cognitive and social success mechanisms like complementary knowledge, engagement, and perspective negotiation (Nokes-Malach et al., 2015). on the right, cumbersome or overloaded information uptake reflects failure mechanisms, such as memory costs and social loafing (Nokes-Malach et al., 2015), leading to *impeded* or *abandoned understanding*. From the perspective of XAI, both explanation and social dynamics are thus important aspects to keep in mind when designing explanations for groups in collaborative settings.

4.1.3. Why individuals performed better in the study tasks but still felt the absence of collaboration

Despite the different advantages and disadvantages that group and individual settings offer in learning contexts (Section 2.4.1), these differences are rarely examined empirically in XAI. We address this gap by comparing in-person focus groups and single interviews to examine whether the social setting impacted participants' understanding. We use a triangulation approach (cf. Section 3.6) by investigating participants' understanding with respect to three aspects: interactions with the explanations, task performance, and self-reported understanding.

Explanation interaction. Participants in single interviews tended to request the same number or more explanations than focus groups on average, despite having less working memory at their disposal. While generally, participants with higher education requested more explanations throughout groups and individual studies, individual interview participants nevertheless tended to request a larger volume of explanations than any group member on their own. Regarding the cognitive load, S5 described: "*Well, maybe [it was overwhelming] at the very beginning […] But I then realized that I could get through it to some extent.*" Single interview participants also regularly *consumed explanations in bulk*, i.e., read through the whole of an explanation category rapidly. This interaction was nearly nonexistent in focus groups. However, single interview participants often stated that they missed the "*exchange with people, with other perspectives*" (S3). S8 explained that this exchange would allow for a different form of understanding:

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

Task performance. Participants in single interviews performed better or equally well in the study tasks compared to groups with a similar educational background (Table 4). Task performance refers to the number of correctly answered questions in the study's fictional job-seeker scenario (Section 3.5). A possible explanation is that participants in single interviews engaged more intensely with the study tasks, as they often calculated the exact employment chance of Harald G. in study task 2. None of the focus groups completed this step, irrespective of education, but rather made educated guesses. This might be explained by the degree of focused attention the settings afforded participants. As single interviews incurred no distractions, participants could immerse themselves in the explanations.

Inductive analysis of understanding. Most participants in both group and individual settings reported unchanged understanding after the explanation phase (Table 4). Paradoxically, the same participants

verbally stated that their understanding improved. E3 commented: "*I don't think I understood it the way you can understand it yet, but it's definitely better than before.*" And S3 explained that: "*I would still say my understanding is "good," but this "good understanding" is much more informed now than the first superficial one.*" This indicates that participants tended to judge their understanding relative to the information available, not necessarily in relation to their previous report. We describe this process as *calibrating understanding*. Previous research in cognitive science has documented similar effects (Keil, 2006), which were also reproduced in an XAI study on white-box explanations (Cheng et al., 2019). In total, participants' verbal reports, their feedback on the explanations (described in the following Section), and the calibration process itself indicate that the explanations in fact improved understanding. Including additional measures, such as information gain, could further capture the calibration process, which is discussed in Section 5.2.

Deductive analysis of understanding. To compare individual and group interactions with the explanations, we lastly draw from the "six facets of understanding" (Wiggins & McTighe, 2005). The framework describes that understanding is represented by the ability to "explain, interpret, apply, take perspective, empathize, and self-reflect" with respect to a topic. The more facets are covered, the better the understanding. Seeing that individuals seemed to have an advantage in solving the study tasks (when compared to groups of equal educational background) suggests that the individual setting supported the "apply" facet. In contrast, the group settings often led participants to "explain" information to others, to "interpret" it by expressing their views and opinions, to "take perspective" by articulating criticism, and to "empathize" through the sharing of anecdotes and experiences. As explanations aim to improve understanding of a given AI system, combining both settings to cover more facets of understanding could thus be a fruitful approach. Furthermore, explanations for individuals can benefit from information that addresses facets typically dependent on social interaction. Our design aimed to implement this through explanations such as *What chance would the system calculate for me?* (interpret: making it personally relevant), and *How could the system make the work of counselors easier or more difficult?* (take perspective: provide multiple angles and arguments).

Synthesis of individual and group settings. Previous work in HCI has found that group interaction boosts task performance compared to individual settings (Karadzhov et al., 2023). In our study, individuals surprisingly performed better in the study tasks than groups, but also stated multiple times that they would appreciate the opportunity to exchange information with others. Part of the discrepancy in performance may be attributed to age and educational differences between groups and individual participants; however, even limiting the comparison to participants with similar demographics (e.g., S7-S12 to Group G and H) still surfaces differences in explanation interaction, developed understanding facets, and number of explanations requested. Further, group settings provide advantages for team cognition and reciprocal encouragement, as described in Sections 4.1.1 and 4.1.2. We argue that both group and individual settings can contribute to participant understanding and should ideally be combined. In particular, focused attention can facilitate the application of information, while *shared understanding* and the exchange of opinions and arguments (Section 4.2) aid encouragement, reflection, and collective action. Considering this tradeoff between settings can inform how explanations can be combined with social settings to cover as many facets of understanding as possible.

4.1.4. Reflections on the explanation design: Modularity, levels of detail, and most important information

To examine how the explanation design was received, we asked participants for feedback on the explanations' structure, content, style of expression, and information coverage. These questions were asked as part of an interview, to which participants responded with verbal statements. In groups, each member was invited to contribute their feedback openly, allowing all group members to discuss the statements in conversation. We report and summarize the participants' criticisms as a basis to formulate design improvement suggestions in Section 5.

Strengths and weaknesses of the design. Positive comments described the explanations' structure as "nicely presented" (A2, C2) and "good to get an overview" (C3, H4) while being "active and controllable" (S8). Critical comments described the information coverage as "too much" (D1, S4), and the structure as "confusing" (B1, D1) and "demanding" (D2). Participants saw the design's strengths in its four-category structure, question-driven presentation, active selection, and information scope. However, the scope and depth of information also led to information overload and loss of overview. Furthermore, the explanations' numerous and complicated texts were described as "very difficult" (E2, F2). E2 compared the language to "*letters […] from the court. I understand every single word, but I don't understand the context.*" Previous work has found that explanations which heavily rely on a textual format can effectively convey information but tend to raise aversion with users (Schmude et al., 2023; Szymanski et al., 2021). However, Weitz et al. (Weitz et al., 2024) paradoxically found that users preferred textual over graphical formats, as they could "work faster with text." This highlights the need for further research on textual formats in XAI, such as the automated adaptation of text to different levels of difficulty.

Most helpful and influential information. Participants in focus groups stated that all explanation categories helped their understanding and influenced their decision evenly (Figure 6), often mentioning that "*all of them [are relevant] … I don't think you can leave anything out, really*" (D3). In contrast, participants in single interviews found *data* much less helpful and less influential, stating, e.g., that they prioritized another category in the time available. Notably, participants emphasized that two categories were central: *system details* and *context. System details* were perceived as "tangible" (S6) and "concrete" (S8), and explanations about the features and weighting were perceived as especially important: "*That is the central point, the basis of the whole system.*" (G4) In turn, explanations from the category *context* were requested the most (Figure 7). Here, participants appreciated explanations that described the decision subjects' inability to contest decisions and the system's political background. Drawing from the concept of "intelligibility types" (Lim & Dey, 2009), we argue that *system details* provided descriptive information to the question "What did the system do? ," while *context* provided normative information to the question "Why did the system do [this]? ." Future research should investigate how both types of information can be integrated into explanations for AI novices.

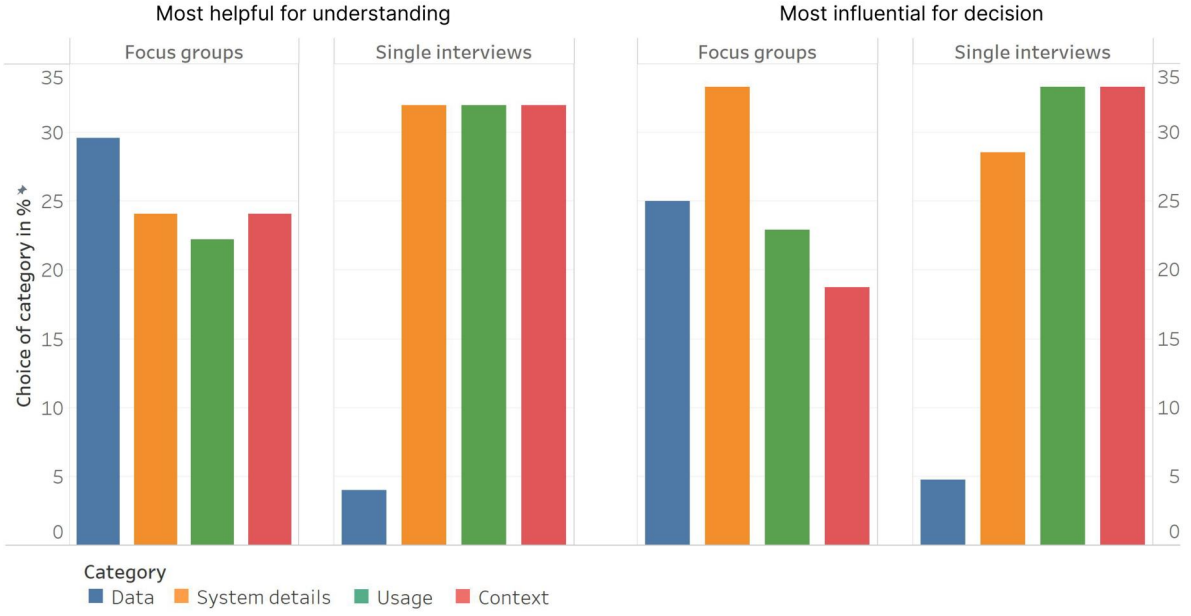


Figure 6. Most helpful explanation categories for understanding and most influential categories for participants' decisions. Depicted are bar charts comparing participants' reports on which categories of explanations (data, system details, usage, and context) they perceived as most helpful for understanding (left) and most influential for their decisions (right). Participants could select any number of explanation categories for both questions, including none and all four. Focus group participants found all categories helpful for understanding, but reported that system details were more influential in their decisions. Participants in single interviews found data both less helpful and less influential and prioritized other categories in the given time (Section 3.3).

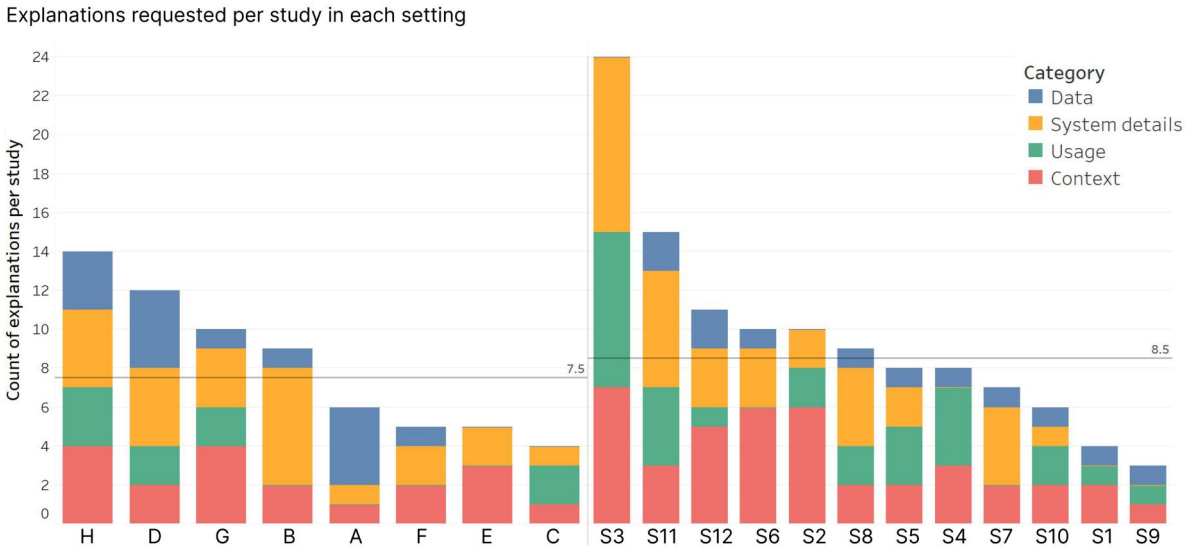


Figure 7. Number of explanations requested. Depicted is a bar chart displaying the count of explanations requested per study across group (left) and individual (right) settings. The categories of explanations, *data*, *system details*, *usage*, and *context*, are stacked to show the total number of explanations requested in each session. Horizontal reference lines indicate the median of requested explanations in the group and individual settings. While groups were able to process many explanations by splitting the reading, several single interview participants went through equal or even higher counts. *Context* explanations were the only category requested in every study.

4.1.5. Summary RQ1-explanations: How does a question-driven, modular explanation design support AI novices' understanding in groups and individual settings?

The findings in this section indicates tha explanations can support both individual and collective settings but differ in the understanding they develop. In groups, we found that explanations facilitated interactions that produced *shared understanding* and involved cognitive and social mechanisms of "collaborative success" (Nokes-Malach et al., 2015). When the social dynamic was trusting, this collaboration of participants acted against discouragement. When groups had a negative social dynamic, the explanation design could become overwhelming, and understanding issues were left unchecked, leading to "process loss" (Kerr & Tindale, 2004) and *abandoned understanding*. Participants in single interviews interacted with the explanations in a more focused and self-directed manner, which had advantages for task performance and engagement with the explanations. However, the positive effects of aggregated knowledge (Navajas et al., 2018) and peer discussion (Smith et al., 2009) in group settings should not be disregarded. Our findings showed that group settings can help bridge understanding issues by boosting morale and allowing participants to share knowledge, interpretations, and experiences. We thus argue that individual and group settings support different *understanding facets* (Wiggins & McTighe, 2005), meaning that they provide different grounds for understanding AI systems. While individual settings can make it easier to understand technical and numerical details that require much attention ("apply"), group settings can support understanding of the deployment context and consequences through the exchange of expertise and lived experiences ("interpret," "take perspective," "empathize"). Consequently, individual and group settings should be combined to leverage their different modes of interaction and understanding facets when explaining AI systems. In cases where both social settings cannot be provided, explanations of AI systems should aim to reinforce facets that are not covered in the corresponding setting.

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

Before and after the explanation phase, participants decided if the study's case should be deployed, and groups additionally made a collective decision (Table 5). We compared participants' decisions and decision confidence, as well as their deliberation process, to examine the impact of explanations and the social setting on deployment. We first describe participants' confidence and decision changes in single

interviews (4.2.1) and then present three cases of group discussion illustrating "elements of deliberation" (Stromer-Galley, 2007), including reasoned arguments (4.2.2), disagreement (4.2.3), and groupthink (4.2.4). Section 4.2.5 summarizes the findings.

Table 5. Individual and collective decisions about deploying the study's use case.

Focus groups

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

Single interviews

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

The tables summarize individual and group decision patterns as well as decision confidence ratings for both focus groups and single interviews. The columns list participants' IDs, their initial decision (decision I), the group decision where applicable, their subsequent decision after the explanation phase (decision II, colored red when changed), and a color-coded confidence score indicating positive (green), neutral (yellow), or negative (red) change. In most focus groups, decision confidence increased after the explanation phase, with the exceptions of groups E and F (Section 4.1.2). In single interviews, participants' confidence increased except for S3 and S11, who explained that their confidence decreased due to strong adjustments to their mental models of the use case (Section 4.2.1).

4.2.1. Explanation phase led to increased decision confidence and decision swings

For most participants, deciding about the *AMS algorithm's* deployment was a clear choice: 7 out of 8 groups and the majority of single participants voted "No" (Table 5). Many participants reported increased decision confidence after the explanation phase and stated that they felt better informed due to the explanations and, where applicable, the group discussion. Reasons for these increases included a better understanding of the system's "*fundamental idea*" (S4) and the "*exchange of different opinions and things that catch your eye*" (G2). Participant B3 emphasized that the explanations, although they only contained factual information, provided a stark contrast to *public narratives*:

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

This contrast highlights how explanations of an AI system can impact decision-making by correcting lay understandings (DeVito et al., 2018) and is in line with previous work that demonstrates how making both the model's workings (Lee et al., 2019b) and its context of use (Weitz et al., 2024) visible to users can be helpful to tackle lay understandings. It further directly connects the explanations and changes in participants' decision confidence. Few participants reported decreased decision confidence, and only S3 and S11 reported a stronger decrease of −2 (Table 5). S3 explained that the system might have benefits, but it depended too much on the *conditions for deployment*. In particular, it should be deployed "*responsibly, with a pilot project, in a selected group, for three months*," and not haphazardly, where "*you sit around for a day or eight hours and then training is finished*" (S3). S11, who together with S3 requested the most explanations out of all participants, paradoxically stated that their confidence decreased because "*I'm still*

missing so much information. Especially […] how tedious it is for the counselors if they have to disagree with the system." In consequence, S11 changed their decision from "Yes" to "No." Notably, this *fear of an algorithmic imprint* was a prevalent theme throughout all studies and was often connected to past experiences with digitization projects and the corresponding *institutional deficiencies*. S5, who also changed from "Yes" to "No," similarly stated that the explanations helped them to *scrutinize the system*: "*You don't have to introduce anything that's extra bad.*" While the explanations thus made the *decision more uncertain* for some, they undoubtedly encouraged critical reflection about the use case and triggered decision changes. Despite only having "*their own mind*" (S8), single participants could make use of the explanations to *weigh pros and cons* and *adjust their mental model*. This form of "internal deliberation" is supported by exposure to different views, as provided through argumentative explanations in *usage* and *context*. This suggests that explanations can substitute at least small parts of public deliberation, which is thought to be the more salient driver for "reasoning towards good outcomes" (Mercier & Landemore, 2012). To contrast these findings with those from focus groups, we illustrate public deliberation with three conversation excerpts that showcase elements of deliberation in the focus groups.

4.2.2. Case 1 - Reasoned arguments: Group B discusses whether to deploy the system

Group B consisted of staff members and volunteers from a civil society organization. Three participants in this group changed their votes from "Yes" in the first report to "No" in the second report. We found this change to be driven by three main deliberation elements (Stromer-Galley, 2007): "sourcing" information, "reasoned arguments" (opinion claims grounded in the information), and "engagement" with the topic and between participants. In the excerpt, B2 and B3 *weigh pros and cons* of deployment. B3 grounds their arguments in explanations about the system's features and weightings (*system details*), changing the discussion's course:

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

B3 *I was originally in favor of 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 the duty of care only applies to women. This comes from a time that should be long gone.*

B2 *Those are strong arguments.*

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

B1 *Did you vote yes first?*

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

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

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

The excerpt highlights how the explanations led B3 to *change their deployment decision* and served as *discussion triggers*. In the resulting discussion, participants state both arguments (*discrimination*, *what's the benefit?*) and opinions (*disagreement with policy choices*, *AI can assist in decisions*). Note that there is a difference between arguments (expression of reasoning processes that can be defended against critique) and opinions (expression of the speaker's belief) (Mercier & Landemore, 2012; Stromer-Galley, 2007). While conceiving arguments to persuade interlocutors can result in confirmation bias (interpreting evidence such that it confirms existing beliefs) (Mercier & Sperber, 2011), the fact that B3 changed their attitude, in fact, indicates that the explanations acted against this bias. We argue that the excerpt thus shows a positive synergy in that the explanations provided grounds for "arguments," which entered the discussion and provoked "collective reasoning" and three decision swings. However, considering the large argumentative influence of B3, it should also be considered how the discussion would unfold if B3 had advocated *for* deployment. A case with comparable dynamics is described in Section 4.2.4.

4.2.3. Case 2 - Disagreement: Group D debates normative positions regarding the algorithmic representation of people

Group D consisted of participants who had previously been job-seeking. When the group discussed the AI system's deployment, the conversation shifted to how features representing job-seekers' profiles should be selected and weighted. This produced disagreement, an "important marker for deliberation" (Stromer-Galley, 2007) that displays heterogeneity of viewpoints, acts against polarization, and involves close examination of others' reasoning. In the excerpt, D3 argues for the system's deployment, while D1 argues against, and D2 acts as a mediator:

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

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

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

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

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

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

D3 *But the human decision is always subjective.*

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

D1 *Or a false image.*

D2 *An incomplete one, I would say.*

The group here *discusses diverging views* and expresses opinions. While these opinions are meant to persuade and defend, they are grounded in *lived experiences* rather than in the explanations. The central conflict develops between D1's belief that the system *misrepresents reality* and D3′s viewpoint that it can *increase objectivity* and *assist in decisions*. The discussion did not lead to a consensus on the deployment decision within the given timeframe, but resulted in a majority vote of "No." We argue that it still illustrates an important process in the deliberation on public AI systems: Participants again "sourced" information that was turned into arguments, but the debate led to a more fundamental topic that surfaced discrepancies which would impede finding a collective decision. The fact that participants then engaged in "disagreement" is a sign of productive deliberation, as it shows that there were diverse viewpoints, that no polarization or "groupthink" (Janis, 1971) occurred, and that the proposal was closely examined based on the information given (Stromer-Galley, 2007). In a real setting, this form of debate could serve as a fruitful basis to investigate whether the system is in the "public interest" (Züger & Asghari, 2023) and to host "early-stage deliberations" (Kawakami et al., 2024) on the system during development. The merit of this debate was further acknowledged by D1, who found the explanations confusing but stated that these exchanges were the study's "centerpiece" and most intriguing part. We argue that the interplay between explanations and group discussion here supported a (simulated) evidence-informed policy-making process (Mair et al., 2019).

4.2.4. Case 3 - Groupthink? Group a follows a minority position and votes for system deployment

Group A was composed of volunteers from a civil society organization. Three participants in this group changed their decisions, shifting from "No" to "Yes" after the explanation phase. We explain these changes in three aspects: First, Group A focused on the explanation category *data* and interacted minimally with other categories. This meant that less attention was paid, for example, to the system's feature selection and weightings, which were decisive in Cases 1 and 2. Second, participants of Group A stated that they were not directly affected by the system, as they were retired, implying low "engagement": "*It doesn't affect me anymore and I think to myself, yeah…*" (A1). Third, participants prioritized group

concurrence above a "careful, critical scrutiny" (Janis, 1971). The following excerpt illustrates the tipping point for the collective decision:

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

A2 *So rather 'no'?*

A1 *Yes, as A3 says, it's… I don't know.*

A3 *Yes and no…* […]

A2 *I mean, it can't be avoided, it will happen. I'm convinced of that, whether we like it or not, it's done.*

A1 *It won't affect us anymore, at least not in the employment office. […] I agree with the majority.*

A2 *But that's difficult now.*

A3 *I'll stick with 'yes'. My daughter would say I shouldn't think so negatively, especially when it comes to AI. […]*

A2 *I say 'yes' too. […] You, A1 and A4, can tip the scales.*

A1 *I say 'yes' now too, but not because I've changed my mind, but because I want an overall solution.*

A4 *I say 'yes,' but I'm leaning towards 'no'.*

Despite their articulated reservations, all participants ultimately decided to vote in favor of deployment. We compared this excerpt with characteristics of "groupthink," a "mode of thinking" in which people value concurrence higher than consideration of alternative courses of action (Janis, 1971). This mode produces defective decision-making processes due to three key aspects: strong social identification with the group, salient norms, and a perceived low self-efficacy to make the decision (Baron, 2005). The excerpt clearly demonstrates two of these aspects: A1 changes their decision due to a desire for group harmony, and A4 follows suit (group identification). Further, both A1 and A3 express their uncertainty and sway between options (low self-efficacy). While A2's statement that *AI is inevitable* is an opinion rather than an argument (neither "sourced" nor the product of evident reasoning), it triggers the group to make a quick decision that disregards any remaining disagreement.

Notably, even though participants in Group A were colleagues in their volunteer roles, their usual roles within the civil society organization did not involve making collective decisions in a professional capacity. Consequently, the lack of familiarity with such decision-making processes could mean that participants had less experience in tolerating constructive disagreement processes, which are vital to deliberation (Stromer-Galley, 2007). This point is emphasized when comparing the deliberation processes to those of groups composed of colleagues with experience in making team-based decisions, which involved disagreement while maintaining positive group dynamics, as seen, for example, in Groups B, C, and H.

However, although the conversation excerpt shows aspects of "groupthink" (Janis, 1971), such as rationalizations of flawed logic and self-censorship, these aspects are not nearly as pronounced as in the literature (Baron, 2005; Janis, 1971; Janis, 1972). For example, the group did not share an illusion of unanimity, and the uncertainty among participants suggests that there were no guiding, salient norms. Still, as participants avoided "disagreement" and instead *followed decisions of others*, the excerpt presents a suboptimal deliberation process (Baron, 2005). In part, this can be attributed to the explanation's failure to make all fundamental information easily available and to not encourage analytical thinking over intuitive, heuristical thinking (Buçinca et al., 2021). In addition, groups might benefit from explanations that highlight opposing viewpoints to fuel discussion. The implications of these findings are discussed in Section 5.

4.2.5. Summary RQ2-deliberation: How do AI novices use explanations to form opinions and make decisions about AI systems in groups and individual settings?

The findings in this section demonstrate how explanations supported deliberation in focus groups and single interviews. Many participants reported improved decision confidence and changed their deployment decisions based on the explanations, often due to a disillusionment regarding the *AMS algorithm's* assumed merits. These changes occurred in both settings, suggesting that the explanations supported

public and internal deliberation (Mercier & Landemore, 2012). In groups, participants used explanations when discussing deployment, as illustrated in Case 1. Case 2 further highlights that explanations surfaced discrepancies in personal beliefs and produced productive conflict. In contrast, Case 3 shows a deployment decision based more on concurrence-seeking than on "collaborative reasoning" (Moshman & Geil, 1998). However, we hesitate to label the exchange as "groupthink," as it does not align with all factors that characterize the phenomenon (Baron, 2005). Based on these findings, we argue that explanations can support people in considering if AI systems are in the public interest and to discuss "*whether* and *under what conditions* to move forward with developing or deploying" them (Züger & Asghari, 2023). To achieve this, both the explanations and the group setting need to i) be designed so that they allow for the easy sourcing of information for arguments, ii) make all relevant information available as soon as possible, and iii) include mechanisms that encourage participants to examine both the proposal and their positions closely. Matching explanations and social setting to support "elements of deliberation" (Stromer-Galley, 2007) thus presents promising starting points for future research on how explainable AI can promote public deliberation on AI.

5. Discussion

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

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

5.1.1. Explanations in individual and collective settings

In Section 4.1, we described that explanations produced *shared understanding* in groups, involving both cognitive and social mechanisms of "collaborative success" (Nokes-Malach et al., 2015). Section 4.2 further showed that explanations improved participants' decision confidence and provided grounds for different elements of deliberation (Stromer-Galley, 2007), such as reasoned arguments and disagreement. In the best cases, focus groups in our study had a familiar (Johnson & Johnson, 1985) and solution-oriented (Niculae & Danescu-Niculescu-Mizil, 2016) atmosphere that facilitated the sharing and discussion of information. In these settings, the modular explanation structure showed its strengths by allowing for the distribution of tasks among group members, providing high levels of detail and breadth if needed, and offering different viewpoints that could be used as argumentative and conversational starting points. In this sense, the explanations fulfilled their aim of supporting learning and deliberation about a public AI system (Kawakami et al., 2024). The interaction between group members is the differentiating factor compared to "one-to-one" (Naiseh et al., 2021) explanation settings. In our study, single interviews allowed for more focused engagement with explanations and a form of "internal deliberation" (Mercier & Landemore, 2012) but lacked the exchange of knowledge and perspectives with others that is deemed central for deliberation about public AI (Züger & Asghari, 2023). Regarding learning and deliberation, XAI would thus benefit from researching how group settings can be used to leverage collective reasoning (Moshman & Geil, 1998), wisdom of the crowds (Navajas et al., 2018), and performance increases through peer discussion (Smith et al., 2009). However, the benefits of group settings have several preconditions, such as the containment of cognitive biases (groupthink (Janis, 1971), equality bias (Naiseh et al., 2024)) and, crucially, a trusting social dynamic (Chiang et al., 2023).

5.1.2. Importance of group dynamics

Social dynamics were essential for the realization of either collaborative success or failure in focus groups, and were shaped by group composition (e.g., occupational background, age, familiarity among group members) and interaction (e.g., affirmation, encouragement, disagreement, perceived hierarchy).

In groups G and H, for example, the social dynamic bridged understanding issues of individual participants and acted against discouragement. In groups E and F, in contrast, these understanding issues eventually led participants to abandon understanding, as the social atmosphere did not support them in overcoming them. Here, a lack of trust or simply unfamiliarity between participants likely led participants not to share understanding issues openly, which resulted in impeded or abandoned understanding. This underscores the importance of creating trust between group members in collaborative XAI settings (Johnson & Johnson, 1985) and of treating composition and social dynamics as conditions that support or impede collective outcomes. 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 (Wason, 1968). Another measure could be the introduction of explanations that introduce opposing viewpoints to encourage changes in perspective and facilitate discussion, as has been done with LLM-generated conversational explanations that acted as "devil's advocate" in previous work (Chiang et al., 2024). Complementary methods may include facilitation to manage turn-taking, prompts that surface the voices of quieter participants, and deliberate composition to ensure diversity of perspectives. Future work could thus examine how explanations can be designed to work with and support different social dynamics and how the information structure presented here could be combined with LLM-based explanations to contribute to team cognition and encourage critical deliberation.

5.1.3. Groupthink

Lastly, regarding cognitive biases, we observed an effect resembling some aspects of "groupthink" (Janis, 1971) when participants in Group A changed their vote to "Yes" to reach a group decision. We argue that this effect originated in the lack of detailed interaction with explanations and, possibly, a perceived low degree of personal affection by the system's deployment. It may also have been reinforced by a dominant actor in the social dynamic since a vocal member guided consensus and there was limited tolerance for disagreement. However, this is contrasted by participants in Group D, who debated at length about the system's deployment without reaching a consensus. Here, persistent disagreement and a more balanced participation pattern acted as counterweight, even though members were not directly affected by the system. Potential measures to avoid groupthink in discussion could thus be to encourage debate, which again could be the introduction of roles to improve "dialectic argumentation" (Mercier & Sperber, 2011), and to explain the system in a way that makes it more personally relevant to participants (Wiggins & McTighe, 2005), e.g., by emphasizing connections to their own experiences.

5.2. Did the explanations improve participants' understanding?

In Section 4.1.5, we described that the explanations helped participants develop different "facets of understanding" (Wiggins & McTighe, 2005). In groups, participants were encouraged to "explain" information to each other and "empathize" with others' experiences, while individuals could better "apply" information in the study tasks. We further described that groups' interactions with explanations realized mechanisms of "collaborative success" (Nokes-Malach et al., 2015). We thus conclude that the explanations had a positive effect on understanding. However, a more complete answer requires that we consider the difference between measurement methods and true cognitive states.

5.2.1. Calibrating understanding

In Section 4.1, we described that a majority of participants reported unchanged understanding after the explanation phase (Table 4) but, paradoxically, described verbally that their understanding improved; two seemingly incongruent pieces of evidence. Recall that understanding involves the flexible use of knowledge across contexts (Wiggins & McTighe, 2005), and that meta-cognition (judging one's own knowledge) is notoriously difficult (Veenman et al., 2006). In consequence, self-reports may be unreliable, task performance captures only the "application" facet, and decision confidence is merely a proxy for a felt working understanding (Keil, 2006). Together, these measurements allow for a triangulation (Carter et al., 2014) but not a perfect representation of understanding. Based on these relations, we explain the contradictory findings with a process we call *calibrating understanding*. The term describes

that participants tend to report understanding not in absolute terms, or even in relation to past understanding, but in relation to the currently available information. Participants explicitly stated that they calibrated their interpretation of "good understanding" according to their knowledge of the information basis, which differed before and after the explanation phase. The calibration process can be traced by using concepts from the cognitive sciences: Participants i) reported their initial understanding after reading the use case introduction, they then ii) saw the explanations and realized that they had understanding gaps (Rozenblit & Keil, 2002), which they iii) proceeded to locate and close (Keil, 2006), however, they iv) also realized that they could not look at every available explanation and would develop at most a "partial understanding" (Keil, 2019), which they v) rated accordingly in the second self-reports.

5.2.2. Positivity and negativity bias

A complementary explanation involves the "illusion of explanatory depth" (Rozenblit & Keil, 2002): the feeling that one understood a phenomenon with greater precision and coherence than is actually the case. As estimating one's understanding is a notoriously difficult process (Keil, 2006), it is prone to cognitive biases such as positivity and negativity biases. While the exact effects of these biases are still subject to research, studies in social psychology and neuroscience suggest that people recall negative (undesirable, unpleasant) information better, weigh it heavily, and respond more to it (Norris, 2021; Unkelbach et al., 2020), while they process positive (desirable, beneficial) information faster and form stronger associations from it (Unkelbach et al., 2020). The valence attached to information could thus interfere with participants' self-assessments of their understanding. Future work could benefit from incorporating the differentiated dimensions of good and bad understanding as identified in the empirical psychological research literature (Unkelbach et al., 2020).

5.2.3. Additional methods to capture understanding

Previous studies in XAI documented similar effects caused by explanation approaches. Cheng et al. (2019) report on a quantitative evaluation of a white-box explanation design, which, due to its high information density, led to increased "objective" user understanding but decreased self-reported understanding. Papenmeier et al. (2022) found similar mismatches in a quantitative evaluation of their explanation interface, where participants who received *no* explanation gave higher understanding scores than participants who received faithful explanations; a discrepancy our findings might explain. In the same study, similar discrepancies also occurred between participants' self-reports about trusts and observations of their behavior (Papenmeier et al., 2022). In line with these findings, we argue that the calibration effect should be accounted for when measuring understanding, for example, by eliciting an additional metric that captures the perceived scope of available information. A potential reporting question could be "How much information do you feel you currently have about the presented AI system?" combined with a 5-point scale ranging from "very little" to "very much." Self-reported understanding could then be compared with self-reported information scope and verbal responses to acquire a more complete picture. Recent work in XAI has further proposed understanding measurement based on participants' abilities (Speith et al., 2024). This approach appears promising, as the ability to calculate study task 2 was a relevant metric in our study. We thus see eliciting understanding via multiple measures and exploring how these measures can be combined in individual and group settings as a direction for future research.

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

In Section 2.4.2, we described several settings where citizens gather to discuss and form opinions on matters of public interest. These included referendums, forums, and community-based spaces. This paper investigates settings suitable for deliberating the deployment of public AI systems, an issue that we frame as a matter of public interest due to the scope and severity of its potential consequences. Having established that using explanations in group settings benefits participants' understanding, decision confidence, and decision-making processes, it is worthwhile to consider how this approach can be applied in real-world contexts.

5.3.1. Transferability to other domains

Based on our findings, we consider three design elements as readily transferable to other public AI domains: organizing information into four categories, including value- and norm-based justifications, and using a guiding structure for levels of detail and complexity. By contrast, the exact question wording and presentation format presented here are specific to the employment domain and should be tailored to each system's purpose and audience. Further adaptations need to be made when the system in question is a technical or proprietary black-box, as our explanation design were designed for an interpretable system

5.3.2. Educational interventions

Participants repeatedly stated that training in their job agency should employ a similar format of collaboration to educate about AI. Notably, this feedback was provided by domain experts (Groups C and G) and decision subjects (Group E), suggesting that the setting would be suitable for both stakeholder groups in an educational intervention. Similarly, P3 explained that the explanation approach could be helpful if a similar system were implemented in their care facility by embedding it in the team's regular meetings, where difficult cases are discussed and joint decisions are made. These insights align with previous research on XAI in public institutions. Notably, Lee et al. (2019b) and Weitz et al. (2024) conducted participatory workshops to design explanations with end-users in the public sector, finding that co-designing explainable AI helps in considering the needs of both clients and end-users.

5.3.3. Participatory public formats

We envision that collaborative settings and "mini-publics" (Fung, 2003) could be useful in many contexts that aim to strengthen participatory democracy with respect to AI. Potential areas of application could be professional consultation workshops for citizens affected by algorithmic decisions, comparable to legal clinics (Legal Aid Ontario, 2025), community-based education and training interventions, such as "contestation cafés" (Collins et al., 2024), or union forums that inform and organize employees' voices about the use of AI in their institution (Kaur et al., 2022). On a different note, Crivellaro et al. (2019) found that participatory formats that aim to connect communities to public institutions can suffer from a lack of crucial information (e.g., budgets), which could be alleviated by an information structure such as the presented explanation design. In short, we propose that explainable AI in collective settings could be a valuable engagement format for contexts in which public AI has the potential to impact people's lives. Future work could explore how collective XAI settings can be implemented in these contexts as part of responsible AI initiatives, in connection to both institutionalized (Costanza-Chock et al., 2022) and user-based (Shen et al., 2021) auditing practices.

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

5.4.1. Improvements to the explanation design

In Section 4.1.4, we described that participants appreciated the explanations' comprehensive and flexible information selection and self-directed exploration. However, they also noted that the explanations have a high access threshold and that the modular structure makes oversight difficult. A digital version of the explanation design could improve the overview through summaries and navigation while allowing for simple language options and cross-references. As the presentation of the large amount of available information also seemed to overwhelm participants, approaches to reduce the scope and prioritize selection would be beneficial. An example could be a recommendation system that suggests explanations from different categories and levels of detail to participants based on their stated interests and technical knowledge.

5.4.2. Reconsidering levels of detail in explanations

However, the separation of information into levels of detail can be challenging. Our explanations' structure is designed to provide varying levels of detail, allowing us to communicate the priority of information effectively. Higher levels of detail were intended to include more peripheral information, and in the case of *system details*, more technical and thus more complex information. Deciding how the

available information is allocated into this structure requires subjective choices. Some participants stated that the most critical information for them, in fact, resided in level 3 of an explanation, highlighting the differences in assigned priority. Future work should investigate how the structure of information in explanations can be designed to make fundamental information readily available from the start, while providing more detailed information upon user demand. This could contribute insights into balancing information soundness and completeness (Kulesza et al., 2013) with the actual needs of AI novices.

5.4.3. Explanation format

The presented explanation design primarily conveyed information textually, using visual anchors with color-encoding and formatting. Previous work has found mixed effects of textual information on users, including user aversion despite understanding benefits (Szymanski et al., 2021), user aversion and understanding barriers (Schmude et al., 2023), and user appreciation (Schellingerhout et al., 2023; Weitz et al., 2024). Factors such as vocabulary, text length, and user familiarity can all impact the format's effects. In our study, several participants judged the texts to be more complex and lengthy than necessary (Section 4.1.4). An improved explanation design would aim to keep texts as short as possible and to provide a *plain language* option: a specific style of language that ensures that the reader understands as quickly and easily as possible (Cooper, 1989). Furthermore, explanation approaches that build on the design presented here could utilize participatory design methods with AI novices to formulate user requirements for visual and interactive information formats. For instance, previous work has identified visual graph-based information as preferred by some stakeholders (Schellingerhout et al., 2023), which could be helpful in conveying procedural information about system deployment. Interactive explanation formats have been found to benefit user experience (Guesmi et al., 2024) and could be used to show users examples of the system's calculation logic. Implementing the presented explanation design as a digital working prototype that enables these forms of interaction would thus be a fruitful avenue for future research.

5.4.4. Cognitive load of explanations

In Section 4.1.4, we described that participants frequently commented on the high effort of interacting with the explanations, which is likely caused by its design. Since the structure was intended to provide multiple levels of information and allow participants to choose from a selection of topics, the resulting design requires a high degree of repeated user choice. For each choice, participants needed to keep track of what they had already read, what they still wanted to read, and the amount of time remaining. Proceeding through these cognitive steps repeatedly, and being confronted every time with the large selection of explanations, likely leads to a form of decision fatigue: "the impaired ability to make decisions and control behavior as a consequence of repeated acts of decision-making" (Pignatiello et al., 2020). An important design requirement for explanation designs aiming to provide comprehensive information about an AI system is therefore the balancing of conscious thinking effort with the depletion of the users' cognitive resources. Users suffering from decision fatigue tend to use mental shortcuts – cognitive heuristics – to make decisions, which can potentially affect the quality of the decision outcomes (Pignatiello et al., 2020). Importantly, explanations that are intended to support critical decision-making tasks, such as whether to deploy a high-risk AI system, must be crafted to avoid depleting the cognitive resources that users would need for the actual decision-making task. While approaches exist that suggest implementing cognitive-forcing mechanisms (Buçinca et al., 2021), which require users to engage in critical thinking and avoid mental shortcuts, we thus argue that explanation design must prioritize cognitive efficiency as much as user understanding, so as not to defeat the purpose of the explanation. A relevant direction for future work is thus to identify designs that contain the easiest possible explanations for a specific purpose, such as evaluating or contesting an AI system, such that users feel confident to take action but are neither overwhelmed nor unduly biased by over-specific selection of information.

5.5. Summary of explanation design suggestions

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

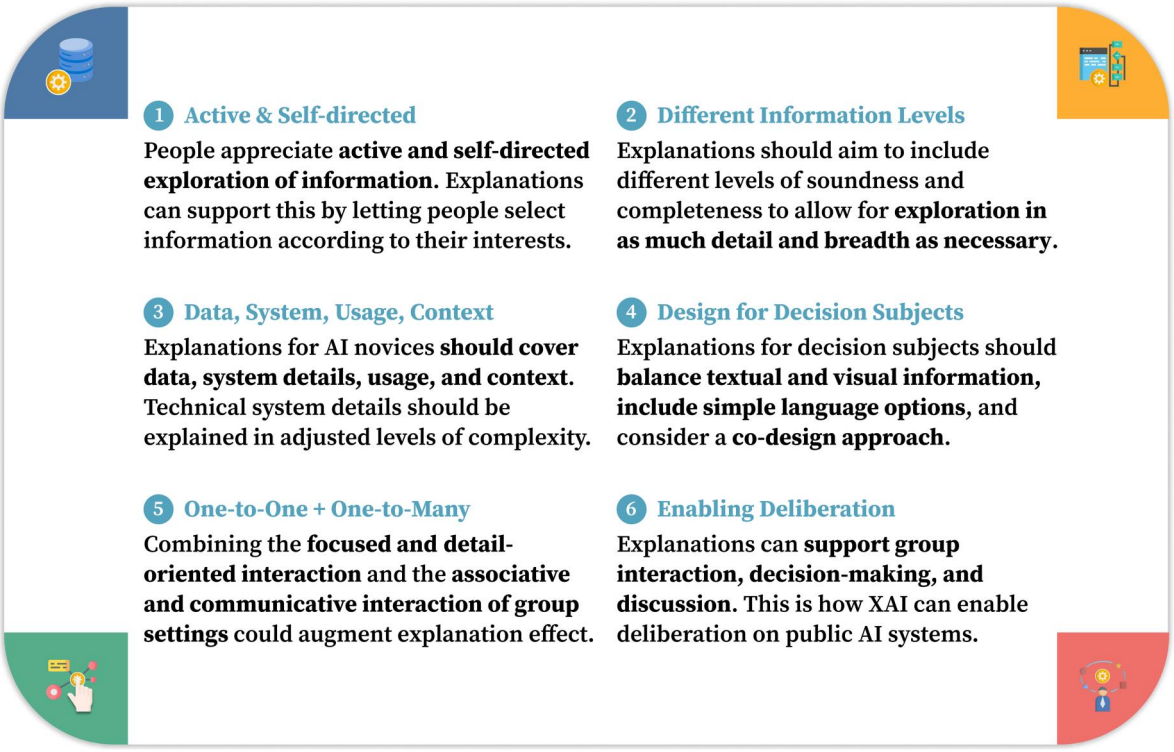


Figure 8. Summary of implications regarding the design of explanations for individual and collaborative settings based on our findings.

6. Limitations

In the following, we list and discuss the limitations of our study.

6.1. Influence of demographics

Due to the limited sample size, we did not analyze the impact of sex and/or gender on our results, limiting the results' external but not internal generalizability regarding these aspects (Maxwell, 2010). We further note that our participant sample is biased toward university education in the single interviews, which we addressed by comparing these participants mostly with university-educated participants in the focus groups.

6.2. Domain specificity

The presented use case is embedded in a specific sociotechnical context (Ehsan et al., 2023a) 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.

6.3. Lack of diversity in verbalizations in single interviews

In comparison to focus group participants, single interview participants exhibited fewer statements and articulations during interaction with the explanations. The prevalence of certain themes and articulations could thus be slightly reduced in comparison to focus groups. To offset this, participants were asked to think-aloud during their interaction and were periodically asked for their perceptions and understanding.

6.4. Recruitment

Since our participants were recruited from organizations and networks in the same geographical region, this could have resulted in regional or cultural biases. We are further aware that the cooperation with civil society organizations in the recruitment of participants could have led to selection biases, especially in the form of convenience sampling (over-representation of readily available participants), self-selection bias (over-representation of strongly motivated participants), and interviewer bias (over-representation of agreeable or compatible participants) (Collier, 1995). Concretely, this could mean that the participants of our study i) were more motivated and engaged, ii) expressed more socially desired viewpoints, iii) over-represented specific subgroups of domain experts and decision subjects, and iv) under-represented the intersectionality of experiences in comparison to a potential random sample. We aimed to mitigate these biases by defining research goals and methods upfront and ensuring the approach was not adjusted post-recruitment based on the participant population. This was achieved by diversifying group composition and reflecting on possible sampling influences in the analysis of results. While recruitment from personal networks can introduce further biases due to the effect of personal relationships on the study outcomes, we note that the participants recruited through these channels were few (Group D, S2, S4, S5, S12) and that these participants did not appear to provide more positive reports nor to avoid providing critical feedback on the explanation design more than other participants. In contrast, Group D articulated several outspoken points of criticism, as is described in Section 4.2.3. Lastly, we note that decision subjects from support organizations are often underrepresented in XAI research due to recruitment and ethical challenges. By partnering with these organizations, we ensured that professionals could make informed decisions about which clients to invite. Although this approach introduces certain biases in the selection, excluding decision subjects would disproportionately limit the study's findings and omit this critical stakeholder group's perspective.

6.5. Approach does not aim for generalizability

This study takes a qualitative approach, meaning that the recruitment strategy further did not aim for statistical generalizability but instead intended to cover a variety of "theoretically relevant cases" (Collier, 1995) and "careful contextualization" (Collier & Mahoney, 1996) to examine our research questions.

7. Conclusion

This paper tested a question-driven, modular explanation design with AI novices in groups and individual settings. We conducted an interview study involving 8 focus groups and 12 single interviews. We analyzed them to examine the effect of explanations on understanding, decision-making, and decision confidence, as well as participants' perceptions of key information and the interaction processes in both settings. We found that explanations supported participants' understanding and decision-making in different ways, encouraging focused interaction in individual settings and shared understanding in group settings. Even though individuals could not exchange with others, the explanations still led to increased decision confidence and changes by supporting internal deliberation. In groups, the explanation design facilitated a set of interactions that allowed participants to support each other's understanding and provided grounds for exchanging arguments about key aspects of the system's deployment. For groups that experienced collaborative failure, we suggest modifying the explanation's design to highlight essential information and measures that create a more productive social dynamic. With this work, we aim to showcase the potential of combining explanations with group settings to enable AI novices to understand and deliberate about public AI systems.

Notes

1. We use the term "AI system" to describe algorithmic systems with machine learning components. The terminology follows research on explainable AI (Langer et al., 2021) and research on AI in the context of society (Collins et al., 2024; Züger & Asghari, 2023) and regulation (Panigutti et al., 2023).
2. We intend the term to mean factual or testable understanding following Cheng et al. (2019) and Bove et al. (2022).

3. Prioritizing group harmony over true argumentation.
4. AMS stands for the Public Employment Agency (Arbeitsmarktservice).
5. The four icons used in this figure and throughout the paper are provided by Freepik (data, usage), Flat Icons (system details), and noomtah (context) through Flaticon.com.
6. The case example was inspired by Allhutter et al. (2020) and adapted to this study, as depicted in Supplementary Appendix A.
7. We note again that with "explanation" we mean a question and answer pair and with "explanations" we mean the collection of all 36 explanations (Section 3).
8. Details on the inspiration for this case example where omitted to adhere to anonymization policy, but will be re-inserted for the final version.

Acknowledgements

We thank Jan Wiegner and Mireia Yurrita for their feedback and support. We further thank Volkshilfe Wien, Wiener Hilfswerk, and Job-TransFair for their support in the realization of this study.

Author contributions

CRediT: **Timothée Schmude**: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft; **Laura Koesten**: Methodology, Supervision, Writing – review & editing; **Torsten Möller**: Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing; **Sebastian Tschiatschek**: Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

AI usage statement

During the final preparation of this manuscript, we utilized ChatGPT and Grammarly for copy-editing. These tools were used solely to improve clarity and readability without altering the paper's intellectual content, methodology, or findings.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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

ORCID

Timothée Schmude http://orcid.org/0009-0006-8276-4670
Laura Koesten http://orcid.org/0000-0003-4110-1759
Torsten Möller http://orcid.org/0000-0003-1192-0710
Sebastian Tschiatschek http://orcid.org/0000-0002-2592-0108

References

Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (2023). *Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute* [Paper presentation]. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23), New York, NY, USA. Association for Computing Machinery (Article 8, p. 16) . https://doi.org/10.1145/3544548.3580984

Allhutter, D., Mager, A., Cech, F., Fischer, F., & Grill, G. (2020). *Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*. Technical Report. Österreichische Akademie der Wissenschaften. epub.oeaw.ac.at.

Ammitzbøll Flügge, A. (2021). *Perspectives from Practice: Algorithmic Decision-Making in Public Employment Services* [Paper presentation]. In Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (pp. 253–255), Virtual Event USA. ACM. https://doi.org/10.1145/3462204.3481787

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science (New York, N.Y.)*, *359*(6373), 325–329. https://doi.org/10.1126/science.aao4408

Baron, R. S. (2005). So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. In *Advances in experimental social psychology* (Vol. 37, pp. 219–253). Elsevier. https://doi.org/10.1016/S0065-2601(05)37004-3

Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemolgy and philosophy of science* (pp. 1–34). Routledge.

Bertrand, A., Eagan, J. R., & Maxwell, W. (2023). *Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making* [Paper presentation]. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23) (pp. 943–958), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3593013.3594053

Biran, O., & McKeown, K. (2017). *Human-Centric Justification of Machine Learning Predictions* [Paper presentation]. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17 (pp. 1461–1467). https://doi.org/10.24963/ijcai.2017/202

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.2307/1175554

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2022). *Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users* [Paper presentation]. In Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22) (pp. 807–819), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3490099.3511139

Bove, C., Laugel, T., Lesot, M.-J., Tijus, C., Detyniecki, M. (2024). *Why do explanations fail? A typology and discussion on failures in XAI*. http://arxiv.org/abs/2405.13474

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). *Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services* [Paper presentation]. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–12), Glasgow Scotland Uk. ACM. https://doi.org/10.1145/3290605.3300271

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Bundesagentur für Arbeit. (2021). *Bearbeiten von Bewerberdaten durch Träger*. arbeitsagentur.de/datei/dok_ba013193.pdf

Byrne, R. M. J. (2023). *Good Explanations in Explainable Artificial Intelligence (XAI): Evidence from Human Explanatory Reasoning* [Paper presentation]. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (pp. 6536–6544), Macau, SAR China. https://doi.org/10.24963/ijcai.2023/733

Capel, T., & Brereton, M. (2023). *What is Human-Centered about Human-Centered AI? A Map of the Research Landscape* [Paper presentation]. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–23), Hamburg Germany. ACM. https://doi.org/10.1145/3544548.3580959

Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The use of triangulation in qualitative research. *Oncology Nursing Forum*, *41*(5), 545–547. https://doi.org/10.1188/14.ONF.545-547

Chatti, M. A., Guesmi, M., Vorgerd, L., Ngo, T., Joarder, S., Ain, Q. U., & Muslim, A. (2022). *Is More Always Better? The Effects of Personal Characteristics and Level of Detail on the Perception of Explanations in a Recommender System* [Paper presentation]. In Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (pp. 254–264), Barcelona Spain. ACM. https://doi.org/10.1145/3503252.3531304

Chen, C., Feng, S., Sharma, A., & Tan, C. (2023). *Machine Explanations and Human Understanding* [Paper presentation]. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT '23) (p. 1), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3593013.3593970

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). *Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders* [Paper presentation]. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–12), Glasgow, Scotland, UK. ACM. https://doi.org/10.1145/3290605.3300789

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2023). *Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk*

Assessment [Paper presentation]. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–18), Hamburg, Germany. ACM. https://doi.org/10.1145/3544548.3581015

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2024). *Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate* [Paper presentation]. Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 103–119), Greenville, SC, USA. ACM. https://doi.org/10.1145/3640543.3645199

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Collier, D. (1995). Translating quantitative methods for qualitative researchers: The case of selection bias. *American Political Science Review*, *89*(2), 461–466. https://doi.org/10.2307/2082442

Collier, D., & Mahoney, J. (1996). Insights and pitfalls: Selection bias in qualitative research. *World Politics*, *49*(1), 56–91. https://doi.org/10.1353/wp.1996.0023

Collins, R. P., Redström, J., & Rozendaal, M. (2024). The right to contestation: Towards repairing our interactions with algorithmic decision systems. *International Journal of Design*, *18*(1), 95–106. https://doi.org/10.57698/V18I1.06

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*(2021), 103503. https://doi.org/10.1016/j.artint.2021.103503

Convertino, G., Billman, D., Pirolli, P., Massar, J. P., & Shrager, J. (2008). The CACHE study: group effects in computer-supported collaborative analysis. *Computer Supported Cooperative Work (CSCW)*, *17*(4), 353–393. https://doi.org/10.1007/s10606-008-9080-9

Cooke, N. J., Cohen, M. C., Fazio, W. C., Inderberg, L. H., Johnson, C. J., Lematta, G. J., Peel, M., & Teo, A. (2024). From teams to teamness: Future directions in the science of team cognition. *Human Factors*, *66*(6), 1669–1680. https://doi.org/10.1177/00187208231162449

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–285. https://doi.org/10.1111/cogs.12009

Cooper, R. L. (1989). *Language planning and social change*. Cambridge University Press.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). *Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem* [Paper presentation]. 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1571–1583), Seoul Republic of Korea. ACM. https://doi.org/10.1145/3531146.3533213

Crivellaro, C., Anderson, R., Lambton-Howard, D., Nappey, T., Olivier, P., Vlachokyriakos, V., Wilson, A., & Wright, P. (2019). Infrastructuring public service transformation: Creating collaborative spaces between communities and institutions through HCI research. *ACM Transactions on Computer-Human Interaction*, *26*(3), 1–29. https://doi.org/10.1145/3310284

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial Intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & Society*, *35*(4), 917–926. https://doi.org/10.1007/s00146-020-00960-w

Desiere, S., & Struyven, L. (2021). Using Artificial Intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy*, *50*(2), 367–385. https://doi.org/10.1017/S0047279420000203

DeVito, M. A., Hancock, J. T., French, M., Birnholtz, J., Antin, J., Karahalios, K., Tong, S., & Shklovski, I. (2018). *The Algorithm and the User: How Can HCI Use Lay Understandings of Algorithmic Systems?* [Paper presentation]. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI EA '18) (pp. 1–6), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3170427.3186320

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021). *Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle* [Paper presentation]. Designing Interactive Systems Conference 2021 (pp. 1591–1602), Virtual Event USA. ACM. https://doi.org/10.1145/3461778.3462131

Donghee Shin, D. (2023). *Algorithms, humans, and interactions: How do algorithms interact with people? Designing meaningful AI experiences* (1st ed.). Routledge. https://doi.org/10.1201/b23083

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2024). *The who in XAI: How AI background shapes perceptions of AI explanations* [Paper presentation]. arXiv:2107.13509 [cs]. https://doi.org/10.1145/3613904.3642474

Ehsan, U., Saha, K., De Choudhury, M., & Riedl, M. O. (2023a). Charting the sociotechnical gap in explainable AI: A framework to address the gap in XAI. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–32. https://doi.org/10.1145/3579467

Ehsan, U., Wintersberger, P., Watkins, E. A., Manger, C., Ramos, G., Weisz, J. D., Daumé Iii, H., Riener, A., & Riedl, M. O. (2023b). *Human-Centered Explainable AI (HCXAI): Coming of Age* [Paper presentation]. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–7), Hamburg Germany. ACM. https://doi.org/10.1145/3544549.3573832

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). *First I "like" It, Then I Hide It: Folk Theories of Social Feeds* [Paper presentation]. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (San Jose, California, USA) (CHI '16) (pp. 2371–2382), New York, NY, USA. In. Association for Computing Machinery. https://doi.org/10.1145/2858036.2858494

European Commission. (2024). Laying down harmonised rules on Artificial Intelligence and amending regulations.
Fiesler, C., Brubaker, J. R., Forte, A., Guha, S., McDonald, N., & Muller, M. (2019). *Qualitative Methods for CSCW: Challenges and Opportunities* [Paper presentation]. Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing (pp. 455–460), Austin TX USA. ACM. https://doi.org/10.1145/3311957.3359428
Freiesleben, T., & König, G. (2023). Dear XAI Community, we need to talk! In L. Longo (Ed.), *Explainable Artificial Intelligence* (pp. 48–65). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44064-9_3
Fung, A. (2003). Survey Article: Recipes for public spheres: Eight institutional design choices and their consequences. *Journal of Political Philosophy*, *11*(3), 338–367. https://doi.org/10.1111/1467-9760.00181
Gamper, J., Kernbeiß, G., Wagner-Pinter, M. (2020). *Das Assistenzsystem AMAS: Zweck, Grundlagen, Anwendung*. https://www.ams-forschungsnetzwerk.at/downloadpub/2020_Assistenzsystem_AMAS-dokumentation.pdf
Grimm, S. R. (2019). *Varieties of Understanding* (pp. 1–14). Oxford University Press. https://doi.org/10.1093/oso/9780190860974.003.0001
Guesmi, M., Chatti, M. A., Joarder, S., Ain, Q. U., Alatrash, R., Siepmann, C., & Vahidi, T. (2024). Interactive explanation with varying level of details in an explainable scientific literature recommender system. *International Journal of Human–Computer Interaction*, *40*(22), 7248–7269. https://doi.org/10.1080/10447318.2023.2262797
Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT Press.
Henin, C., & Le Métayer, D. (2022). Beyond explainability: Justifiability and contestability of algorithmic decision systems. *AI & Society*, *37*(4), 1397–1410. https://doi.org/10.1007/s00146-021-01251-8
Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research*, *27*(4), 591–608. https://doi.org/10.1177/1049732316665344
Hoffman, R. R., Mueller, S. T., Klein, G., Jalaeian, M., & Tate, C. (2023). Explainable AI: Roles and stakeholders, desirements and challenges. *Frontiers in Computer Science*, *5*, 1117848. https://doi.org/10.3389/fcomp.2023.1117848
Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2018). *Das AMS-Arbeitsmarktchancen-Modell*. Arbeitsmarktservice Österreich.
Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196. https://doi.org/10.1145/353485.353487
Innes, J. E., & Booher, D. E. (2003). Collaborative policymaking: Governance through dialogue. In M. A. Hajer, & H. Wagenaar (Eds.), *Deliberative policy analysis* (1 ed., pp. 33–59). Cambridge University Press. https://doi.org/10.1017/CBO9780511490934.003
Janis, I. L. (1971). Groupthink. *Psychology Today*, *5*(6), 43–46.
Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes* (pp. viii, 277–viii, 277). Houghton Mifflin Harcourt.
Johnson, D. W., & Johnson, R. T. (1985). *The internal dynamics of cooperative learning groups* (pp. 103–124). Springer US. https://doi.org/10.1007/978-1-4899-3650-9_4
Karadzhov, G., Stafford, T., & Vlachos, A. (2023). DeliData: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–25. https://doi.org/10.1145/3610056
Kaur, H., Adar, E., Gilbert, E., & Lampe, C. (2022). *Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory* [Paper presentation]. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 702–714), Seoul Republic of Korea. ACM. https://doi.org/10.1145/3531146.3533135
Kawakami, A., Coston, A., Zhu, H., Heidari, H., & Holstein, K. (2024). *The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24) (Article 749, p. 22), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642849
Keil, F. (2019). How do partial understandings work? In *Varieties of understanding* (pp. 191–208). Oxford University Press. https://doi.org/10.1093/oso/9780190860974.003.0010
Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373. https://doi.org/10.1016/S1364-6613(03)00158-X
Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100
Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*(1), 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009
Kramer, M. F., Schaich Borg, J., Conitzer, V., & Sinnott-Armstrong, W. (2018). *When Do People Want AI to Make Decisions?* [Paper presentation]. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (New Orleans, LA, USA) (AIES '18) (pp. 204–209), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3278721.3278752

Krueger, R. A. (2004). *Focus groups: A practical guide for applied research* (3. ed., 6. print. ed.). Sage.
Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). *Principles of Explanatory Debugging to Personalize Interactive Machine Learning* [Paper presentation]. In Proceedings of the 20th International Conference on Intelligent User Interfaces (pp. 126–137), Atlanta Georgia USA. ACM. https://doi.org/10.1145/2678025.2701399
Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). *Too much, too little, or just right? Ways explanations impact end users' mental models* [Paper presentation].2013 IEEE Symposium on Visual Languages and Human Centric Computing (pp. 3–10), San Jose, CA, USA. IEEE. https://doi.org/10.1109/VLHCC.2013.6645235
Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (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*(2021), 103473. https://doi.org/10.1016/j.artint.2021.103473
Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019a). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26. https://doi.org/10.1145/3359284
Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., & Procaccia, A. D. (2019b). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–35. https://doi.org/10.1145/3359283
Legal Aid Ontario. (2025). *What is a legal clinic? - Legal Aid Ontario*. legalaid.on.ca. https://www.legalaid.on.ca/faq/what-is-a-legal-clinic/. [Accessed 03-04-2025].
Liao, Q. V., Gruen, D., & Miller, S. (2020). *Questioning the AI: Informing Design Practices for Explainable AI User Experiences* [Paper presentation]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20) (pp. 1–15), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376590
Liao, Q. V., Pribić, M., Han, J., Miller, S., & Sow, D. (2021). *Question-driven design process for explainable AI user experiences*. arXiv. https://doi.org/10.48550/arXiv.2104.03483
Lim, B. Y., & Dey, A. K. (2009). *Assessing Demand for Intelligibility in Context-Aware Applications* [Paper presentation]. Proceedings of the 11th International Conference on Ubiquitous Computing (Orlando, Florida, USA) (UbiComp '09) (pp. 195–204), New York, NY, USA. In. Association for Computing Machinery. https://doi.org/10.1145/1620545.1620576
Lima, G., Grgic-Hlaca, N., Jeong, J. K., & Cha, M. (2023). *Who Should Pay When Machines Cause Harm? Laypeople's Expectations of Legal Damages for Machine-Caused Harm* [Paper presentation]. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT '23) (pp. 236–246), New York, NY, USA. In. Association for Computing Machinery. https://doi.org/10.1145/3593013.3593992
Long, D., Padiyath, A., Teachey, A., & Magerko, B. (2021). *The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences* [Paper presentation]. In Creativity and Cognition (pp. 1–10), Virtual Event Italy. ACM. https://doi.org/10.1145/3450741.3465264
Lopez, P. (2019). Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. In *Proceedings of the 18th Annual STS Conference* (pp. 289–309). Graz. https://doi.org/10.3217/978-3-85125-668-0-16
Lundberg, S. M., & Lee, S.-I. (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)* (pp. 4768–4777). Curran Associates Inc.
Lupia, A. (2024). By design: How people adapt to cognitive limitations in politics. *Topics in Cognitive Science*, *16*(2), 175–186. https://doi.org/10.1111/tops.12690
Maarten A. Hajer and H. Wagenaar (Eds.). (2003). *Deliberative policy analysis: Understanding governance in the network society*. Cambridge University Press.
Mair, D., Smillie, L., La Placa, G., Schwendinger, F., Raykovska, M., Pasztor, Z., & van Bavel, R. (Eds.), European Commission. (2019). *Understanding our political nature: How to put knowledge and reason at the heart of political decision-making*. Publications Office. https://doi.org/10.2760/374191
Mandviwalla, M., & Olfman, L. (1994). What do groups need? A proposed set of generic groupware requirements. *ACM Transactions on Computer-Human Interaction*, *1*(3), 245–268. https://doi.org/10.1145/196699.196715
Maxwell, J. A. (2010). Using numbers in qualitative research. *Qualitative Inquiry*, *16*(6), 475–482. https://doi.org/10.1177/1077800410364740
Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, *33*(2), 243–258. https://doi.org/10.1111/j.1467-9221.2012.00873.x
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*(2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007
Mohseni, S., Zarei, N., Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. arXiv:1811.11839 [cs.HC] https://arxiv.org/abs/1811.11839

Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3rd ed.). https://christophm.github.io/interpretable-ml-book/

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, *4*(3), 231–248. https://doi.org/10.1080/135467898394148

Myers, C. W., Cooke, N. J., Gorman, J. C., & McNeese, N. J. (2024). Introduction to the emerging cognitive science of distributed human-autonomy teams. *Topics in Cognitive Science*, *16*(3), 377–390. https://doi.org/10.1111/tops.12744

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendation: When design meets trust calibration. *World Wide Web*, *24*(5), 1857–1884. https://doi.org/10.1007/s11280-021-00916-0

Naiseh, M., Jiang, N., Ma, J., & Ali, R. (2020). Personalising explainable recommendations: Literature and conceptualisation. In Á. Rocha, H. Adeli, L. Reis, S. Costanzo, I. Orovic, & F. Moreira (Eds.), *Trends and innovations in information systems and technologies*. WorldCIST 2020. Advances in intelligent systems and computing (vol. 1160; pp. 518–533). Springer. https://doi.org/10.1007/978-3-030-45691-7_49

Naiseh, M., Webb, C., Underwood, T., Ramchurn, G., Walters, Z., Thavanesan, N., & Vigneswaran, G. (2024)., 17/07/24 - 19/07/24). *XAI for group-AI interaction: Towards collaborative and inclusive explanation* [Paper presentation]. World Conference for Explainable Artificial Intelligence. https://eprints.soton.ac.uk/493227/

Narayanan, R., Cohen, M. C., Feigh, K. M., & Cooke, N. J. (2025). Two sides of the same coin? Joint perspectives from shared mental models and interactive team cognition theories on human-AI team cognition. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *69*(1), 412–417. https://doi.org/10.1177/10711813251358788

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126–132. https://doi.org/10.1038/s41562-017-0273-4

Niculae, V., & Danescu-Niculescu-Mizil, C. (2016). *Conversational Markers of Constructive Discussions* [Paper presentation]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (pp. 568–578), San Diego, California. https://doi.org/10.18653/v1/N16-1070

Niklas, J., Sztandar-Sztanderska, K., & Szymielewicz, K. (2015). *Profiling the unemployed in Poland: Social and political implications of algorithmic decision making*. panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf

Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When is it better to learn together? Insights from research on collaborative learning. *Educational Psychology Review*, *27*(4), 645–656. https://doi.org/10.1007/s10648-015-9312-8

Norris, C. J. (2021). The negativity bias, revisited: Evidence from neuroscience measures and an individual differences approach. *Social Neuroscience*, *16*(1), 68–82. https://doi.org/10.1080/17470919.2019.1696225

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). *The Role of Explainable AI in the Context of the AI Act. In* [Paper presentation]. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT '23) (pp. 1139–1150), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3593013.3594069

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction*, *29*(4), 1–33. https://doi.org/10.1145/3495013

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed., p. 532). Sage Publications, Inc.

Pignatiello, G. A., Martin, R. J., & Hickman, R. L. (2020). Decision fatigue: A conceptual analysis. *Journal of Health Psychology*, *25*(1), 123–135. https://doi.org/10.1177/1359105318763510

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). *The Fallacy of AI Functionality* [Paper presentation]. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 959–972), Seoul Republic of Korea. ACM. https://doi.org/10.1145/3531146.3533158

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "*Why Should I Trust You?*": *Explaining the Predictions of Any Classifier* [Paper presentation]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16) (pp. 1135–1144), New York, NY, USA. In. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Sato, B. K., Hill, C. F. C., & Lo, S. M. (2019). Testing the test: Are exams measuring understanding? *Biochemistry and Molecular Biology Education: a Bimonthly Publication of the International Union of Biochemistry and Molecular Biology*, *47*(3), 296–302. https://doi.org/10.1002/bmb.21231

Schellingerhout, R., Barile, F., & Tintarev, N. (2023). A co-design study for multi-stakeholder job recommender system explanations. In *World Conference on Explainable Artificial Intelligence* (pp. 597–620). Springer. https://doi.org/10.1007/978-3-031-44067-0_30

Schmude, T., Koesten, L., Möller, T., & Tschiatschek, S. (2023). *On the Impact of Explanations on Understanding of Algorithmic Decision-Making* [Paper presentation]. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA) (FAccT '23) (pp. 959–970), New York, NY, USA. In. Association for Computing Machinery. https://doi.org/10.1145/3593013.3594054

Schmude, T., Koesten, L., Möller, T., & Tschiatschek, S. (2025). Information that matters: Exploring information needs of people affected by algorithmic decisions. *International Journal of Human-Computer Studies*, *193*, 103380. https://doi.org/10.1016/j.ijhcs.2024.103380

Scott, K. M., Wang, S. M., Miceli, M., Delobelle, P., Sztandar-Sztanderska, K., & Berendt, B. (2022). *Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective* [Paper presentation]. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) (2138–2148), Seoul Republic of Korea. ACM. https://doi.org/10.1145/3531146.3534631

Shen, H., DeVos, A., Eslami, M., & Holstein, K. (2021). Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–29. https://doi.org/10.1145/3479577

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2023). 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*, *39*(7), 1455–1482. https://doi.org/10.1080/10447318.2022.2095705

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science (New York, N.Y.)*, *323*(5910), 122–124. https://doi.org/10.1126/science.1165919

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

Speith, T., Crook, B., Mann, S., Schomäcker, A., & Langer, M. (2024). Conceptualizing understanding in explainable artificial intelligence (XAI): An abilities-based approach. *Ethics and Information Technology*, *26*(2), 40. https://doi.org/10.1007/s10676-024-09769-3

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation*, *3*(1). https://doi.org/10.16997/jdd.50

Sutcliffe, A. (2005). *Applying small group theory to analysis and design of CSCW systems* [Paper presentation]. In Proceedings of the 2005 Workshop on Human and Social Factors of Software engineering - HSSE '05 (pp. 1–6), St. Louis, Missouri. ACM Press. https://doi.org/10.1145/1083106.1083119

Swiss Confederation. (2025). *The referendum*. ch.ch/en/votes-and-elections/referendum

Szigetvari, A. (2018). AMS bewertet Arbeitslose künftig per Algorithmus. *Der Standard*. https://www.derstandard.at/story/2000089095393/ams-bewertet-arbeitslose-kuenftig-per-algorithmus

Szymanski, M., Millecamp, M., & Verbert, K. (2021). *Visual, textual or hybrid: The effect of user expertise on different explanations* [Paper presentation]. 26th International Conference on Intelligent User Interfaces (pp. 109–119), College Station TX USA. ACM. https://doi.org/10.1145/3397481.3450662

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy Artificial Intelligence. *Electronic Markets*, *31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Unkelbach, C., Alves, H., & Koch, A. (2020). Chapter three - Negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. In *Advances in experimental social psychology* (Vol. 62, pp. 115–187). Academic Press. https://doi.org/10.1016/bs.aesp.2020.04.005

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Wang, X., & Yin, M. (2021). *Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making* [Paper presentation]. 26th International Conference on Intelligent User Interfaces (IUI '21) (pp. 318–328), New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3397481.3450650

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Weiss, R. S. (1995). *Learning from strangers: The art and method of qualitative interview studies* (1. free press paperback ed.). Free Press.

Weitz, K., Schlagowski, R., André, E., Männiste, M., & George, C. (2024). *Explaining It Your Way - Findings from a Co-Creative Design Workshop on Designing XAI Applications with AI End-Users from the Public Sector* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24) (Article 745, p. 14), New York, NY, USA. In. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642563

Wenzelburger, G., König, P. D., Felfeli, J., & Achtziger, A. (2024). Algorithms in the public sector. Why context matters. *Public Administration*, *102*(1), 40–60. https://doi.org/10.1111/padm.12901

Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. *Data & Policy*, *5*, e2. https://doi.org/10.1017/dap.2022.39

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (expanded 2nd ed.). Association for Supervision and Curriculum Development.

Xu, W. (2019). Toward human-centered AI: A perspective from human–computer interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Zagzebski, L. (2019). Toward a theory of understanding. In *Varieties of understanding* (pp. 123–136). Oxford University Press. https://doi.org/10.1093/oso/9780190860974.003.0007

Züger, T., & Asghari, H. (2023). AI for the public. How public interest theory shifts the discourse on AI. *AI & Society*, *38*(2), 815–828. https://doi.org/10.1007/s00146-022-01480-5

About the authors

Timothée Schmude is a PhD student at the University of Vienna. He holds master's degrees in Information Processing and Professional Writing from the University of Cologne. His research focuses on human-centered explainable AI, using empirical and interdisciplinary approaches to make algorithmic decision systems understandable, contestable, and subject to public deliberation.

Laura Koesten is an assistant professor at MBZUAI, affiliated with the University of Vienna, and senior scientist at AIT. Her research examines how people engage with data, sensemaking, discovery and reuse, and collaborative data work. She received her PhD from University of Southampton and won the 2024 Hedy Lamarr Prize.

Torsten Möller is a professor of computer science at the University of Vienna. Previously, he was a faculty member at Simon Fraser University, received his PhD from Ohio State University and a Vordiplom from Humboldt University of Berlin. His research covers topics at the intersection of data analysis, visualization, computer graphics, and HCI.

Sebastian Tschiatschek is Associate Professor for Machine Learning at the University of Vienna. Previously, he was a senior researcher at Microsoft Research Cambridge and a postdoctoral researcher at ETH Zurich. He received his PhD from Graz University of Technology, and specializes in machine learning, structured data, and human–machine interaction.