# On the Impact of Explanations on Understanding of Algorithmic Decision-Making


Timothée Schmude
timothee.schmude@univie.ac.at
University of Vienna, Faculty of Computer Science, Research Network Data Science, UniVie Doctoral School Computer Science DoCS
Vienna, Vienna, Austria

Laura Koesten
laura.koesten@univie.ac.at
University of Vienna, Faculty of Computer Science, Research Group Visualization and Data Analysis
Vienna, Vienna, Austria

Torsten Möller
torsten.moeller@univie.ac.at
University of Vienna, Faculty of Computer Science, Research Network Data Science, Research Group Visualization and Data Analysis
Vienna, Vienna, Austria

Sebastian Tschiatschek
sebastian.tschiatschek@univie.ac.at
University of Vienna, Faculty of Computer Science, Research Network Data Science, Research Group Data Mining and Machine Learning
Vienna, Vienna, Austria



## ABSTRACT

Ethical principles for algorithms are gaining importance as more and more stakeholders are affected by "high-risk" algorithmic decision-making (ADM) systems. *Understanding* how these systems work enables stakeholders to make informed decisions and to assess the systems' adherence to ethical values. Explanations are a promising way to create understanding, but current explainable artificial intelligence (XAI) research does not always consider existent theories on how understanding is formed and evaluated. In this work, we aim to contribute to a better understanding of understanding by conducting a qualitative task-based study with 30 participants, including users and affected stakeholders. We use three explanation modalities (textual, dialogue, and interactive) to explain a "high-risk" ADM system to participants and analyse their responses both inductively and deductively, using the "six facets of understanding" framework by Wiggins & McTighe [63]. Our findings indicate that the "six facets" framework is a promising approach to analyse participants' thought processes in understanding, providing categories for both rational and emotional understanding. We further introduce the "dialogue" modality as a valid explanation approach to increase participant engagement and interaction with the "explainer", allowing for more insight into their understanding in the process. Our analysis further suggests that individuality in understanding affects participants' perceptions of algorithmic fairness, demonstrating the interdependence between understanding and ADM assessment that previous studies have outlined. We posit that drawing from theories on learning and understanding like the "six facets" and leveraging explanation modalities can guide XAI research to better suit explanations to learning processes of individuals and consequently enable their assessment of ethical values of ADM systems.



## CCS CONCEPTS

• **Human-centered computing** → **Field studies**.

## KEYWORDS

XAI, learning Sciences, algorithmic decision-making, algorithmic fairness, qualitative methods

**ACM Reference Format:**
Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2023. On the Impact of Explanations on Understanding of Algorithmic Decision-Making. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3594054


## 1 MOTIVATION

"Algorithmic decision-making" (ADM) systems analyse data to derive information used to support or facilitate decisions [20]. As such, they are increasingly used in public institutions and administration and thus affect our daily lives. Examples include systems for recidivism prediction in criminal justice [15], refugee resettlement advice in immigration policy [7], and employability estimation in public employment [2, 49]. The EU classifies ADM systems that are used to decide over human individuals as "high-risk" and proposes to regulate them strictly [16, 59], for example by prescribing adherence to standards of "trustworthy artificial intelligence" (TAI) [39]. These standards state that a system should be, among other criteria, *transparent*, *fair*, *accountable*, and have *human oversight*, in order to be deemed "trustworthy". However, two problems pose a challenge in fulfilling these criteria: First, no definition is given of when a system is, for example, transparent or fair, and second, a system *can be* transparent or fair, without being *perceived* as such [31].

How individuals perceive a system's ethical values depends not only on the system's characteristics, but also on individual factors, such as personality traits and demographics [42, 52], as well as on the relation between the individual stakeholder and the ADM




*FAccT '23, June 12–15, 2023, Chicago, IL, USA*



ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594054

system [26, 30, 54]. Stakeholders that are involved in a systems' development, deployment, day-to-day usage, or regulation are known to have very different information needs and priorities in assessing ADM systems [8, 26, 30, 32]. For example, while an ADM system can produce benefits for an employer, such as informing employee decisions [7] and reducing costs [37], the same system can negatively impact stakeholders that are the decision targets by discriminating against certain population groups [13, 64], thus creating two divergent perspectives.

Explanations can aid different stakeholders in acquiring a basic *understanding* of ADM systems in order to "assess" them in terms of ethical values [30, 54]. To this end, numerous studies in Explainable Artificial Intelligence (XAI) examine how people's understanding of a system can be increased by using e.g., input influence, sensitivity, counterfactuals, case-based, and white box explanations [14, 18, 51, 57, 61]. Further, understanding can be affected by the *explanation modality*, meaning the presentation of information in e.g., textual, visual, and interactive form [14, 57]. How to create an explanation that will address every stakeholder's individual information needs, however, remains an open challenge. [18, 52].

To acquire a concept of the mental processes involved in understanding, we employ definitions that are established and used in the learning sciences research. Wiggins & McTighe [63], by referring to Bloom's widely-known "taxonomy of educational objectives" [4, 11], suggest that understanding is essentially *transfer*: "to take what we know and use it creatively, flexibly, fluently, in different settings or problems, on our own". Students can demonstrate their understanding by showing their ability to *perform* specific things with their knowledge, which Wiggins & McTighe [63] describe as the "six facets of understanding". Novel explanation methods could benefit significantly from adopting theories such as the "six facets" framework from the learning sciences in order to better construct and evaluate explanations along existent conceptualisations of understanding and learning. Recent XAI studies already began to leverage different theories of understanding to this end [27, 28].

In this paper, we follow up on these approaches by investigating how a one-on-one explanation presented in three modalities (textual, dialogue, an interactive) creates understanding in different individuals. To this end, we conducted a task-based qualitative study with 30 participants. We analyse their responses using both inductive and deductive approaches, leveraging the "six facets of understanding" [63] to examine if participants can *explain*, *interpret*, *apply*, *empathise*, *take perspective*, and *self-reflect* after receiving the explanation. We further provide a practical analysis of the assumption that understanding is a prerequisite for ethical assessment [30, 54], by relating participants' ability to discuss the fairness of algorithmic decisions to their understanding. As a case study for high-risk ADM, we use the *AMS algorithm*– a system that was planned to be deployed in Austria to predict job-seekers' employability but which was stopped before its actual implementation [3]. The *AMS algorithm* represents a high-risk system that incited public discourse when it was planned [2] and which generalises to other ADM systems used in "Public Employment Services" [49] due to the prevalence of individual scoring based on personal attributes.

We are guided by the following research question, including three sub-questions:

- **How does a "global" explanation using textual, dialogue, and interactive modality impact participants' understanding?**
  - RQ1: Which "facets of understanding" emerge in the responses of participants after receiving the explanation?
  - RQ2: How is the explanation modality correlated to understanding?
  - RQ3: To what degree do participants demonstrate the ability to engage in meaningful discourse about the algorithm, for instance in evaluating the algorithms' fairness with regard to decisions about job-seekers?

Our findings demonstrate that the explanation method chosen for this study successfully gives participants the opportunity to articulate their thought processes that underlie their understanding of the algorithm. To capture and evaluate these thought processes, we highlight the utility of leveraging learning sciences frameworks, such as the "facets of understanding" [63], to gain insight on participants' understanding. Pertaining to the study design, we validate the one-on-one explanation and interview setup which allows for the gathering of "evidence based on response processes" [5] and for an "interactive dialogue" which can "more fully capture understanding" [46]. Lastly, we observe that participants can successfully articulate fairness assessments of the given ADM system after proceeding through the explanation, which however vary in detail and argumentative reasoning depending on participants' understanding. We thus practically illustrate the link between understanding and assessing a system's ethical values as posited conceptually in recent XAI studies [30, 54].

## 2 RELATED WORK

### 2.1 Relevance of algorithmic decision-making

In this work, we focus on the *AMS algorithm*, a system that is used for "algorithmic decision-making" (ADM) i.e., processing data to support or drive decisions in a public institution [13, 14, 59, 61]. high-risk ADM systems [16] are increasingly used throughout all countries and sectors, including the COMPAS[1] model to score defendants' recidivism probability in US courts [15], the GeoMatch refugee resettlement algorithm [7], the Dutch, German, and Austrian classification systems for public employment [49], and systems to decide on child welfare services [13]. Many of these systems suffer shortcomings, including unreliability of predictions [13], a lack of transparency [17], missing stakeholder involvement [49], and biased training data [15], resulting in negative effects on larger parts of the population due to ADM. Current literature further shows that ADM systems rarely comply with standards such as "trustworthy artificial intelligence" [39] or "value-based engineering" [55] for multiple reasons [6, 9, 17, 18, 36]. This is critical since research into the relationship between the use of high-risk ADM systems and societal values suggests that perceptions of such systems as unequal, untrustworthy, or unjust can erode trust in democratic institutions if a large number of people is affected [10, 21–23, 40, 43]. Explanations of ADM systems are seen as one of the possible solutions to these problems, as they in theory lead to more transparency and thus more trustworthy systems [39].

[1] Correctional Offender Management Profiling for Alternative Sanctions.

## 2.2 Stakeholders and explanation design

We orient our explanation approach towards the high-level goals of explainable artificial intelligence (XAI), which as a research field is dedicated to "amend the lack of understanding of AI-based systems" to enable different groups of people to assess whether a system's output is, e.g., accurate, fair, just, or beneficent [30, 54]. However, the degree of understanding that explanations produce has been shown to vary depending on *who* the explanation's recipient is. XAI literature defines individuals involved in the development, deployment, regulation, or use of an ADM system as "stakeholders". Stakeholders have different information needs and attitudes depending on their relation to the system [8, 30]. For example, a "deployer" might expect an explanation to tell them whether the system can inform employee decisions [7] and reduce costs [37], while affected stakeholders might expect to learn if the system discriminates against certain population groups [13, 64].

We use three different explanation *modalities* to present information: textual, dialogue, and interactive modality. In this we are guided by several works that find that explanation modalities can vary in their impact on understanding [14, 57]. We are further guided by works featuring in-person empirical studies [31, 32, 41, 49, 51, 64], as they enable a direct interaction with the participant and, in our case, the introduction of the dialogue explanation modality, in which information is conveyed verbally. A flow chart serves as the basis for all three explanation modalities, as it can depict the complete algorithmic decision-making process [29], including the "human-in-the-loop", an individual overseeing the algorithm and an essential factor in many ADM analyses [9, 18, 32, 64]. Using Speith's [54] taxonomy of explanation methods, our explanation flowchart can be described as a both result- and functioning-focused, model-specific explanation with a visual output format that aims to globally explain the whole decision-making process [54]. The three modalities build upon this base form and add information via textual, verbal, and interactive presentation.

## 2.3 Building and assessing understanding

The purpose of an explanation arguably entails producing understanding in the explanation's recipient. Many studies discuss and analyse how explanations can influence participants' understanding [6, 44, 50, 51, 57], but what *constitutes* understanding and how it can be evaluated is not always discussed. Seeing this as preliminary to our discussion, we will provide a brief description of how understanding is discussed in the learning sciences, before introducing Wiggins' & McTighe's [63] framework on understanding and outlining the "six facets of understanding".

In Anderson's and Krathwohl's [4] revised version of Bloom's [11] taxonomy of educational objectives, understanding is lined up as one of six categories in the "cognitive process dimension", which is counterposed with the "knowledge dimension" to produce the "cognitive" taxonomy of learning objectives. Bloom's original taxonomy was later complemented by the "affective" and "psychomotor" domains; this separation however was criticised "because it isolates aspects of the same objective – and nearly every cognitive objective has an affective component" [4]. Wiggins and McTighe's [63] framework is based on the revised taxonomy of educational objectives, but focuses on the process of understanding. In this work, we therefore use the definition of understanding given by Wiggins and McTighe [63].

According to Wiggins and McTighe [63], when someone truly understands a topic, they can: a) "explain", generalise, and make connections; b) "interpret", translate, or make the subject personal through analogies or anecdotes; c) "apply" or "do" the subject in different contexts; d) "take perspectives" on the topic and see the big picture; e) "empathise" with values that others might find odd and perceive sensitively; and f) "self-reflect" on their own beliefs and habits that shape and impede understanding [63]. This list of "understanding facets" aims towards "transferability" of knowledge [63]. We utilise Wiggins' & McTighe's [63] framework for multiple reasons: First, it includes both the cognitive and affective domain, as well as a notion of "metacognition" [48] – meaning to reflect on one's knowledge and understanding. Second, their framework is well applicable to our study design, allowing us to compare our inductive analysis of understanding in participants' responses with a deductive approach. Third, Kawakami et al. [28] outline a concept of using this framework to produce "learner-centered" XAI, which we follow up on by transferring the theoretical considerations into an empirical study. Lastly, Wiggins & McTighe [63] also provide a categorisation of "barriers to understanding", which we adapt to our use case, consisting of: i) forgetting, ii) being unable to use what we learn, and iii) not knowing that we do not understand.

We further base our one-on-one interview study design on studies from the learning sciences, which posit that "evidence based on response processes" [5] allows us to observe how understanding emerges in the learner's responses and to distinguish it from knowledge or recall [12]. We further include a task in our study design where participants are asked to explain the *AMS algorithm* in their own words, drawing from Duckworth et al. [19], who underline that letting learners explain in their own words provides insight into their understanding.

## 2.4 Analysing perceptions of algorithmic fairness

Fairness is seen as one of the central criteria for "trustworthy AI" [39] or "ethical AI" [21]. Similar to other high-level criteria, the meaning of fairness as a moral value shifts depending on who is asked [26, 42, 52, 64], and whether it applies to a human or to a machine [31, 56]. In this work, we focus on what Langer et al. [30] call the "epistemic" satisfaction of fairness, meaning that we examine whether participants can engage in discourse about the fairness of the *AMS algorithm* after receiving our explanation. In contrast, we are *not* focussing on the "substantial" satisfaction of fairness [30], meaning our discussion will not cover whether the *AMS algorithm* actually acts fairly or not. We thus use the fairness assessment as an indicator of whether the explanation served to increase participants' understanding and enabled them to assess the system in ethical terms.

# 3 METHODOLOGY

To gain insight into stakeholders' understanding of ADM, we conducted a task-based qualitative study with 30 participants using

three explanation modalities of the *AMS algorithm* (textual, dialogue, and interactive). In the study, an explanation of the algorithm (Section 3.2.1) was followed by two tasks (Section 3.2.2) and a short interview about the deployment of the algorithm in society. We analyse participants' responses inductively and deductively, using the "six facets of understanding" framework [63]. An overview of the study procedure is depicted in Figure 3. See Section Table 1 for the participant sample.

## 3.1 The algorithm

In our study, we used the *AMS algorithm*[2] as a prototypical example of an algorithmic decision-making system in Public Employment Services [49]. The algorithm was developed between 2015 and 2021 by a private company for the Public Employment Agency and was piloted for a short time in the autumn of 2018, but was never used as a live system and is currently put on hold due to legal objections [3]. The case has been covered by several academic studies and incited public discourse over the benefits and risks of its deployment [1, 3, 35, 49].[3] As a great number of people could be affected by the implementation of such a system, and as the stakes generalise well to other high-risk settings, we use the *AMS algorithm* as a case example for our study.

*The algorithm's predictions and model.* The *AMS algorithm* was constructed to assign job-seekers to one out of four categories ("high", "medium", or "low" employment chances, plus special cases), depending on their personal attributes, such as age, gender, and education. Every prediction would be confirmed or corrected by an employee of the Public Employment Agency. The groups of employability were defined as follows:

- "Medium": Job-seekers would receive regular support measures[4] to improve their chances of finding employment.
- "High": Job-seekers were expected to find new employment quickly and would receive fewer support measures.
- "Low": Job-seekers were expected to require more support and would be referred to another facility.
- Other: Teenagers, people with disabilities, and people over 50 would receive additional support measures independent of their employability scoring [3].

The algorithm's model was trained on several years of job-seekers' data, mainly consisting of personal attributes (features): gender, age, citizenship, education, impairment, obligations of care (only women), occupational group, prior occupations; as well as a representation of the local job market.[5] People with similar personal attributes would be grouped and compared to the "standard group" [24] – young men with secondary school education – to be assigned a short-term[6] and long-term[7] employability score. For further details please refer to the supplementary material and Allhutter et al. [3].

*Biases in the algorithm.* The weighting of personal features in the algorithm's classification can be considered biased in that several attributes such as gender and nationality led to decreased employability predictions [24]. However, these biases would in theory lead to higher support measures for job-seekers possessing these attributes, in effect supporting those that were potentially disadvantaged in the job market [25]. Nonetheless, Allhutter et al. [2] point out how the algorithm's practical implementation could have detrimental effects on the support and treament that job-seekers would receive. In summary, the *AMS algorithm* is an example of how personal and systemic considerations can lead to value conflicts and ethical dilemmas in algorithmic decision-making, which is why we chose it as a suitable example for the study. Explanation and task examples were chosen such that these issues were brought to the participants' attention, without however preempting any judgement or value statement.

## 3.2 Study setup

*3.2.1 Explanation modalities.* We used a between-subjects study design, showing each participant one of the three distinct explanation modalities (textual, dialogue, interactive) of the *AMS algorithm*. All modalities built upon a visual flowchart of the algorithmic decision-making pipeline, depicted in Figure 1[8]. In the explanation, the fictional job-seeker *Hannah* reports to the Employment Agency and is assigned to group "medium" by the algorithm, which is confirmed by the employee. The modalities, depicted in Figure 2, enriched the flowchart with additional information, which was identical between modalities but was presented in different ways[9]:

- **Textual**: Textual descriptions were added to the basic flowchart in the form of comments. Participants continued through the explanation slides at their own pace and asked questions at the end.
- **Dialogue**: The study examiner verbally added information to the flowchart, explaining depicted components in each slide. Participants could ask questions throughout the whole process.
- **Interactive**: The flowchart was implemented as a simple interactive web version that featured buttons which showed textual descriptions when clicked. Participants could ask questions at the end.

We chose these modalities to achieve multiple aims: First, showing the whole "global" decision-making process without requiring participants to have technical or specific domain knowledge, following the setups of Wang & Yin [61] and Logg et al. [34]. Second, giving learners the opportunity to close gaps in their understanding by asking questions ("inquiring") and interacting verbally in the dialogue modality, which can lead to more effective understanding [33, 45, 46, 53]. Third, examining the contrast between static and interactive interfaces, motivated by Cheng et al. [14], who found that interactive interfaces led to increased understanding.

*3.2.2 Tasks: predicting employability and explaining the algorithm.* Participants had to complete two task sections that were meant to

[2]The abbreviation AMS stands for the Public Employment Agency.
[3]An extended discussion of the public discourse and perception of the *AMS algorithm* is provided in Section A in the supplementary material.
[4]Such as further training or application coaching.
[5]Attributes are listed in detail in Section A in the supplementary material.
[6]At least 90 days unsupported employment in seven months. Unsupported employment is not subsidised by the Employment Agency.
[7]At least 180 days unsupported employment in 24 months.
[8]Please note that alt-text for Figures 1 and 2 is provided in Section G of the supplementary material.
[9]For a more detailed description of modalities confer Section C in the supplementary material.

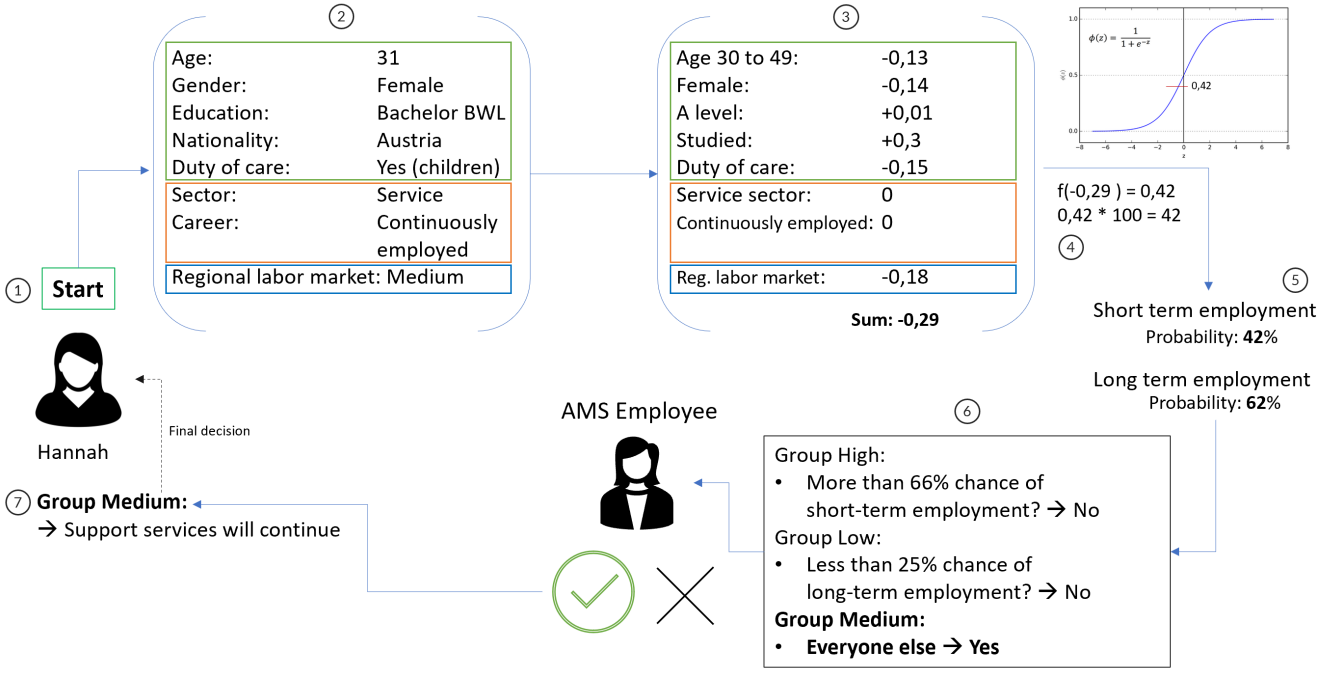


**Figure 1: The flowchart that every explanation modality built upon. The different parts of the explanation are presented sequentially according to the indicated numbering. ① *Hannah* is a fictional job-seeker reporting to the Employment Agency. ② Some of her personal attributes are recorded for the employability prediction. ③ *Hannah's* attributes are compared to the "standard group": young men with secondary school education. Based on this comparison, the weight and value of *Hannah*'s attributes are calculated. ④ The sum of these values is put into a logistic regression function that maps it to a scale of 0 to 1. Multiplied by 100 this gives the short-term employability chance of *Hannah*. ⑤ The long-term chance is calculated in a similar fashion with a different model. ⑥ According to three simple rules, *Hannah* is assigned to one of three groups, which is then confirmed or corrected by the employee of the agency. ⑦ *Hannah* receives the final decision and group assignment.**

probe their understanding and fairness assessment of the algorithm, as depicted in Figure 3.[10]

In the first task section, participants were presented with three example cases of job-seekers and were asked to (1) propose measures that could help the person find work again and (2) estimate their chances for short-term and long-term employment[11]. Participants then received the algorithmic decision for the job-seeker and the employee's decision (accepting or correcting the algorithm's scoring plus any additional measures), and indicated whether they perceived the (3) algorithmic and (4) human decision as fair.

The second task section was split into two: In the first sub-task (2.1), participants were provided with a job-seeker case and were asked to explain to the study examiner how the algorithm would handle the case. In the second sub-task (2.2), participants received a case similar to the first one but ranked higher in terms of employability. Participants should then indicate why the two job-seekers were classified differently.[12]

[10]Example cases are described in detail in Section D and E in the supplementary material. Cases were taken from a detailed report on the *AMS algorithm* by Allhutter et al. [3].
[11]Short-term in the *AMS algorithm* is defined as being employed at least 90 days in the next seven months, long-term as at least six months in the next two years
[12]The first case was female and had duties of care, while the second was male and did not have duties of care (for further detail see supplementary material).

## 3.3 Analysis

For our thematic analysis, we applied two levels of qualitative coding to the data: inductive analysis in the first pass, and deductive analysis in the second.[13]

In our first pass, we examined the data for understanding and perceptions of fairness, letting the overarching themes and theory emerge from the data, following Thomas [58]. We created 18 categories capturing different aspects of understanding, such as the level of detail (e.g., whether participants attended more to the technical details or the societal consequences), the connection to personal knowledge and experience, emotional engagement, and instances in which understanding was impeded. We created 18 more categories to capture participants' perceptions of algorithmic fairness, including statements about the algorithm's precision, the importance of the human in ADM, and issues of inequality. We then compared statements addressing algorithmic fairness with the four dimensions of algorithmic fairness perceptions by Starke et al. [56] to find intersections and create overarching themes.

In our second pass, we produced six deductive code categories according to the "six facets of understanding" framework by Wiggins and McTighe [63], which we split, according to the sub-processes

[13]Both approaches are documented in the supplementary material.

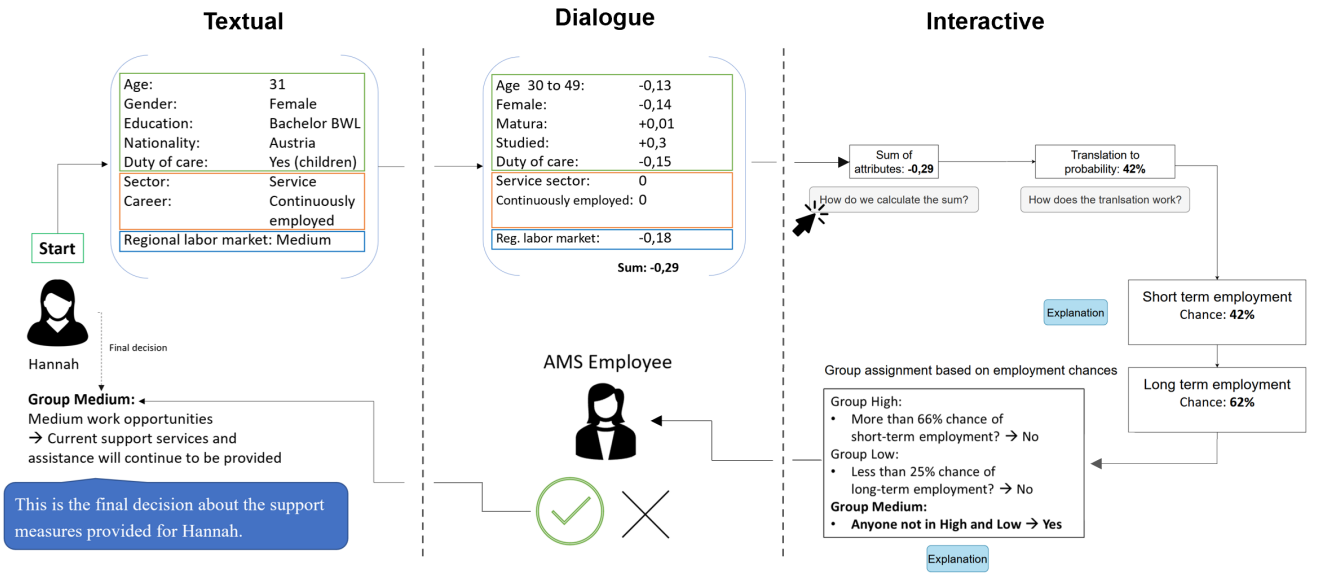


**Figure 2: The flowchart split into all three modalities: textual, dialogue, and interactive. Information was added by textual comments, verbal comments, and interactive controls, respectively.**

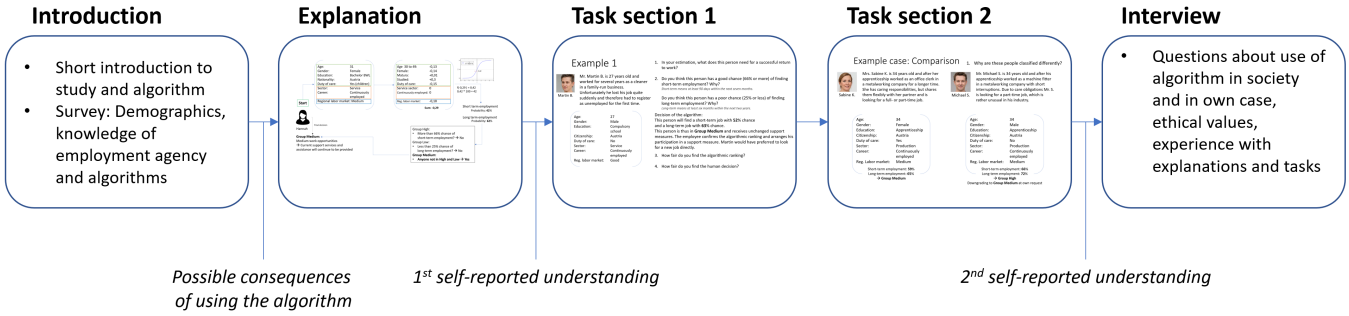


**Figure 3: Depiction of the study procedure. Participants received a short introduction about the goals and scope of the study, filled out a questionnaire, and then were asked about the possible consequences of implementing the *AMS algorithm*. They then received one of three explanation modalities and were asked about their self-reported understanding. Participants then proceeded through both task sections, were asked again for their self-reported understanding and finally answered several interview questions.**

involved in every category, into a total of 21 codes. We analysed the interviews once more using this deductive framework with the aim to examine whether participants' responses could be mapped to these pre-defined categories and to examine whether the framework captured aspects that our inductive codes missed. We found that while most inductive codes from our first pass could be assigned to one of the "six facets of understanding", the framework also introduced additional differentiations that were useful for the analysis, which will be further described in Section 4. A detailed description of both inductive and deductive codes is further included in the supplementary material.

## 3.4 Participants

In Table 1, we present a description of the 30 study participants, whom we recruited in four locations: a café near a university, a café in another city district, an auto repair garage, and a local employment agency's office. We conducted three of the studies online; all others were conducted personally in place. In our recruitment, we aimed for the inclusion of the two stakeholder groups "users" and "affected stakeholders", a balance in terms of age, gender, and occupation, as well as a focus on people with no expert knowledge of algorithms. While of course not representative of the general population, the participant sample allowed us to gather a wide

**Table 1: Details on the participants of our study. 5 participants were employees of the Public Employment Agency or similar institutions, whom we define as "domain experts" and who are indicated with a *d* attached to their ID. 5 participants were job-seeking at the time of the study, whom we define as "affected stakeholders".**

| ID | Explanation | Age | Gender | Education | Occupation |
|---|---|---|---|---|---|
| 1 | textual | 32 | F | Master | Researcher |
| 2 | textual | 26 | M | Master | PhD Student |
| 3 | textual | 23 | D | Bachelor | Master's student |
| 4 | textual | 36 | F | Master | UX Researcher |
| 5 | textual | 49 | M | PhD | Financial Advisor |
| 6 | textual | 25 | M | 1st state examination | PhD Student |
| 7 | textual | 47 | F | Academy | Leisure pedagogue |
| 8d | textual | 41 | M | Academy | Social worker |
| 9 | dialogue | 26 | M | Bachelor | Student |
| 10 | textual | 24 | M | Apprenticeship | Car mechanic |
| 11 | textual | 24 | M | Apprenticeship | Car mechanic |
| 12 | dialogue | 26 | M | Apprenticeship | Car mechanic |
| 13 | dialogue | 58 | M | Bachelor | Clerk |
| 14 | dialogue | 69 | F | University | Pension |
| 15 | dialogue | 42 | F | University | Job-seeking |
| 16 | dialogue | 52 | F | Apprenticeship | Job-seeking |
| 17 | dialogue | 51 | M | Vocational college | Job-seeking |
| 18d | dialogue | 60 | F | Master | Application trainer |
| 19d | dialogue | 57 | M | A level | Personnel consultant |
| 20d | dialogue | 55 | F | University | Personnel consultant |
| 21 | interactive | 28 | F | Bachelor | Student |
| 22 | interactive | 23 | F | A level | Student |
| 23 | interactive | 37 | M | University | Employed |
| 24d | interactive | 48 | M | University | Trainer |
| 25 | interactive | 26 | F | Apprenticeship | Short-term worker |
| 26 | interactive | 60 | M | University | Job-seeking |
| 27 | interactive | 35 | M | University | Self-employed |
| 28 | interactive | 27 | M | Master | Journalist |
| 29 | interactive | 29 | F | Master | Consultant |
| 30 | interactive | 31 | M | Bachelor | Consultant |

range of perspectives on how the *AMS algorithm* was understood and perceived in terms of fairness.

As motivated in Section 1, algorithmic decision-making systems can be detrimental to democracy if their implementation is faulty or unaware of the public's stance towards the system [10, 21, 23, 40, 43]. For this reason, we aimed to conduct the study with a balanced sample of the "general public", with the only condition being that participants would optimally not be experts on algorithms. The sample included both employees in the public employment sector and job-seekers, but also individuals of different educational and cultural backgrounds who were not directly affected by public employment. While our sample is not representative, it includes a range of different perspectives on the explanation, which provided a solid reasoning ground for examining the explanation's effects on understanding.

## 4 FINDINGS

In this section, we present the findings from our study, structured according to our RQs covering the following topics: understanding (Section 4.1), the effect of explanation modality (Section 4.2), and enabling the ethical assessment of ADM (Section 4.3).[14]

[14]Please note that some participant responses touch on sensitive topics, such as discrimination and self-harm.

### 4.1 Which "facets of understanding" emerge in the responses of participants after receiving the explanation? (RQ1)

*4.1.1 "Emotional" and "analytical" facets of understanding:* Participants' understanding of the explanation surfaced in different ways. The framework of Wiggins & McTighe [63] contains both "analytical" (explain, apply, take perspective) and "emotional" (interpret, empathise, self-reflect) facets of understanding. Participants tended to show facets of either one of the two sides, and the degree to which they felt personally affected by the algorithm made a difference in which facets they were most likely to use. For example, participants who were looking for employment at the time of the study rather responded to the question of whether the *AMS algorithm* should be used to score job-seekers with the facet "empathise":

> *These are just numbers, you don't know the background of why they lost the job. It's not fair because it's just a program and the human side is completely gone.* (P17)

In contrast, participants who never had contact with the Public Employment Agency and did not know anyone who was officially unemployed responded to the same question rather by critically "taking perspective" on the algorithm's application:

> *I think it's good as a recommendation. I think it turns into a procedure that is otherwise purely a human decision and provides a percentage probability, though I wouldn't use it on its own. I think it's good because the case worker has something to go by, but for some examples, it's not complex enough, that's what the human case worker is for.* (P6)

*4.1.2 Relating to the information in more than one way indicates higher understanding.* According to Wiggins & McTighe [63], successful understanding must show *all* of the six facets: "explain," "interpret," "apply," "empathise," "take perspective," and "self-reflect." While no participant used all six facets, we observe that several participants combined up to four facets in a single answer. In the following quote, the participant "explains" the current labour market situation, "interprets" what this would mean for the job-seeker, and as a result "takes perspective" by critically questioning the algorithm's employability scoring:

> *I disagree with the algorithm in this case because now there are shortages of employees in specific sectors like catering. So employers are forced to bend their requirements and will have a positive attitude towards applications from people such as Harald. He has a lot of experience, and he is motivated, so they will disregard his age, which would be bad in normal circumstances.* (P23)

Notably, neither stakeholder group, demographic background, nor prior knowledge seemed to affect if someone would show the ability to use multiple facets of understanding to make sense of the algorithm. This stands in contrast to previous studies, which found that measurable understanding increases with domain expertise [14, 57] and the level of education[52].

*4.1.3 High understanding co-occurs with self-reflection.* Participants that used multiple facets of understanding and thus showed a

higher understanding were able to give comprehensive answers to the tasks and often supplemented their responses by reflecting on their own understanding, beliefs, or circumstances. This occurred, for example, in response to the idea of being classified by the *AMS algorithm*:

> *In my case it would not be a disadvantage, because I am privileged in terms of education and I have no care obligations. [...] But if I were less privileged I wouldn't want that and if I had support needs I wouldn't want that either.* (P1)

Reflections tended to occur at a "turning point" of the interview, consisting of a question asking if the algorithm should be used on everyone, followed by a question asking if it should be used in the participant's own case. Often when participants responded to these questions disparately, for example by voting against the general usage, but agreeing to the usage on themselves, they felt inclined to reflect on their understanding and their beliefs without being prompted.

*4.1.4 Barriers to understanding.* Wiggins & McTighe [63] define three cognitive processes that can hamper understanding and learning: i) forgetting, ii) being unable to use what we learn, and iii) not knowing that we do not understand. Forgetting was an issue distributed throughout the participant sample, as participants could not refer back to the explanation while proceeding through the tasks. Notably, participants with lower levels of education encountered more issues of forgetting and were not always able to apply the learned information, which found its expression in incorrect recollections of the explanation and less emergence of understanding facets:

> *I didn't understand the conversion, I don't know the formula. Otherwise, I halfway understood it, they simply take the data and... I don't know how to explain it.* (P10)

In contrast, we did not find many instances of a participant not realising their lack of understanding. This, however, is more likely due to the difficulty of distinguishing it from the inability to apply knowledge. Finding a way to better identify this barrier to understanding could be a topic for future work.

*4.1.5 Facets not covered by the framework.* The comparison between inductive and deductive thematic analysis shows that while most inductive themes of understanding could be mapped to the theoretical framework of Wiggins & McTighe [63], we also identified themes that the framework did not include. First, the "six facets of understanding" do not provide a clear categorisation of expressions that reflect a strong value statement, such as "I believe that algorithms actually have no place in this field." (P18d). The closest match is the facet "interpret," which however aims more towards making the topic "personal or accessible through images, anecdotes, analogies" [63], and less towards value statements. It is debatable whether these statements bear evidence for understanding, but seeing that Langer et al. [30] describe value statements as an indicator for stakeholder understanding needs, this lack of categorisation stood out.

Second, our application of the framework does not allow us to distinguish between different domains of knowledge, which in our analysis sometimes blurred the lines between understanding the algorithm and understanding the public employment system. A clearer distinction between the subjects of understanding would be useful for future analyses.

## 4.2 Which, if any, correlation to the explanation modality can be seen in the understanding facets? (RQ2)

The different modalities of how the explanation was presented (textual, dialogue, and interactive) had a limited effect on the emergence of specific understanding facets. However, we point out two findings related to the textual and dialogue explanation modality.

*4.2.1 Textual modality leads to understanding barriers.* The "textual" modality group showed more and deeper barriers to understanding than other groups. Several participants commented on the amount of text in the explanation and their learning process:

> *I have to be honest, I find it easier when someone explains something to me. Then I can also ask until I get it. It is hard for me to understand something like that just by reading it.* (P10)

In contrast, participants expressed their satisfaction both with the dialogue and interactive modality, praising the option to ask questions and the engagement with the explanation. Of all 20 participants in the dialogue and interactive modality groups, only one stated that they would have rather liked a textual explanation.

*4.2.2 Dialogue modality leads to increased expression of some understanding facets.* Compared to the textual and interactive modality, the dialogue explanation showed an overall increase in the number of words spoken by participants and more usage of the facets "interpret", "empathise", and "self-reflect". Factual questions by participants mostly served the purpose of confirming information that was already present in the flowchart (e.g., "*So, having studied actually has the largest effect on your score?*" (P13)). In contrast, questions that pertained to personal concerns or systemic issues often touched on the "emotional" facets of understanding (e.g., "*Do women automatically get less points? Where does discrimination begin, where does it end? (P15)*"). These questions were not answered in detail during the explanation, but instead served as points of entry to the later interview. The dialogue modality might have thus helped to later start the in-depth conversation about the algorithm by introducing a verbal interaction directly at the beginning and giving participants more opportunities to correct their understanding. Other possible reasons for these findings will be discussed in Section 5.

## 4.3 Do participants demonstrate the ability to engage in meaningful discourse about the algorithm, for instance in evaluating the algorithms' fairness with regard to decisions about job-seekers? (RQ3)

After each case example in task section 1, participants were asked to assess and discuss the fairness of both the algorithmic and human decisions regarding the case. We used these fairness assessments as a form of proxy for their ability to engage in discourse about the algorithm's ethical dimensions. We find that all participants

were able to articulate a basic fairness assessment, but that their statements differed strongly in detail and argumentative reasoning.

*4.3.1 Participants demonstrate the ability to individually assess algorithmic fairness.* The "epistemic" satisfaction of a trustworthy AI criterion means that people are able to assess on their own grounds whether a system fulfills a certain ethical criterion, such as being fair or not [30]. This does not mean that the system *is* fair, only that people are able to *discuss* it. Participants fulfilled this "epistemic" criterion as they showed a rich and diverse range of fairness assessments of the *AMS algorithm*, including topics such as the influence of the "human-in-the-loop", the perceived gender inequality, the perception of "algorithmic objectivity" and, in parallel to findings from Scott et al. [49], the importance of using the system for "orientation purposes, not to deny access to resources".

We further observe that the fairness assessments changed depending on whether participants made more use of analytical or emotional facets of understanding, as evident in these responses to the algorithm predicting "low" employability for a job-seeker:

> *I see him almost like me. He's 49, so for me, he would already count as 50+ and should get extra support. [...] I find the algorithm good and the human decision bad, because it goes by numbers and is not accommodating. The algorithm just sees 49 and makes the decision.* (P17)

> *I understand why the algorithm says Group "Low" in this case. If he retrains and can explain his career well, I think he could find work, but needs support to do so. [...] Personally, I would perhaps not rate him that way, but I think it's good that he gets into the group.* (P6)

The first participant relates the decision to his own personal circumstances and speaks about the difference between human and algorithmic decision-making, while the second takes a more analytical approach in speaking about the consequences of the decision to justify his assessment. Despite their different reasons, both participants were able to argue why they agreed with the algorithm, thus fulfilling the "epistemic" criterion [30]. At the same time, this case exemplifies that individuality in understanding also leads to individuality in the fairness assessment.

*4.3.2 Understanding barriers lead to less nuanced fairness assessments.* As expected, participants who encountered more understanding barriers assessed the algorithm's fairness in much less detail. Often, some form of blanket statement was used that addressed neither dimensions of fairness nor facets of understanding:

> *I think the algorithmic classification is fair, because I was of the same opinion.* (P10)

The difference to more detailed fairness assessments becomes apparent when we compare this to statements from participants with the same demographics, mostly similar education, but a different explanation modality and largely no barriers to understanding:

> *I'll try to leave out the current situation in the labour market. I mean, he has his problems. I understand that the chance is low, but this is too low, the algorithm and especially the employee are not fair. Perhaps the employee is new and has no experience in how to place people in the labour market.* (P12)

The latter statement shows three different facets of understanding while the participant reasons about the fairness assessment: taking a critical perspective on the algorithm, reflecting on the particularities of the post-COVID labour market, and including the perspective of the employee.

*4.3.3 Unwillingness to engage in discourse about the algorithm's fairness.* Besides understanding barriers, the most common reason for a lack of fairness assessments was a general objection to assessing the system as "fair" or "unfair". These objections were often accompanied by the (flawed) argument that an "objective" algorithm was not able to act fairly or unfairly:

> *What does fair mean? I can understand the algorithmic decision. Fair is such a strong moral word, I don't find it immoral, I rather find it "justified", "understandable", "realistic".* (P29)

Similar responses were given when participants were prompted to rate the algorithm as "just", "legitimate", "social", "biased", and "democratic". Several participants commented on the close meaning between words, which points to the need of differentiating concepts when asking for complex moral value judgements.

## 5 DISCUSSION

In this section, we discuss our findings with regard to the research questions and suggest directions for designing explanations that create increased understanding in different stakeholders and enable the ethical assessment of ADM.

### 5.1 Designing explanations to address all six facets of understanding

The "understanding by design" framework [63] states that the primary objectives in creating understanding are to convey a topic's central ideas, address all "facets" of understanding, and uncover misunderstandings. Applying the framework to our explanation setup allowed us to gather insights into the participants' mental processes involved in understanding, i.e., which facets emerged for which participant and how many facets emerged simultaneously. In particular, we want to highlight that participants showed "emotional" understanding facets, such as "empathise" and "interpret", which are not always addressed in ADM explanations, despite their known importance in human perception of ADM systems [31, 49, 64]. The facet "self-reflect" is another valuable dimension that could improve future explanations, as several studies suggest that "metacognition", in the form of reflecting on one's knowledge and understanding, is a "most powerful predictor of learning" [48, 60].

Theories on learning and understanding like the "six facets" [63] can thus guide explanation design by identifying distinct learning goals and by outlining mental processes that support understanding. We chose Wiggins' & McTighes' [63] framework due to the practical interpretation of understanding and the well-established theoretical foundation in Bloom's taxonomy [4, 11]. Other promising sources for future studies include the concept of "responsive teaching" [45], "knowledge building and knowledge creation" [47], and "sensemaking theory" [62]. We posit that drawing from these theories in the design and development of explanations can guide

research to better support "humans in learning about particular AI systems and how to work with or around them" [28].

## 5.2 Effect of explanation modality on understanding

Our findings suggest that the textual and dialogue modality had the most effect on understanding. The textual modality led to less emergence of understanding facets for several participants, some of whom stated that they did not usually rely on text to learn information. Compared to Szymanski et al. [57], where participants disliked textual information but actually performed better using them, participants in our case did not show any advantages in understanding after receiving the textual explanation.

The dialogue modality, in contrast, led to an overall increase in observed facets of understanding, which could have multiple reasons: i) the higher amount of words spoken, ii) the option to ask questions during the explanation, and iii) the additional personal interaction. Although the dialogue modality in theory allowed for a higher amount of factual input, participants seldom chose to ask more than four or five brief factual questions, which mostly pertained to secondary details of the algorithm (e.g., the weighting of certain features) and in terms of information differed little from the textual modality. At the same time, several participants used the dialogue modality to express their opinions and attitudes towards specific information and to ask more profound questions about e.g., the intention behind the algorithm's deployment and the selection of the "standard group". The dialogue modality might thus serve as a conversational ice-breaker due to the direct interaction between "explainer" and participant, encouraging participants to share their own thoughts and, in turn, increase their understanding. This also connects to findings from Miller [38], who states that explanations between humans are "social" and "presented as part of a conversation." This is striking considering that verbal or dialogue explanations are seldom used in contemporary XAI research. Our findings establish this explanation modality as a valid alternative to be explored in future ADM explanations.

## 5.3 Enabling participants to engage in discourse about the algorithm's ethical values

We used the fairness assessment of the *AMS algorithm* as a form of proxy to observe whether people are able to articulate value assessments after receiving the explanation. We find that i) the explanation provided most participants with the necessary information to form a detailed fairness assessment, and thus enabled them to successfully engage in discourse about the algorithm's ethical values, and ii) in contrast to other participants, domain experts were able to discuss the algorithm's implementation even before the explanation, stating that they had come in contact with automated tools. On the other hand, participants who showed barriers to understanding articulated less nuanced fairness assessments. Further, some participants stated that an algorithm simply could not be judged in terms of fairness, despite showing multiple facets of understanding in their responses. This means that understanding the system might not be the only prerequisite for stakeholders to make a value assessment, but that understanding what specific ethical values mean when being applied to the system might be just as important. Future explanations should thus consider adding an explanation of the ethical values or finding stand-ins that convey the core of the value in a more accessible manner.

## 5.4 Limitations

In this section, we touch on four limitations of our study. First, our assessment of understanding relied largely on qualitative analysis of participants' responses, which were self-reported and thus might limit objectivity. For future studies, we therefore consider including a quantitative evaluation of understanding. Second, some inductive codes such as strong value statements were not covered in the "six facets" framework [63]. Seeing that value statements can "serve as an orientation for when stakeholders are more likely to demand higher degrees of understanding" [30], future analyses should consider including a facet covering these forms of statements. Third, the study design relied on memorisation of information, as no reference was given to participants while proceeding through the tasks. While this impacted some participants, it also highlighted a difference in memorability of the textual, dialogue, and interactive modality. To offset over-reliance on recall, we consider providing participants with a way to review the explanation in future studies. Lastly, we noticed a bias towards higher education in our participant sample, with 22 participants having a university degree. However, we aimed to collect different perspectives by speaking to employees from the local employment agency, job-seekers and individuals with different educational backgrounds.

# 6 CONCLUSION

In this paper, we inductively and deductively analyse participants' understanding of an algorithmic decision-making system after they received one of three explanation modalities (textual, dialogue, and interactive). We find that all of the six "facets of understanding" [63] (explain, interpret, apply, empathise, take perspective, self-reflect) emerge in participant responses throughout the study, with some participants expressing high understanding by combining multiple facets at once. We argue that incorporating theories from the learning sciences can significantly improve the design of ADM explanations by adapting them to the underlying thought processes of learning and understanding in individuals. We further highlight the "dialogue" explanation modality as a valid alternative to convey information and gather in-depth insights on how participants understand and contextualise explanations. Lastly, while we observe that most participants are able to articulate a fairness assessment of the explained ADM system and that a more pronounced understanding supports this articulation, it also becomes evident that participants have general difficulties considering the meaning of "fairness" in the context of algorithmic systems. We posit that letting stakeholders independently assess algorithmic systems in terms of ethical values such as fairness, accountability, or transparency, could require an additional explanation of how to apply these values in an algorithmic context, in addition to increasing stakeholders' understanding of the algorithm itself.


# ACKNOWLEDGMENTS

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

# REFERENCES


[1] Doris Allhutter. 2021. Ein Algorithmus zur effizienten Förderung der Chancen auf dem Arbeitsmarkt? *WISO – Zeitschrift für Sozial- und Wirtschaftswissenschaften* 44.JG (2021), 82–95.

[2] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data* 3 (2020). https://doi.org/10.3389/fdata.2020.00005

[3] Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. *DER AMS-ALGORITHMUS: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS).* Technical Report. https://epub.oeaw.ac.at/0xc1aa5576_0x003bfdf3.pdf

[4] Lorin W. Anderson and David R. Krathwohl (Eds.). 2001. *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives* (complete ed ed.). Longman, New York.

[5] American Educational Research Association, American Psychological Association, and National Council on Measurementin Education (Eds.). 2014. *2014 - Standards for education and psychological testing.* American Educational Research Association, Washington, D.C. OCLC: ocn826867074.

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

[7] Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. 2018. Improving refugee integration through data-driven algorithmic assignment. *Science* 359, 6373 (Jan. 2018), 325–329. https://doi.org/10.1126/science.aao4408

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[9] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 248–266. https://doi.org/10.1145/3531146.3533090

[10] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3173574.3173951

[11] Benjamin Bloom, M. Engelhart, E. Furst, W. Hill, and D. Krathwohl. 1956. *Taxonomy of Educational Objectives, Handbook 1_ Cognitive Domain.pdf.* Addison Wesley Publishing Company.

[12] John D. Bransford and Marcia K. Johnson. 1972. Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall. *Journal of Verbal Learning and Verbal Behaviour 11* (1972).

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

[14] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300789

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

[16] European Commission. 2021. Laying Down Harmonised Rules on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

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

[18] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285. https://doi.org/10.1145/3301275.3302310

[19] Eleanor Duckworth (Ed.). 2001. *"Tell me more": listening to learners explain.* Teachers College Press, New York.

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

[21] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28, 4 (Dec. 2018), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[22] Yoan Hermstrüwer and Pascal Langenbach. 2022. Fair Governance with Humans and Machines. *SSRN Electronic Journal* (2022). https://doi.org/10.2139/ssrn.4118650

[23] César Hidalgo, Diana Orghain, Filipa de Almeida Jordi Albo Canals, and Natalia Martin. 2021. *How Humans Judge Machines.* MIT Press, Cambridge, MA.

[24] Jürgen Holl, Günter Kernbeiß, and Michael Wagner-Pinter. 2018. *Das AMS-Arbeitsmarktchancen-Modell.* Technical concept. Synthesis Forschung, Vienna.

[25] Jürgen Holl, Günter Kernbeiß, and Michael Wagner-Pinter. 2019. *Personenbezogene Wahrscheinlichkeitsaussagen (»Algorithmen«): Stichworte zur Sozialverträglichkeit.* Technical concept. Synthesis Forschung, Vienna.

[26] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 310–323. https://doi.org/10.1145/3531146.3533097

[27] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 702–714. https://doi.org/10.1145/3531146.3533135

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

[29] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, San Jose, CA, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

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

[31] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1035–1048. https://doi.org/10.1145/2998181.2998230

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

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

[34] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[35] Paola Lopez. 2019. Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. (2019), 21.

[36] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445

[37] Marco Marabelli and Sue Newell. 2019. Algorithmic Decision-making in the US Healthcare Industry: Good for Whom? *Academy of Management Proceedings* 2019 (08 2019), 15581. https://doi.org/10.5465/AMBPP.2019.15581abstract

[38] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[39] High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. (April 2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[40] Cathy O'Neil. 2016. *Weapons of math destruction: how big data increases inequality and threatens democracy* (first edition ed.). Crown, New York.

[41] Evan M. Peck, Sofia E. Ayuso, and Omar El-Etr. 2019. Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania. In *Proceedings of the*

*2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300474

[42] Emma Pierson. 2018. Demographics and discussion influence views on algorithmic fairness. http://arxiv.org/abs/1712.09124 arXiv:1712.09124 [cs].

[43] Ciprian N Radavoi. 2020. The Impact of Artificial Intelligence on Freedom, Rationality, Rule of Law and Democracy: Should We Not Be Debating It? *Texas Journal on Civil Liberties & Civil Rights* 25, 2 (2020), 24.

[44] Lydia Reader, Pegah Nokhiz, Cathleen Power, Neal Patwari, Suresh Venkatasubramanian, and Sorelle Friedler. 2022. Models for understanding and quantifying feedback in societal systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1765–1775. https://doi.org/10.1145/3531146.3533230

[45] Amy D. Robertson, Rachel Scherr, and David Hammer (Eds.). 2015. *Responsive Teaching in Science and Mathematics* (0 ed.). Routledge. https://doi.org/10.4324/9781315689302

[46] Brian K. Sato, Cynthia F. C. Hill, and Stanley M. Lo. 2019. Testing the test: Are exams measuring understanding? *Biochemistry and Molecular Biology Education* 47, 3 (May 2019), 296–302. https://doi.org/10.1002/bmb.21231

[47] Marlene Scardamalia and Carl Bereiter. 2014. Knowledge Building and Knowledge Creation: Theory, Pedagogy, and Technology. In *The Cambridge Handbook of the Learning Sciences* (2 ed.), R. Keith Sawyer (Ed.). Cambridge University Press, 397–417. https://doi.org/10.1017/CBO9781139519526.025

[48] Gregory Schraw, Kent J. Crippen, and Kendall Hartley. 2006. Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Research in Science Education* 36, 1-2 (March 2006), 111–139. https://doi.org/10.1007/s11165-005-3917-8

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

[50] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1330–1340. https://doi.org/10.1145/3531146.3533189

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

[52] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Enhancing Fairness Perception – Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. *International Journal of Human–Computer Interaction* (July 2022), 1–28. https://doi.org/10.1080/10447318.2022.2095705

[53] M. K. Smith, W. B. Wood, W. K. Adams, C. Wieman, J. K. Knight, N. Guild, and T. T. Su. 2009. Why Peer Discussion Improves Student Performance on In-Class Concept Questions. *Science* 323, 5910 (Jan. 2009), 122–124. https://doi.org/10.1126/science.1165919

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

[55] Sarah Spiekermann. 2021. From value-lists to value-based engineering with IEEE 7000™. In *2021 IEEE International Symposium on Technology and Society (ISTAS)*. 1–6. https://doi.org/10.1109/ISTAS52410.2021.9629134 ISSN: 2158-3412.

[56] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2021. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. (2021).

[57] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. https://doi.org/10.1145/3397481.3450662

[58] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *The American journal of evaluation* 27, 2 (2006), 237–246.

[59] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (April 2018), 1–14. https://doi.org/10.1145/3173574.3174014

[60] Marcel V. J. Veenman, Bernadette H. A. M. Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning* 1, 1 (April 2006), 3–14. https://doi.org/10.1007/s11409-006-6893-0

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

[62] Karl E. Weick, Kathleen M. Sutcliffe, and David Obstfeld. 2005. Organizing and the Process of Sensemaking. *Organization Science* 16, 4 (2005), 409–421. https://doi.org/10.1287/orsc.1050.0133 arXiv:https://doi.org/10.1287/orsc.1050.0133

[63] Grant P. Wiggins and Jay McTighe. 2005. *Understanding by design* (expanded 2nd ed ed.). Association for Supervision and Curriculum Development, Alexandria, VA. OCLC: 60756429.

[64] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3174230