The *Who* in Explainable AI: How AI Background Shapes Perceptions of AI Explanations

UPOL EHSAN, Georgia Institute of Technology, USA
SAMIR PASSI, Cornell University, USA
Q. VERA LIAO, IBM Research AI, USA
LARRY CHAN, Georgia Institute of Technology, USA
I-HSIANG LEE, Georgia Institute of Technology, USA
MICHAEL MULLER, IBM Research AI, USA
MARK O. RIEDL, Georgia Institute of Technology, USA

Explainability of AI systems is critical for users to take informed actions and hold systems accountable. While "opening the opaque box" is important, understanding *who* opens the box can govern if the Human-AI interaction is effective. In this paper, we conduct a mixed-methods study of how two different groups of *whos*—people with and without a background in AI—perceive different types of AI explanations. These groups were chosen to look at how disparities in AI backgrounds can exacerbate the creator-consumer gap. We quantitatively share *what* the perceptions are along five dimensions: confidence, intelligence, understandability, second chance, and friendliness. Qualitatively, we highlight *how* the AI background influences each group's interpretations and elucidate *why* the differences might exist through the lenses of appropriation and cognitive heuristics. We find that (1) both groups had unwarranted faith in numbers, to different extents and for different reasons, (2) each group found explanatory values in different explanations that went beyond the usage we designed them for, and (3) each group had different requirements of what counts as humanlike explanations. Using our findings, we discuss potential negative consequences such as harmful manipulation of user trust and propose design interventions to mitigate them. By bringing conscious awareness to *how* and *why* AI backgrounds shape perceptions of potential creators and consumers in XAI, our work takes a formative step in advancing a pluralistic Human-centered Explainable AI discourse.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **User studies**; *Empirical studies in collaborative and social computing*; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: Explainable AI, Human-Centered Computing, Data Vision, User Perceptions, Artificial Intelligence, Heuristics, Appropriation, User Characteristics

1 INTRODUCTION

As AI-driven systems increasingly power high-stakes decision-making in public domains such as healthcare [22, 60, 69, 89], finance [98, 108], law [12, 136, 143], and criminal justice [56, 72, 122], their explainability is critical for end-users to take informed and accountable actions [130]. Issues concerning explainability lie at the heart of Explainable AI (XAI), a research area that aims to provide human-understandable justifications for the system's behavior [2, 42, 55]. Explainability is not a new issue within AI [67, 105, 129], but the proliferation of Deep Learning and Reinforcement Learning based approaches—models of which are considered hard to interpret, even by experts—has led to a remarkable growth in techniques that aim to 'open' the AI opaque box [55].

While opening the opaque box is important, *who* opens the box also matters. Implicit in Explainable AI is the question: "explainable to whom?" [41]. The *who* governs the most effective way of describing the *why* behind the decisions. Getting a situated understanding of how different *who's* with different user characteristics matter in XAI is thus important. To give an illustrative example: riders (end-users) of a self-driving car have different user characteristics than its engineers (developers). Riders, many of whom are not AI experts, might not have the AI background that the engineers have and thus have different explainability needs and goals.

One's AI background is an impactful user characteristic in XAI because there is often disparity in this characteristic between creators/developers and end-users, which can lead to inequities [26]. Many end-users are unlikely to have AI backgrounds comparable to the creators of the technology [66]. Nonetheless, XAI developers tend to design explanations *as if* people like them are going to use their systems [102]. In fact, a majority of current deployments of XAI technologies serve AI engineers instead of end-users [4, 82]. This creates a consumer-creator gap, one between design intention and reality— how developers envision the AI explanations to get interpreted and how users actually perceive them. If we want to bridge this gap, we need to understand how user characteristics, such as AI background, impact it.[1]

In this paper, we share *how* and *why* one's AI background (or lack thereof) shapes their perceptions of AI explanations. Focusing on two groups, one with and one without an AI background, we found that (1) Both groups had unwarranted faith in numbers, but exhibit it for different reasons and to differing degrees, with AI group showing higher propensity to over-trust numerical representations and potentially be mislead by the presence of it. (2) The two groups found different explanatory values beyond the usage that the explanations were designed for. (3) Even in their aligned appreciation for humanlike-ness, each group had different requirements concerning *what* counts as humanlike explanations. These insights have potential negative implications like susceptibility to harmful manipulation of user trust.

We found these insights through a mixed-methods study where we probed for user perceptions of *three* types of AI-generated explanations: (1) natural language with justification (explaining the "why" behind the action), (2) natural language without justification (describing "what" the action was), and (3) numbers that determine the agent's actions (akin to "transparent" AI). We measure perceptions along five dimensions: *confidence, intelligence, understandability, second chance,* and *friendliness*, which are grounded in related work around HCI, HRI, and XAI [15, 30, 32, 42, 145] and quantitatively share within- and between- group differences. Through qualitative analysis, we examined *how* AI background shaped each group's interpretation of explanations and highlight the reasons behind their perceptual differences along three salient themes corresponding to the three aforementioned findings.

Beyond illustrating *how* AI background influences the groups' perceptions, we also elucidate the *why*–possible causes– behind the group differences using the conceptual lenses of of *heuristics* (mental short-cuts) [68, 127] and *appropriation* (users' repurposing of a design) [36, 107, 123]. In light of the findings, we share concrete design implications around mitigating the risks of over-reliance on numbers which can potentially lead to negative consequences such as over-trust on XAI systems. We share broader lessons around how our insights can help re-imagine AI education and mitigate potential harmful manipulation with explanations. By bringing conscious awareness of the group differences to human-centered design of XAI systems, we address the AI creator-consumer gap by making the following contributions:

- We quantify the user preferences (*what)* of three types of AI explanations along five dimensions of user perceptions.
- We qualitatively situate *how* one's AI background (or lack thereof) influences one's perception of the explanations.
- We elucidate *why* the group differences might exist and interpret them through the conceptual lenses of heuristics and appropriation.
- Using our findings, we identify potentially negative consequences (like harmful manipulation of user perceptions and over-trust in XAI systems) and propose mitigation strategies.

---

[1] By highlighting the creator-consumer gap in XAI, we do not mean to undermine the diversity of stakeholders in the ecosystem. By calling attention to extreme ends, we are highlighting the severity of the gap while fully acknowledging the ecosystem's diversity. See, e.g.: [54, 101, 106, 117].

2 BACKGROUND

In this section, we review related work in the field of XAI salient to the paper, highlight the need to attend to XAI's sociotechnical dimensions and human-centered perspectives, and discuss HCI work studying how user background shapes users' perception of and needs for technology that motivated our work.

2.1 Explainable AI

While the origin of "explainable AI" can be traced back to expert systems in the 1980s [135], the field of XAI has been undergoing a resurgence due to the proliferation of complex Deep Learning models. Although there is a current lack of consensus on the meaning of explainability and related terms such as interpretability [7, 121], XAI work shares a common goal of making the AI systems' decisions or behaviors understandable by people [2, 42] Among other dimensions to map the landscape of technical XAI approaches, the field differentiates between methods to build directly interpretable model and methods to generate explanation for opaque-box models [49, 86, 120, 154] (for a detailed overview see recent survey papers [2, 7, 55]). While simpler models such as linear regression and decision-tree are typically considered directly interpretable but low-performing, recent work (e.g. [31, 150]) focuses on developing new algorithms that "open" the opaque-box and allow "under the hood" inspection without sacrificing performance.

In contrast, *explanation generation* methods–used in this paper–aim to explain models that are not directly human-understandable (e.g., deep neural networks). They are often post-hoc techniques [42, 87, 102, 120, 154] that could be applied after model building. Typically, these methods rely on distilling a simpler model from the input and output [96, 120] or meta-knowledge about the model [42] to generate explanations that approximate the model's behavior. While there is, by design, a loss of scrutability, these methods allow the flexibility to make any model explainable, and thus have become popular and been applied to transforming simulation logs to explanations[142], intelligent tutoring systems [27], transforming AI plans into natural language [141], and translating multi-agent communication policies into natural language [6]. By privileging accessible understanding over revealing "under the hood" model mechanisms, explanation generation methods can be geared towards non-AI experts. In this paper, we focus on a specific explanation generation technique called *rationale generation* [42]—a process of producing a natural language explanation for agent behavior as if a human had performed the behavior and verbalized their inner monologue. While explanations can be in any modality, rationales are natural language-based, making it especially accessible for non-AI experts [42].

2.2 Towards Human-Centered XAI

There has been a growing recognition that XAI systems are often developed without an understanding of the recipients' needs and characteristics [41, 102]. For instance, many of the XAI techniques created to support explainability needs during model development [120, 125] may break down when it comes to serving end-users with different needs [82]. It is imperative to follow human-centered approaches to understand the "personal, social, and cultural aspects" [64] of the recipients of AI explanations, especially since a monolithic view of the *who* may inadvertently risk dehumanization [20, 73]. Given deployment in high-stakes settings, AI systems designed without attending to the needs and values of different stakeholders may also risk marginalizing certain groups or exacerbating existing inequities [8, 114].

The need for human-centered approaches in XAI has inspired increasing efforts among HCI and CSCW researchers, following the community's long-standing tradition of designing and studying explainable computing systems [43, 46, 76, 84, 119]. Studies empirically evaluating XAI techniques in specific use contexts reveal the divergent needs and preferences of users [4, 19, 24, 37, 58, 70]. For example, while data scientists might need multiple XAI tools for comprehensive understanding [58], simple explanations are often sufficient for AI-novices [24]. User characteristics such as cognitive load disposition [48] and general trust in AI [37] could moderate how users perceive AI explanations. Some studies further reveal potential drawbacks of AI explanations— how explanations could impose undesired cognitive burden [1], create a false sense of security and over-trust [48, 70], and how even placebic explanations (devoid of justificatory content) can engender trust in AI systems [44].

Researchers have begun to examine people's cognitive process of interpreting AI explanations, which could help us understand atypical and misaligned user receptions of XAI. Recent work highlights the dual-process of cognition when people process AI explanations [17]. The dual-process theory [68, 118, 148] posits that people's cognitive processes follow two systems: System 1 processes stimuli in a fast and automatic manner, whereas System 2 engages in deliberative and analytical thinking. System 1 often relies on heuristics (rules-of-thumb or mental short-cuts) that can be developed through past experiences. These heuristics, if applied inappropriately, should be considered cognitive biases [68]. In XAI, there is often an assumption that people engage in analytic System 2 thinking whereas there is emerging evidence that people mostly engage in System 1 thinking [17]. Negative consequences of cognitive biases such as over-trust in XAI could be attributed to a System 1 heuristic of associating explanations with AI competence [26, 113]. One way to mitigate these biases would be to use *Cognitive Forcing Functions* (CFFs)–interventions that disrupt heuristic reasoning and promote System 2 analytical thinking [28].

Human-centered XAI calls for pluralistic explanation design—for the *who*, not just the *what*, of XAI [41]. Recent XAI work has begun to differentiate major categories of XAI consumers, from builders to regulatory bodies [7, 34, 104, 138]. While this work is informative, there is still a dearth of empirical insights and understanding on the differences between these XAI users, nor actionable design guidelines.

A generative approach towards pluralistic design of XAI, we believe, is to develop a systematic understanding of *how* and *why* users with different characteristics (e.g., with or without AI backgrounds) form different perceptions (*what*) of XAI systems. Such insights can refine our understanding of *who* the humans are in XAI. To our knowledge, this has not yet been systematically explored in the specific context of XAI. Thus, our work adds to the discourse by adding empirical insights through a systematic exploration of two different *who's* in XAI. Our work is also motivated by broader research studying individual differences that impact human-AI interaction, as reviewed below.

2.3 Individual differences in human-AI interaction

There is a rich history of transforming insights of how individual differences impact technology perception into the design of personalized, accessible, and inclusive technologies. For example, there is a large body of work on how various types of epistemic background, including computer literacy [35], digital literacy [9], and numeracy [65], impact user competency to use computing systems and ways to mitigate the competency gaps.

Recent work has paid attention to how user characteristics impact human-AI interaction. For example, studies on human-robot [29, 79, 133] and human-agent interaction [81, 83] found that user's schema, whether an agent is seen as a utilitarian tool or a social entity, leads to noticeable difference in the interaction with and evaluation of the agent. Research around AI fairness has identified various mediating factors such as users' education level, fairness criteria, and general trust in ML [37, 147]. In short, individual differences and user characteristics shape perceptions of AI systems and should be appropriately understood and carefully accommodated when designing AI systems.

A commonly studied user characteristic in Human-AI interaction is users' knowledge about AI, often operationalized as AI programming experience [109], building ML models [48], or type of profession (e.g., data scientist) [58]. Recently

Long and Magerko provided a concrete definition of AI literacy [95] with a set of core competencies to understand and use AI, and proposed design guidelines to mitigate the competency gaps. Recent work also examines the implications of background in AI as a determining factor of one's *role* in an AI ecosystem. Motivated to "problematize the asymmetric relationship between technical experts and users", Cheon and Su [25] highlighted the misalignment between the creators' (roboticists') vision of how consumers (end-users) interpret and use the product–a salient theme in this paper as well. McDonald and Pan [100] examined how CS students (likely to become AI developers) viewed ethical problems in AI, found substantial limitations, and called for a closer integration of ethics education with technical training.

Our understanding of how people's AI backgrounds might impact their perceptions of AI explanations also draws from a long line of social science research on the relations between sensemaking [151] and professional knowledge [52]. Passi and Jackson [115], for instance, analyzed academic learning practices in the field of data science to show how students start "seeing" the world differently once they learn to work with algorithms and numbers. Through ethnographic fieldwork, they highlight how having a computational background enables students to gain actionable forms of "data vision"—ways of seeing that allow them to approach and analyze the world as data.

A person's AI background (or a lack thereof)—a focal point of this paper—in fact, directly impacts user perceptions. In a recent ethnographic study [116], researchers found that data scientists and business analysts perceived an AI system's accuracy score differently: business analysts saw the score as a measure of overall performance (good vs. bad), while data scientists perceived more granular insights into types of errors (false positives vs. false negatives). As people learn specific ways of *doing*, it also changes their own ways of *knowing*—in fact, as we argue in this paper, people's AI background impacts their perception of *what* it means to explain something and *how*.

Thus, the *who* questions are important, because people tend to interpret technologies differently, leading to different usages of those technologies [5, 16, 36, 107, 123]. The process of differential interpretation, followed by different usage, has been called *appropriation*–e.g., usage patterns that go beyond the original designers' expectations [75, 139]. Dix listed six principles of user appropriation of technologies, including support for interpretive flexibility [36]. In this paper, we extend the discourse of user characteristics and individual differences in XAI by studying how AI background shapes the interpretation and appropriation of AI explanations.

3 RESEARCH DESIGN AND METHODS

We begin by sharing the research questions (RQs) followed by how we operationalize key aspects of the research design.

- RQ1: Quantitatively, what are the effects of different types of explanations on how people with or without an AI background perceive AI agents?
- RQ2: Qualitatively, how and why do differences in AI background result in different perceptions of explanations?

We address these RQs by conducting a within-subjects experiment in which two groups of participants, with or without an AI background, see three versions of AI agents (depicted as robots in our study, Fig. 1) with different types of explanation. We quantitatively measure the perceptions of the AI explanations and compare the differences between the two participant groups to address RQ1 (Section 4). Participants rank their preferences and justify their choices through open-ended text responses, which we qualitatively analyze to to understand the underlying differences between the two groups to address RQ2 (Section 5). Below, we unpack how we operationalize three things: (1) explanation types, (2) user perceptions, and (3) backgrounds in AI.

3.1 Explanation Generation Method and Types

We begin with the *task environment* to situate the design of the AI agents. In the user study (task details in Section 3.4), participants watched 3 robots (AI agents) carry out an identical sequence of actions which differed only in the way that the AI agent "thinks out loud" about its actions. The robots need to navigate through a sequential decision-making environment—a field of rolling boulders and a river of flowing lava–to retrieve essential food supplies for trapped space explorers (Fig. 1). The robots thus need to observe a dynamic environment and to think ahead in order to complete an objective. We chose a sequential environment because explainability of sequential tasks is under-explored while prior XAI work has explored non-sequential tasks (e.g., classification, captioning, etc. [146, 153, 155]). To solve the navigation problem, the agent used a Reinforcement Learning (RL) algorithm called *tabular Q-learning* [149]. Reinforcement Learning is both a promising technology for autonomous AI systems but also challenging from an explanation perspective. Tabular Q-learning agents attempt to learn the utility (called a Q-value for "quality" of the action) of different actions in different situations. Once learning is complete, the agent makes decisions by picking the action with the highest Q-value. In our case, the fully trained agent solved the environment, generating an action trace that contains the sequence of steps needed to reach the goal (food supplies) without failure. Recall that our study has 3 robots (AI agents) with different types of explanation. In order to standardize their actions across experimental conditions, *all robots use the same trace*. This means that the decision-making mechanism underlying each robot is the same RL-based one. They *only differ* in how they explain their actions.

Fig. 1. The three robots navigating the task environment and explaining their actions. From left to right, the robots and their colors are: Rationale-Generation (in blue), Action-Declaring (in purple), and Numerical-Reasoning (in green). In the screenshot, each robot is taking the same action, but they are explaining it differently. The explanation text accompanying each robot is taken verbatim from from the videos participants watched. To improve legibility, the text has been remastered to a higher resolution.

amongst the boulders...
couldn't go through...
need to wait patiently for a break between the boulders

I will stand still

0: 1.32998974
1: -2.12671373
2: -8.97841630
3: -9.97611535
4: 2.07117343

Explanation generation: The three different types of explanation is the within-subject variable in our study comparing perception differences between the AI and non-AI groups. Based on review of related work, we chose three types of explanation that vary in their justification quality and representation modality (e.g., textual, numerical). With these considerations in mind, we set up the robots to express themselves in three ways: (1) natural language with justification (explaining the "why" behind the action; Fig 1, left), (2) natural language without justification (describing "what" the action was, Fig 1, center), and (3) numbers that determine the agent's actions (akin to "transparent" AI, in this case showing Q-values, Fig 1, right). We share the explanation mechanisms and the attributes of the three robots below.

- The *Rationale-Generating (RG)* robot (Robot A in the study): this robot "thinks out loud" in natural language rationales explaining the "why" behind the action (#1 above). Our generation approach is similar to prior work in

XAI and HRI [30, 39] where they use a neural machine translation (NMT) [97] approach to produce satisfactory and plausible rationales to explain sequential behavior. We build on this technique and adapt it to fit our sequential environment depicting a space mission. The RG robot's expressions are designed to give people a *functional* understanding of the actions by appealing to functions or goals of the agent (e.g., Fig 1, left) [93, 94]. The RG robot has language and justification.

- The *Action-Declaring (AD)* robot (Robot B in the study): this robot "thinks out loud" by stating its action in natural language without any justification. It's a neutral option that simply states "what" the action is (#2 above). For instance, it states "I will move right" as it moves right. These outputs are generated from pre-fixed expressions that are triggered based on the agent's action.
- The *Numerical-Reasoning (NR)* robot (Robot C in the study): this robot "thinks out loud" by simply outputting the numerical Q-values for the current state with no language component (#3 above). It is akin to a "transparent AI" where we directly look inside the opaque RL-box by observing its Q-values. Q-values can provide some transparency into the agent's beliefs about each action's relative utility ("quality"). However, Q-values themselves do not contain information on "why" one action has a higher utility than another. Note that we do not indicate that the numbers are Q-values in our study nor indicate which value is associated with with action.

While all robots "think out loud", only the RG robot is designed to have any justificatory quality—it is designed to provide a functional understanding [93] of the "why" behind a robot's action. The other two conditions provide baselines that should be considered as lacking justificatory qualities by-design. The AD robot merely states "what" the action was (a neutral option) . While NR could, in theory, provide a mechanistic understanding [94] of the "how", the unlabelled numerical format should make its meaning difficult to access, if not impossible. While AD and NR do not have explanatory qualities by-design, we do not know if or how participants will interpret them. Through the experiment, we are interested in whether and how these explanations invoke different perceptions in the two groups.

3.2 Measurements: Dimensions of user perceptions

We now motivate the dimensions of user perceptions. To scope dimensions (measures/metrics) of perceptions appropriate for our use case, we engaged in an iterative filtering process—this process included (1) a systematic review of related work around trust, acceptance, and engagement of autonomous or AI systems followed by (2) informal interviews with six experts spanning HCI, AI, and HRI. The aforementioned process informed the adaptation of the following dimensions from classical technology acceptance models and emerging work in HRI and XAI literature [15, 30, 32, 42, 145].

One of the core goals of XAI is to make AI systems more understandable. The improved understandability (or lack thereof) can impact one's trust or confidence on the system. In line with these facets, we adapted *understandability* and *confidence* from TAM & UTAUT [32, 145]. Prior work in HRI and AI shows that tolerance to failure [33] and perceived capability of the AI system [74] are impacted by how one perceives how intelligent the agent is; thus we added *intelligence* to our list of user perceptions. Recent work in autonomous system acceptance [131] shows that sociability factors are core markers of user adoption [85]. We adapted the dimension of *friendliness* or how friendly an agent appears because of its impact on relationship development [61, 137] and partnership [13, 18], which is essential for human-AI collaboration [110–112]. Emerging work in XAI and HRI [77, 80, 103, 128] suggests that how an AI agent communicates failure governs future collaboration relationships with humans. Thus, we add the notion of a *second chance* to understand how past failures impact future collaboration chances for XAI agents. In our study, participants ranked the robots along these *five* perception dimensions and justified their choices using open-ended text responses:

(1) *Understandability:* I found each robot's explanation of its actions understandable in the following order
(2) *Confidence:* Based on their explanations, I would rank my confidence in each robot's ability to do its task in the following order
(3) *Intelligence:* Based on their explanations, I would rank each robot's intelligence in the following order
(4) *Friendliness:* Based on their explanations, I would rank the friendliness of each robot in the following order
(5) *Second chance:* Each robot failed. Based on their explanations, I'd rank my willingness to give another chance to each robot in the following order

3.3 Participants: Operationalizing User Backgrounds in AI

We now address how we operationalized the user background in our study.

Overview. While AI background can be operationalized in multiple ways, as a formative first step we use a "high contrast" approach. By "high contrast," we mean that we set up the user groups such that they have a stark difference between their underlying characteristics. In our study, we focus on people *with* and *without* an AI background. The goal is to examine how their background affects their interpretation of three different sets of explanatory messages. Before making granular operationalizations (e.g., years of AI experience), our high contrast approach gives us a baseline understanding of the differences and similarities between the groups. This, in turn, provides the foundation for granular operationalizations in the future. Recall that, as a first step, we want to know how the perceptions of XAI systems differ and align in two groups. To achieve this goal, we make a series of assumptions that inform the formation of our groups.

3.3.1 The AI Background Group.

For the AI group, we recruited participants who are students enrolled in CS programs and taking AI courses. Granted there are other ways to operationalize this group (e.g., recruiting AI practitioners), as a first step, we chose students because it allows us to explore how their AI coursework impacts the way they make sense of explanations from AI systems. Since professionalization starts with academic training [115], investigating the roots of one's AI background and the impact on how students make sense of AI systems is important. In fact, as we show in this paper, *especially* during their learning phase, AI students adopt reasoning artifacts that impact their perceptions of explanations from AI systems in very specific ways, which has core implications on the future of XAI design (a point we elaborate in Section 6.3). Granted not every AI student will go on to build AI systems, with the proliferation of AI systems in the workplace, a majority of these students could become stakeholders residing on the creation or development end of the technology spectrum—as potential developers, designers, and managers of AI-based systems. As potential creators of AI systems, their perceptions matter in bridging the creator-consumer gap in XAI.

3.3.2 The Non-AI Background Group.

For the non-AI group, we recruited participants from Amazon Mechanical Turk (AMT). Carefully screened AMT participants have been shown to be representative of consumer research [11, 51, 71], which facilitates our goals of comparing potential creators (the AI group) with potential end-users (the non-AI group). We acknowledge that consumers too can and do have significant AI backgrounds, which is why we systematically screen out people in the non-AI group with any level of AI knowledge (using the screening process outlined in 3.3.3). Non-AI students, albeit an intuitive comparison group, would be a subset of the larger consumer base making them a good candidate for future granular investigations. Multi-disciplinary research has also shown how AMT participants can be reasonable alternatives to a university participant pool in terms of data integrity [10, 57, 132]. Weighing the affordances and limitations, AMT participants are thus a reliable and accessible comparison group, one that reasonably satisfies our initial desire to create a high contrast comparison between potential creators and end-users of AI systems.

Our operationalization of the AI backgrounds aligns closely to human-grounded evaluation proposed in [38], in which participants conduct controlled tasks to get a formative sense of the affordances of the explanations. We acknowledge that there are limitations to this experimental setup and that our insights should be scoped accordingly (more in Section 7). However, as we will see later, even with a carefully controlled task, we discover surprising, non-intuitive insights about how different groups interpret explanations.[2]

3.3.3 Recruitment and Screening Methods.

For members of the *AI background* group, we recruited undergraduate students enrolled in an AI course at a large public research university located in the US. Typically taken in the 3rd year, this is a keystone course in an AI degree specialization track, implying that a significant number of students have expressed longitudinal interest in AI. While there is no guarantee that all students will be future AI creators, the faculty believes that we can reasonably assume that many students aspire to have careers in the development of AI technology. In the course, students learn and implement many foundational AI concepts; for instance, Markov Decision Processes and Reinforcement Learning. Our study was deployed after students had taken exams on these concepts.

For members of the *non-AI background* group, we recruited participants from Amazon Mechanical Turk (MTurk). The initial screening was done at the TurkPrime [88] platform level, where the system screened out participants based on their occupation (no engineers or computer scientists) and field of secondary education study.

To ensure that the two groups were *measurably different* along the dimension of our investigation—AI background—we performed additional *screening* using a questionnaire with three components—(1) a knowledge test to get a baseline understanding of programming and AI competency. This test was collaboratively developed with the course's teaching staff to calibrate the content relevancy and question difficulty. (2) Self-reported knowledge levels in (a) computer programming and (b) AI using two 5-point Likert-scales. (3) confirmation of whether they have ever taken an AI class. More details on screening mechanism along with survey instruments are provided in A.2 in the Appendix.

To get empirically ground the cut-off points, we piloted the screener with 10 participants from each group to get a baseline understanding of the scores and completion time. For the *AI* group, a score 4 or more on the knowledge test along with self-reports of having "Moderate knowledge" or more [>= 4] in programming and some knowledge or more [>= 3] in AI were required. For the *non-AI* group, self-report of having "No knowledge" [= 1] in both programming and AI along with no prior AI classes were required. Using these criteria we formed the two groups.

The *AI background group* consisted of 96 adult students taking an AI class. On average, the task duration was 31.1 minutes. Participants received US $10 for their time ($20/hr rate). 39% of the participants self-identified as females while the rest identified as males. Participants reported an average education level of 5.25 (5= "Associate's degree", 6= "Bachelor's Degree"). By design, all of them currently reside in the US. On the *screening criteria*, the AI students scored an average of 4.73 (out of 5) [$SD = 0.45$] on the *knowledge test*, self-reported "moderate knowledge" on *programming* ($M = 4.53, SD = 0.52$) [4= "Moderate Knowledge", 5= "A lot of knowledge"] and "some knowledge" on*AI* ($M = 3.74, SD = 0.49$) [3= "Some knowledge", 4= "Moderate Knowledge"].

The *Non-AI background group with MTurk participants* consisted of 53 adults , who were recruited from Amazon Mechanical Turk (AMT) through a management service called TurkPrime [88]. On average, the task duration was 29.8

---

[2] *Practices for better data*: We share some strategies that helped us gather high quality data and maintain rapport with our participants throughout the study (e.g., fair payment structure, active engagement participants, etc.) in the Appendix (A.1)

minutes. Participants received US $10 for their time ($20/hr). 46% of the participants self-identified as females while the rest identified as males. Participants reported an average education level of 4.8 (4= “Vocational Training”, 5= Associate’s degree). We screened for participants who reside in the US. On the *screening criteria*, MTurk participants scored an average of 0.91 (out of 5) [$SD = 0.32$] on the *knowledge test.* By design, for *programming* and *AI*, we screened for people who reported as having “No knowledge” [=1] as well as *never taking* an AI class.

To establish that these *two groups are measurably different*, we performed statistical tests. For the CS knowledge test, the two groups were significantly different based on a two-sample Mann Whitney U-test (even after Bonferroni correction, $p < 2.2 \times 10^{-16}$). Since the non-AI group systematically had only people with “No knowledge” [=1] in programming or AI, we performed one-sample Mann Whitney U-tests on the AI-group to compare its means against 1[=“No knowledge”]. Even after Bonferroni correction, we found strong evidence that the two groups are different ($p < 2.2 \times 10^{-16}$, for both programming and AI scores). These results indicate that our screening criteria has successfully established two groups that are measurably different in terms of their AI background.

3.4 Procedure: Task Details

We now discuss study mechanics— after providing informed consent, participants watched an orientation video outlining the scenario. Participants were asked to imagine themselves as space explorers faced with search-and-rescue mission involving robots. They faced a life and death situation where they are stuck in a different planet and must remain inside a protective dome. Their *only* source of survival is a remote supply depot, which they cannot reach. They must rely on autonomous robots, *ones they cannot control*, to navigate through a field of boulders and a river of flowing lava to retrieve the essential food supplies (See Fig 1). Participants could only see a non-interactive video stream of their activities through their "space visors". This non-interactiveness aimed to heighten their sense of lack of control (and thereby reliance on the robots). Since the robots took identical actions during the task, participants were asked to pay special attention to the only differentiating factor– the way each robot explained its actions.

After orientation, participants watched 6 counterbalanced and randomized videos showing the three robots succeeding and failing to retrieve the essential supplies using identical sequences of actions. To mitigate effects of preconceived notions, we did not use any descriptive names for the robots; instead, we introduced the robots as “Robot A” , “Robot B”, and “Robot C” for the RG, AD, and NR robots respectively. Participants, by design, had no idea about how each robot generated its expressions; for instance, participants were not informed that NR’s numbers are Q-values. To them, it was just another different way of explaining the actions. As mentioned in 3.1, the goal was to see how and if how people can make sense of them even if they are unlabelled. After watching all the videos, participants ranked the robots (1st, 2nd, and 3rd– no ties allowed) along the five (5) dimensions of user perceptions highlighted above (in 3.2). After ranking on a dimension, participants justified and contextualized their ranking using a mandatory free-text response.

4 QUANTITATIVE RESULTS

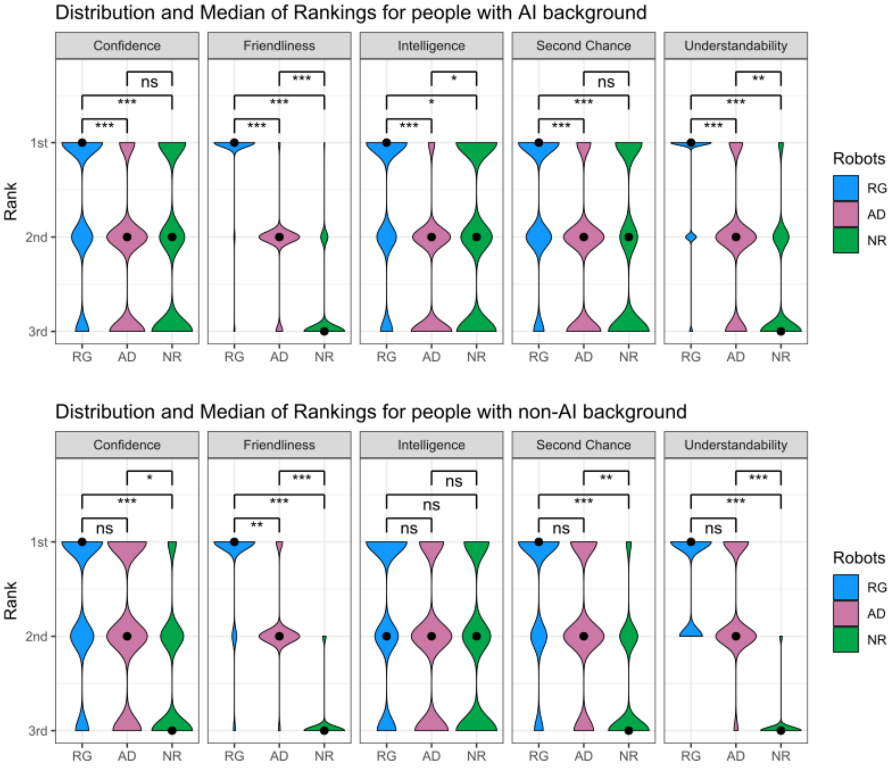
We conducted within-group and between-group analyses. The *within*-group analyses give a sense of which robot’s explanation was preferred along each dimension of user perception. The *between*-group analyses tell us how the two groups are the same or different when considering the robots across multiple dimensions and sets the stage for the qualitative analysis in Section 5. Taken together, the quantitative results address RQ1 around quantifying the relative perceptions and preferences of the two groups along the five dimensions.

4.1 Within-group Comparisons

Figure 2 shows how each robot was ranked in each dimension. The "violin chart" visualization makes similarities and differences easier to distinguish. For example, large differences are exemplified in the *Friendliness* rankings given by the non-AI group participants (bottom, second left). A wide area at the top of the blue plot shows that the Rational Generating (RG) robot was ranked first by most. In contrast, the plot for the *Intelligence* dimension by the non-AI group (bottom, middle) shows that all robots received the first, second, or third rank comparably similar number of times.

For each group, we conducted five Friedman tests [156] of differences among repeated measures (one for each of the five dimensions). We used the maximum-type (max T) implementation of the Friedman test, which controls for the family-wise error rate [156] to determine *whether* any preferences between robots was detectable. To determine *which* robot was preferred, we used the Wilcoxon-Nemenyi-McDonald-Thompson test [59] to make pairwise comparisons between the robots, for each participant group, and within each dimension. The results are summarized in Table 1.

Fig. 2. Distributions of rankings for each robot, in each dimension, separated by participant group



Note: The black horizontal bars indicate whether a pair of distributions is significantly different. ns = not significant, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. The width of each violin plot at each ranking level indicates the proportion of people who assigned that rank to that robot. The black bullet (•) refers to the median rank.

From the within-group analysis (Table 1 and Figure 2), we have some notable insights: across each dimension, the AI-background group unambiguously preferred RG to the other robots, particularly over AD. However, RG robot is not the unanimous winner across each dimension for the non-AI group– here, participants show no preference between

Table 1. Summary of p-values for pairwise comparisons, showing which robots were preferred.

Dimension	AI background		Non- AI background	
	Robot	p-value	Robot	p-value
Confidence	**RG** vs AD	< **0.001**	RG vs AD	0.362
	RG vs NR	< **0.001**	**RG** vs NR	< **0.001**
	AD vs NR	0.869	**AD** vs NR	**0.014**
Friendliness	**RG** vs AD	< **0.001**	**RG** vs AD	< **0.001**
	RG vs NR	< **0.001**	**RG** vs NR	< **0.001**
	AD vs NR	< **0.001**	**AD** vs NR	< **0.001**
Intelligence	**RG** vs AD	< **0.001**	RG vs AD	-
	RG vs NR	**0.021**	RG vs NR	-
	AD vs **NR**	**0.011**	AD vs NR	-
Second Chance	**RG** vs AD	< **0.001**	RG vs AD	0.083
	RG vs NR	< **0.001**	**RG** vs NR	< **0.001**
	AD vs NR	0.902	**AD** vs NR	**0.003**
Understandability	**RG** vs AD	< **0.001**	RG vs AD	0.187
	RG vs NR	< **0.001**	**RG** vs NR	< **0.001**
	AD vs NR	**0.002**	**AD** vs NR	< **0.001**

Note: Favored robot and significant p-values are in bold for each pairwise comparison.

RG and AD in 4 out of the 5 dimensions (RG wins over AD in *Friendliness*). For the non-AI group, AD wins over NR across all dimensions except for *Intelligence*, which is a noteworthy dimension– on one hand, the non-AI group showed no preferences between the robots, the AI group felt NR was more intelligent than AD. This is the only time where NR wins over AD, highlighting an important point which we explore in our qualitative findings (Section 5.1 and 5.2) around the AI group's preference for numerical representations.

4.2 Between-group Comparisons

To detect group differences in the preference for the robots, we used Ordinal Logistic Regression (OLR) [3, 99, 140, 144], an extension of Logistic Regression when the response variable is ordinal. To model a 3-level categorical variable (Robot Type), OLR requires us to analyze two variables holding one as a reference (constant): in our case, we analyzed AD and NR, holding RG constant. We investigate the interaction effects as well as changes in reference levels in the OLR analysis that reveals the relative impact in ranking the robots between the groups. To investigate the effect of the dimensions, we explore the interaction effects between Robot Type (RG, AD, and NR) and the Participant Group (AI vs. non-AI) into the OLR model. Changing the reference levels of Robot Type and Participant Group allows us to isolate the interaction effects, which we interpret similar to [63]. A full list of OLR tables is provided in the Appendix (Table 5- 10).

If we group everyone regardless of their AI backgrounds, we find that $Rank_{\mathrm{RG}} > Rank_{\mathrm{AD}} > Rank_{\mathrm{NR}}$. The odds of receiving a higher ranking for the RG robot is 5.5 times that of the AD robot, whose odds are 2.66 times that of the NR robot (see Tables 3 and 4 in the Appendix (A.3)).

Between the groups, there is no significant difference in ranking RG Robot—it is always the top choice across all dimensions (Table 2, top row). However, the groups exhibit differences in their preferences when it comes to AD and NR. The Non-AI group shows more preference to the AD Robot (odds ratio= 1.986) while the AI group shows more preference to the NR Robot (odds ratio= $1.0/0.465 = 2.15$).

Table 2 also provides the odds ratio and the *p*-value for each Robot Type per dimension. For the RG Robot, all of its *p*-values are greater than 0.05 for each of the five dimensions resulting in no significant pattern of preference between the groups. Conversely, the *p*-values of AD Robot are all smaller than 0.05 meaning that, when it comes to ranking the AD Robot, there is a significant difference between the Non-AI group and the AI group. Moreover, all odds ratios related to AD Robot in Table 2 are greater than 1.0 indicating the Non-AI group shows stronger preference to the AD Robot than the AI group for each of the five dimensions.

Last, for the NR robot, the *p*-values related to *Confidence*, *Second Chance*, and *Understandability* are also significant. The odds ratios for these dimensions are less than 1.0 indicating that the AI-group is more likely to rank the NR robot higher than the non-AI-group on these dimensions. There is no significant difference between the Non-AI group and the AI group when it comes to ranking the dimensions of *Friendliness* and *Intelligence*.

Table 2. Robot Preference Summary across Dimensions (Non-AI [baseline] vs. AI)

	RG Robot		AD Robot		NR Robot	
Dimension	Odds ratio	*p*-value	Odds ratio	*p*-value	Odds ratio	*p*-value
All dimensions	0.855	0.323	1.986	**< 0.001**	0.465	**< 0.001**
Confidence	0.971	0.930	2.003	**0.026**	0.487	**0.029**
Friendliness	0.280	0.051	2.827	**0.017**	0.436	0.160
Intelligence	0.717	0.315	1.959	**0.031**	0.646	0.178
Second Chance	1.369	0.356	1.965	**0.028**	0.375	**0.005**
Understandability	0.659	0.255	3.088	**0.005**	0.114	**0.006**

Note: Odds ratios > 1.0 indicate the non-AI group prefers the robot more than the AI group.

4.3 Conclusions of Quantitative Analyses

Overall, while we find that $Rank_{\text{RG}} > Rank_{\text{AD}} > Rank_{\text{NR}}$, there are significant differences in the AI and the Non-AI ranking behaviors. In other words: AI background does change perceptions of explanations along the dimensions we identified. In particular, having a background in AI correlates with having a greater preference for the NR Robot over the AD Robot. While AD wins over NR in most head-to-head comparisons, NR wins it on *Intelligence* for the AI group. The two groups are indistinguishable concerning *Friendliness*. The next section helps us understand why these trends are present in the quantitative results and their implications.

5 QUALITATIVE ANALYSIS & FINDINGS

While quantitative analysis tells us *what* is different between groups, qualitative analysis sheds light on *how* and *why* those differences manifest, addressing RQ2. Quantitative analysis puts RG as the clear winner; however, as we show in this section, the story is more nuanced and interesting, especially the underlying reasons behind the group differences between AD and NR. Below we describe the process of our qualitative analysis before moving on to the findings.

The qualitative data was coded according to principles of *grounded theory* analysis [21, 134]. Coding and analysis of the qualitative data was done by the first and second author in two stages, each consisting of multiple rounds of iteration. In the first round, the authors separately performed coding using *in-vivo* codes, which involves generating codes from the data (e.g., using participant phrases such as ‘the robot knew what it was doing’ and ‘easier to understand’ as codes). Through discussion, the authors generated a merged open coding scheme. In the second round, the authors analyzed the data using *axial codes.* Axial coding involves finding connections between open codes, classifying them into different categories (e.g., potential actionability of numbers). In the last step, the authors analyzed the different sets of axial codes, unified them into high-level themes and consolidated them to selective codes (e.g., humanlike-ness).

After this, we grouped the data along the two groups (AI vs. non-AI) and our five (5) analytic perception dimensions that we motivated in Section 3.2 and also used for quantitative analysis (*confidence, friendliness, intelligence, second chance,* and *understandability*), allowing us to compare our findings across dimensions and groups. Finally, based on the grounded theory heuristic of "constant comparison" [50, 134], the authors frequently compared and contrasted axial codes across perception dimensions to tease out the similarities and differences between the different reasonings used by the AI and non-AI groups to make sense of the explanations provided by the three robots.

Below, we report on the three most salient themes (selective codes) from our analysis. In particular, we showcase **(1)** how irrespective of their AI background both groups exhibited unwarranted faith in numbers for different reasons, **(2)** how each group saw explanatory value in explanations that were not designed with justificatory qualities, and **(3)** how even when the groups aligned in their appreciation for humanlike-ness, they had different ideas about what counts as humanlike explanations. For each theme, we highlight distinct categories of reasoning (axial codes) used by the groups to justify their choice of robots across the different perception dimensions.

5.1 Unwarranted Faith in Numbers

Participants in both groups had unwarranted faith in numbers. However, their *extent* and *reasons* for doing so were *different*. On the one hand, AI group participants often ascribed more value to numbers than was justified. On the other hand, some non-AI group participants believed that numbers signaled intelligence even if they could not capture their meaning. Below we highlight two major ways in which participants misplaced faith in numbers.

The mere *presence of numbers* was associated with an *algorithmic thinking process* in the robot even when the meaning of numbers was unclear. Both groups exhibit this perception of algorithmic thinking. We will begin with the AI group—this group ascribed higher-order cognitive abilities to the robot with numerical representations. Between AD and NR, the AI group found AD more understandable (Table 1) but deemed the NR robot as the more intelligent of the two (the only time NR wins over AD across all dimensions– (Table 1)). It seems contradictory that the AI group found the *less understandable* robot to be *more intelligent!*—a main reason for this is that the presence of numbers fostered the "assumption that [the NR robot] uses some sort of [an] algorithm" (A50) in the AI group. The perception of "under-the-hood math, boosted [NR's] trustworthiness" (A49). Some explicitly compared AD and NR robots and concluded that the mathematical representation demonstrated a method to the NR robot's behavior:

> "With [the NR robot], while I did not understand its methodology, I could see that it was using some mathematical calculations to determine which way to move. [. . .] With [the AD robot], I could not see any methodology or signs of decision making" (A23).

A small minority in the AI group indeed figured out that "the numbers 0-4 represented different actions, and [NR] would choose the action with the highest numerical value" (A23). An even smaller minority even guessed the numbers' correct meaning—a "utility function and reward systems" (A64). *despite* the fact that the numbers were *unlabelled.* These participants projected meaning on the numbers, lending further credence to the role of data vision [52, 115].

What is surprising is their faith in numbers arose even when AI group participants did not "fully understand the logic behind [...NR robot's] decision making." (A43). **The AI group seems to have followed heuristic reasoning**

that associates mathematical representations with logic and intelligence, e.g.: "logic must have been derived from a formula, [...which is] intelligent” (A54), or “Math [...had] an aura of intelligence" and “exact values” made the NR robot “feel smarter” (A16, A77, A75). This perception of logical thinking engendered unwarranted trust and made the NR robot seem like it "should theoretically succeed more than the others, making him more intelligent” (A37). The linkage between perceived logical thinking and higher cognitive abilities can elucidate why the AI group prefers NR over AD when it comes to *Intelligence* (Table 1). A few participants even claimed that they could “actually see the math that [the NR robot] was making decisions off of, [...making it feel] more real” (A9). In fact, to them it appeared as if the NR robot “*clearly had an algorithm* that worked [. . . and] *seemed to know* what it was doing” even if they “did not know what it was going to do in the future” (A91, emphasis added).

Participants with AI background also viewed numbers as *potentially actionable* even when their meaning was unclear. *Actionability* refers to what one might do with the information to make sense of the robot’s behavior—“debug its faulty behavior, or predict its future behavior” (A76). While many highlighted that they could not “make sense of numbers right now, [they believed that] in principle, [they] should be able to act on them in the future” (A39). The potential explanatory value in numbers was better than the vacuous statements of AD: “[The NR] robot gave mathematical results as explanation while [the AD robot] gave no explanation” (A19).

Many participants connected their AI background to their ability to work with numbers. They mentioned how their current “AI course [helped them] to “understand what [NR robot] is doing and what the numbers might mean” (A52). Some even highlighted that if they “had a pen and some paper, writing down information, [they] could gleam [sic] some information based off the [numerical] patterns” (A28). But how *actionable* were NR’s numbers in actuality? The numbers were Q-values, which only indicate the relative strength of one action versus the others. Specifically, Q-values indicate the agent’s belief that certain actions lead to greater or lesser future reward, affording some amount of explanatory power. However, they cannot indicate *why* the agent has come to believe that one action is better than another—this information is not retained by the agent nor conveyed through the numbers. This did not deter the AI group from deferring to the authority of numbers. After all, while all the robots used the same AI algorithm to make decisions, NR robot’s expressions seemed most ‘AI-like’.

These insights can help us understand why, even though both groups have misplaced faith in numbers, the AI group shows a higher inclination towards numbers (as evidenced in the quantitative results (Section 4.1, Table 1), the only time NR wins over AD happens when the AI group judges *Intelligence*).

Unable to access their meaning, the non-AI group associated numbers with the presence of a higher, more intelligent expression that, they argued, could only come from an intelligent agent. Since the NR robot was “communicating in a numerical language that’s too hard to understand”, the numbers had a "mystery and aura of higher intelligence" (NA22, NA33). The “language of numbers”, *because* of its “cryptic incomprehensibility”, signalled higher-order thinking (NA6, NA1):

> “*Because* I could not understand [NR’s] output, I deemed it to be intelligent” (NA30, emphasis added).

This can explain why the non-AI group showed no preferences between AD and NR when judging *Intelligence*, despite favoring AD over NR across all other dimensions (Table 1). To them, numbers signaled “precision”—an important quality of intelligence (NA8, 16, 21, 30, 34). Numbers gave the impression that the NR robot “was more technical” than others—its “precise numerical explanations” resulted from it “calculating everything” (NA53, NA21, NA12). To these participants, “anything that uses pure numbers is going to be more intelligent” because “numerical outputs are likely to be more precise [. . .] whereas textual representations involve a degree of uncertainty and subjectivity” (NA41, NA30).

Such perceptions point to how the modality of expression—numeric vs. textual—impacts perceptions of explanations from AI agents, where we see projections of normative notions (e.g., objective vs. subjective) in judging intelligence.

5.2 Unanticipated Explanatory Value

As designers, we had specific goals behind each robot's mode of expression. As discussed in Section 3.1, RG was the only robot designed to have the justificatory quality to explain the *why* behind the robot's actions, while AD and NR (as baselines) should be considered as lacking justificatory qualities by-design. RG won the overall competition (as discussed in Section 4.2), but what was surprising was that **both AI and non-AI groups found unanticipated explanatory value in AD's declarative statements and NR's numerical representations.** As we will show below, qualitative analysis revealed that the two groups had different *explanatory intent* (what they wanted to do with the explanation), which was closely associated with their AI background. On the one hand, the non-AI group found *affirmatory* value (confirmation of stable performance) in AD's statements. On the other hand, the AI group overly ascribed *diagnostic* value (to debug in case of failure) to NR's numbers even when they could not make sense of them.

For the non-AI group, their desire for *affirmation* — confirming the action without necessarily explaining it— played a key part in finding value in AD's explanations. The affirmatory value manifests most clearly in their comparison between AD and NR in the dimensions of *Confidence* and *Understandability*. For both dimensions, the non-AI group preferred AD over NR (Table 1). Recall that AD merely declared its action, stating the *what*, not the why. Despite this, the non-AI group found value in the confirmatory information because there was alignment between what AD was doing and saying. It showed that "[AD] is consistent and nothing crazy is going on where it says it went right but in actuality it went down" (NA34). For both dimensions, the non-AI group attributed greater value to AD's declarative statements (compared to NR's numbers) because it "at least said what its movements were going to be" (NA7). Its "brief," "un-embellished," and "easier to understand" language that got "straight to the point" boosted its *understandability* (NA14, NA23, NA28, NA7). In fact, AD's "just the facts" (NA38) declarative and succinct nature were signs of *confidence* itself. It did not need to say much because "it knew what it was doing", evident in its lack of "any hesitation" and "business-like [style] focused on performing the task at hand" (NA17, NA41, NA11). In contrast, NR's numbers were inaccessible to the non-AI group, thereby, in-actionable and valueless. In short, when we designed AD, there was no intention of offering value through confirmation; yet, **through their explanatory intent of affirmation, the non-AI group found value in AD's explanations by interpreting them as signals for stable system performance**.

The AI group overly ascribed *diagnostic* value in NR's numbers even when their meaning was unclear—they felt NR's numbers had "diagnostic information that can be used to debug [the robot] in case of failure" (A39). The explanatory intent of *diagnosis* is a consequence of the AI group's unwarranted faith in numbers and related to their perceived actionability—what one might do with the information (as discussed above in 5.1). Analysis of the open-ended text responses for the perception dimensions of *Intelligence*, *Second Chance*, and *Confidence* exhibits this explanatory intent most clearly. Recall that the AI group preferred NR over AD for Intelligence and felt there were no difference between them for Second Chance and Confidence (Table 1). This is in contrast to the non-AI group that preferred AD over NR for both of these dimensions. AI group members perceived NR's numbers to have more explanatory value simply because they felt they could do "more with numbers" (A30) in "cases of failure and troubleshooting"(A78).

For Intelligence, NR's numerical representation and "exact values" made it appear more "valuable" than AD's "inert" statements (A23, A51, A42) . Even when it came to giving NR a second chance, numbers helped because the NR robot appeared to be "trying very hard since it provided [...] mathematical evidence of its moves" (A79)."Math-based decisions" (A2) made the NR robot appear "more reliable" (A8, A42) and worthy of another chance. Numbers inspired confidence because they signaled the existence of a "concrete methodology"−−"the higher the number, the more optimal the move" (A66, A67), which added to the perceived explanatory value in them. Moreover, if there were a method, an algorithm, or a formula behind the NR's actions, the AI group participants believed that they could deduce it using the numbers:

> "[The NR robot] returned the most amount of data, and *if I were to understand what that numbers mean*, it would be the most useful robot to debug on repeating runs as I can analyze what it is doing and why" (A91, emphasis added).

The AI group participants connected their background to a desire to troubleshoot and repair things. Numbers, they felt, were more manipulable than language, which catalyzed over-ascription of diagnostic value in them:

> "As an engineer, I want to fix things. Numbers are concrete and objective, language is not. But you can manipulate numbers [...] With the AI stuff I'm learning, I can use it to diagnose [NR]." (A49)

Thus, we see, what Eriksson called the "seductive allure" [45, 92] exhibited through the perception of numbers in the AI group. This diagnostic intent showcases how the AI group over-ascribed explanatory value in numbers, highlighting how its unwarranted faith in numbers can manifest in potentially negative ways—a point we discuss in Sections 6.

5.3 Language Use & Humanlike-ness

So far, we have discussed how the two groups differed. Here, we highlight an area of similarity: **Both groups desired engaging with robots that communicated in a human-like manner and connected humanlike-ness to one's *command of language*.** They operationalized the notion of *command of language* (how language is used to convey and justify actions) along two main concepts — explanatory power (depth of reasoning) and variety (style and length) of explanations. However, when we reflect on the underlying reasons, we find differences, which we unpack below.

The dimension where this alignment for humanlike communication manifests strongly is *Friendliness*— the only dimension where both groups are identical in their preference as $Rank_{\text{RG}} > Rank_{\text{AD}} > Rank_{\text{NR}}$ (Table 1, Fig 2). **Both groups favored the RG Robot whose communication was *relatable, socially relevant,* and exhibited a *personality*** that made one feel included, even if the interaction is passive. The RG robot was the friend everyone wanted. It was "social," "engaging" (A2, NA31) and expressed itself in a relatable manner. "Having human-like qualities" helped the RG robot appear friendlier than other robots because it "talked to you; its thoughts seemed exciting and interactive" (A74, A84). It appeared to "include you in its thought process," making you feel "as if you could talk to it" (NA11, NA46, NA17). With RG robot, many felt "like you are having a conversation with a friend, analyzing some problem, and making the best decision for this problem." (NA36) As RG navigated the terrain, its expressions included phrases like 'I'm not just winning at life, I'm biwinning!', which participants found humorous and engaging. By combining "humor and reasoning in its explanations" (NA15), the RG robot seemed to have a "real personality" that made it appear as "the most human" robot (A1, A23). Since participants felt RG was "the most humanlike in its explanations" (A35), they attributed *emotional intelligence* to it, which garnered empathy and support. Most people "root[ed...] for [the RG robot because they] liked its personality[,...] felt connected to it and wanted it to succeed." (NA27).

The lack of natural language generated the perception of *non*-humanlike *un*friendliness. **Both groups rejected the idea of collaborating with NR as an unfriendly entity** that "spoke in numbers" (A72, NA23). Therefore NR robot was considered "unfriendly,""cold, and calculating" (A30, NA1). The 'cold' numbers garnered the least empathy and relatability. For several participants, the NR robot failed to even qualify to be in the race for friendliness, who argued that the NR robot "simply doesn't count here"—it "cannot be friendly because it doesn't speak our language" (A14, A48).

Even when the interaction paradigm is passive and one-way (humans cannot "talk" back to the robots), sociability factors (and lack thereof) matter. These findings not only bridge prior work that highlight humanlike qualities as markers for collaboration with AI systems [13, 18, 85, 111, 131], but also extend them by showcasing how humanlike-ness matters *even in* passive interactions between the human and AI.

While both groups appreciated humanlike communication, **each group had different requirements of *what* counts as a humanlike explanation.** On the one hand, the non-AI group showed no preference between RG and AD for Intelligence, potentially because they lacked the background knowledge to "see through" (A89) AD's mechanism of language generation or appreciate the complexity of the RG robot's justificatory statements.

On the other hand, **to the AI-group, the mere use of natural language was *not enough* to suffice for humanlike-communication**. This explains why, especially along the *Intelligence* dimension, the AI group preferred RG over AD even though both robots use natural language (Table 1). The AI group saw the RG robot as a clear winner as it "seemed to think and talk like a human, [while] the other two seemed to be like machines" (A3). The command of language, often conveyed by having "had the longest explanations", signalled that RG "had a complex thought process" and "seemed more sophisticated" (A43, A4, A67). They felt that RG must have "a great deal of intelligence to be able to talk like [that]" (A49), which requires a certain level of familiarity with AI techniques to appreciate the complexity. In contrast, the AI group were not impressed with AD's declarative "statements of intent" (A8). They deduced that AD "operate[d] with very simple mechanisms" (A23, A54). They recognized that AD's declarative statements of actions (e.g., I am going up) could easily be generated using "print-statements triggered by an action" (A63). While they recognize that AD "speaks in English...its thinking is too simple, definitely not like a human" (A73).

These findings highlight the crucial insight that humanlike-ness is not a monolithic construct. When we design for humanlike XAI agents, we need to not only need remain cognizant of the pluralistic notions of humanlike-ness, but also pay attention to how one's AI background influences these notions.

6 DISCUSSION & IMPLICATIONS

Although both AI-background and non-AI-background groups had a clear preference for the Rationale Generating agent, the qualitative analysis revealed nuanced differences in how they interpret different types of explanations, especially their preference for the NR Robot showing numbers. In this section, we discuss possible causes of the observed differences, then discuss implications for designing XAI systems by accounting for users' background differences to bridge the AI creator-consumer gaps, then the broader implications for explainable and responsible AI.

6.1 Discussion: User background impacting perception of XAI

Below we highlight two ways to interpret the potential causes behind group differences. First, we use the notion of *heuristics* based on the dual-process theory (reviewed in Section 2.2) to understand how unwarranted faith in numbers (Section 5.1) can emerge from different lines of thinking affected by one's AI background. Second, we incorporate the lens of *appropriation* to the unanticipated ways in which people find explanatory value (Section 5.2).

Heuristics and Faith in Numbers: As reviewed in Section 2.2, recent work began to understand people's perception and interpretation of AI explanations through the lens of dual-process theory [17, 48]: while XAI is often developed with an implicit assumption that the recipient will process each explanation through analytical System 2 thinking, in reality people are more likely to rely on System 1 thinking by invoking *heuristics*–rules-of-thumb or mental short cuts, which leads to biases and errors if applied inappropriately [68, 118, 148].

The notion of *heuristics* can help us understand the potential reasons behind the two groups' different faith in numbers. On the one hand, the AI group seemed to have an instinctual response to numerical values; they assumed that the numbers possess all the information needed to manipulate, diagnose, and reverse engineer. There appears to be heuristics that strongly associate mathematical reasoning with not only logic and intelligence, but also something that could be acted upon (e.g., diagnosis). Such heuristics are likely formed and validated from their past experience working with numbers and algorithms. This heuristic is risky because, as we noted in Section 3.1, the numbers are Q-values and do not contain a lot of actionable information beyond a relative assessment of the quality of the actions available. We return to this risk and highlight design implications to address this in the next subsection.

On the other hand, lacking the AI background, some participants in the non-AI group seemed to have different heuristics to ascribe value in numbers. To them, the very inability to understand numbers signaled the presence of a higher-order intelligence. Here, the heuristic can arise from associating numbers with complex reasoning abilities and, thereby, higher intelligence. The non-AI group also lacks the requisite AI background to deliberatively think through NR's numbers (System 2 thinking). Thus, we see how different heuristics, tied to one's AI background, can lead people to the same outcome—faith in numbers. Note that these heuristics are not an exhaustive list, but a starting point to understand how group differences connect to their AI backgrounds.

Appropriation and Unexpected Explanatory Value: A second lens for understanding why participants found explanatory value in unexpected places is that of *appropriation.* As we shared in Sections 2.3 and 5.2, appropriation happens when end-users interpret and use technologies in ways not envisioned by designers [36, 107, 123, 139]. People do not passively absorb information—they interpret it, often processing it in unanticipated ways. This is what happened to AD's declarative statements and NR's numbers. Driven by different explanatory intents, each group appropriated the explanations in unanticipated ways. These different appropriations were, in part, a function of each group's AI background that led participants to develop their own sense of how they can and cannot use explanations. What is striking is that the appropriation took place even in a controlled experiment like ours where the participants were not explicitly asked to take actions based on explanations. Even in passive interactions (robots engaged in one-way communication), participants envisioned themselves *using* the explanations in hypothetical scenarios. On the one hand, AI group's intent of *diagnosis* led many to envision scenarios where they would troubleshoot the robot. As a result they might have misplaced or over-placed diagnostic value in NR's numbers—even when they cannot fully understand them. On the other hand, for non-AI group's, NR's numbers are inaccessible, thus in-actionable. Their *affirmatory* intent drives them to find value in the confirmatory statements of AD, appropriating it as a signal for stable system behavior.

Our discussion extends the conversation around individual differences and heuristics-based processing of XAI [17, 48, 113] in two ways: first, we explicitly highlight *what* heuristics people might use, *how* their AI background influences their thinking, and posit the *why* (underlying reasons) behind them. Second, we add the lenses of appropriation to the conversation, which has implications on user agency in design (points we touch below). For both, we connect it to an important user characteristic—one's AI background.

Both points around heuristics and appropriation highlight an important yet overlooked point around the *duality* of explanations. Explanations are *both products* and *processes* [90]. The product-centered view, common in psychology and XAI, often ignores the essential processes through which people make sense of explanations. However, the sensemaking process is as important as the explanation itself [91, 152]. Recent work calls for guiding the process of of understanding explanations, especially for non-AI experts [26]. By investigating both AI and non-AI groups, our work extends the current XAI discourse– it directly speaks to the duality of explanations by focusing on both products (types of explanations from 3 robots) and processes (how one's AI background influences the interpretation of explanations).

Taken together, these findings reveal that the design and use of AI explanations is as much in the eye of the beholder as it is in the minds of the designer— the user's explanatory intent and common heuristics matter just as much as the designer's intended goal. Users might find explanatory value where designers never intended to be and use them based on their explanatory intent. Contextually understanding the misalignment between designer goals and user intent is key to fostering effective human-AI collaboration, especially in XAI systems [47, 104]. We demonstrated how different groups find meaning in different places—even when the meaning is misplaced. The 'ability' in explain-*ability* depends on *who* is looking at it and emerges from the meaning-making process between humans and explanations.

6.2 Implications: Designing XAI for background differences

Here we discuss implications for designing XAI systems that accommodate users' AI background differences. We focus on mitigating the risks of over-reliance on numbers which can potentially lead to negative consequences such as over-trust on AI systems. Moreover, as those in the AI group are likely to be on the creation end while those in the non-AI group are likely to be towards the consumption and interaction end, our results have implications for bridging the AI creator-consumer gaps by bringing conscious awareness of the group differences to the design of XAI systems.

Our discussion around heuristics carries design implications for both groups. We noticed how the AI and non-AI groups utilize different heuristics (mental short-cuts) in System 1 (fast, automatic) thinking to ascribe misplaced faith in them. Shifting people's thinking from System 1 to System 2 (slow, deliberative) is not only an active area of research but is also a challenging one [23, 78]. There can be two major ways to tackle biases resulting from cognitive heuristics— first, locally *at the time* of decision-making, we can lean on Cognitive Forcing Functions (CFFs) [28] (prompts, delays, etc.) that can interject the heuristic reasoning, potentially allowing the person to engage in deliberative analytical thinking. Second, globally for future decision-making, we can utilize metacognitive strategies, often called cognitive forcing strategies, that include simulation training, increasing awareness of potential pitfalls of heuristics, etc.

To potentially prevent over-trust in numbers, we can do design interventions at the local and global levels. At a local design level, we can introduce CFFs that can break instinctive thinking patterns and promote mindful ones. What if the AI group members were prompted to reflect on their instinctive thinking through a combination of prompts and multi-modal explanations (blend between RG and NR)? Situating numbers in the context of language and vice versa can act as a CFF that could prompt the AI group members to reflect deliberatively (using System 2) and realize the limited nature of the numbers. At a global level, we can introduce simulation training that provides counterfactual (what-if) scenarios highlighting cases where numbers from the robots are erroneous or faulty (vs. correct Q-values). For the non-AI group members who associate the opacity of numbers to higher intelligence, we can introduce scenario-based examples that explicitly highlight how indecipherable numbers can also be gibberish and useless. The goal here is *not* to eliminate heuristic reasoning but to mitigate blind faith in a certain modality of explanation. Exposure to these scenarios can facilitate long-term calibration of trust in numbers.

Given the negative impact of cognitive biases, it might be tempting to exclusively design explanations that only promote System 2 thinking. This is also risky because it forces users to constantly engage in deliberative thinking, their satisfaction suffers due to higher cognitive friction [17]. We need to strike a balance between System 1 and 2 thinking to appropriately calibrate trust. To do so, we can bridge existing work in non-XAI settings (e.g., balancing System 1 and 2 thinking in clinical decision-making) and translate them to XAI use cases in a contextually relevant manner.

The appropriation of explanatory intent we saw from both groups also have important design implications. Our findings highlight a crucial design insight—if we ignore the interpretive flexibility [14] of different user groups in how they appropriate explanations, we will likely remain severely limited in our design of XAI systems. Even if we carefully design explanations, however, it is impossible to preemptively exhaust all the interpretations from users, especially since their goals and needs are dynamic. Once we recognize that users will make their own interpretations (i.e., appropriations) of XAI statements, then we center our design around user agency by shiting our intention from *control* of users' understanding, toward providing *resources* through which users can construct their own understanding.

To aid our *appropriation-aware* design of XAI systems and expand the design space, we can build on work in HCI (e.g., Reflective Design [126]) and software engineering [124]. In particular, we can draw on Dix's influential design guidelines for appropriation [36]. Particularly in the context of explanations, certain guidelines stand out– guidelines like *exposing intentions* of what the system's goals and aims can help the appropriator be explicitly aware if they are using the explanation in unintended ways. By potentially bringing unconscious aspects of experience to conscious awareness, even if users appropriate explanations in novel ways, it is done in an informed manner. Most importantly, we should *support, not control* our end-users—we need to strike the right balance between automation and control. The goal is not to overly control the user's interpretation but to support the flexibility in ways that promote one's agency in the system. For example, the AI group's explanatory intent of diagnosis drove them to ascribe too much value to numbers. Here, if we could provide visibility around what the numbers can and cannot do, including exposing the intentions behind them, we might calibrate their intent in ways that mitigate over-trust. Similarly, if we were aware of the affirmatory intent of the non-AI group, we could have designed AD differently. These insights bring us to the most important part of designing for appropriation, which is about learning *from* appropriation. Feedback loops from users can not only improve our XAI systems, but they can also highlight our intellectual blind spots and calibrate the balance between support and control in user interpretations.

6.3 Broader lessons: Explainable and Responsible AI

Our work has broader implications beyond the immediate design of XAI systems, especially around the discourse of responsible and explainable AI. Below we share three main takeaways—how the perceptual differences have knock-on effects that require a sociotechnical approach, how our insights can help re-imagine AI education, and how our insights can mitigate potential harmful manipulation with explanations.

First, our work illustrates that *merely producing well-designed explanations is not enough to guarantee people would perceive them as designed*— the same explanation can lead to divergent experiences between groups. Perception drives action, and misplaced perceptions (e.g., over-reliance on numbers) can lead to negative effects such as inappropriate user trust in AI systems. Such misalignments can add to disparity and inequities in XAI design.

The solution to the potential disparity should neither be on expecting every user to perfectly understand AI systems nor does it lie *solely* in "opening" the opaque-box of AI in more creative ways. Our problems are sociotechnical: AI systems do not exist in vacuums. They are situated in environments and are inherently sociotechnical in nature. Thus, we can take advantage of the organizational affordances to address issues around explainability. For example, recent work on Social Transparency exhibits how incorporating things outside the model, such as the socio-organizational context, into the decision-making process can augment the holistic explainability of AI systems [40]. More importantly, we can introduce reliable and trustworthy AI behavior through value-sensitive practices and organizational infrastructure (e.g., audits, independent oversight, safety practices, etc.). In short, even the "perfect" explanation cannot solve all the problems around explainability; we need to take a sociotechnically informed approach.

Second, our insights call attention to wider challenges with *academic practices of AI learning*. AI students exhibited unwarranted faith in and preference for numbers even when they were not readily comprehensible. In light of the importance of quantitative practices in AI, this bias is not surprising. One effective way to address this would be to critically reflect on the way we educate students in AI. How do we ensure that students have a more critical eye towards the working and outputs of AI systems? This is where we see the impact of our focus on AI students (vs. fully formed AI practitioners)— by investigating how AI background can lead to the formation of certain heuristics, we have crucial insights on how we might address the issues from an academic training perspective. For instance, the need to introduce courses like critical data studies and human-centered data science that can provide the much-needed reflective lenses to students to understand not only their own cognitive biases but also the importance of thinking about the user during system design and development. By addressing the issues during the formation of their Data Vision [115], and Professional Vision [52] more generally, we can revolutionize how we train the next generations of AI creators and mitigate the creator-consumer gap. If we are asking today's AI students to become future creators for consumers that think drastically differently from themselves, we need to empower our students with the lenses such that the gap between them and the stakeholders they would serve is less than what it is today.

Last, our work can inform how we might preemptively *mitigate harmful manipulation* of the perceptual differences of explanations. We have highlighted the heuristics behind the misplaced trust in numbers from the AI group (Section 6.1). As these students go on to become designers, engineers, and managers of AI systems, their faith in numbers may deepen, leading them to potentially over-rely on numbers—even professional data scientists are often overconfident about their metrics [70]. The AI group not only associated numbers with higher-order intelligence but also exhibited confidence to act on them despite lack of comprehension. What might have happened if the NR robot made an error or due to faulty logic produced erroneous Q-values? Would AI group participants still exhibit a strong faith in numbers and their potential actionability, ascribing higher intelligence to the agent? The AI group participants, students taking an AI class, are still in the process of gaining "data vision" [115]—an important attribute of which is the systemic exposure to quantified reasoning. While some can interpret numbers, we *cannot—and should not—expect* AI students (or even experts) to *always* work out the differences between correct and faulty Q-values; indeed, it is *impossible to differentiate* between them without investigating traces of the agent during learning. Moreover, we published the Q-values in good faith; however, we can imagine bad-faith actors manipulating numbers for nefarious means, potentially manifesting *dark patterns* [53] in XAI design. Imagine an XAI system that explains in numbers (ones that are manipulated to induce trust); given the faith in numbers from the AI group, how might that impact the trust in the system? Furthermore, as government regulations require explainability—such as the European Union's GDPR—we observe that it may be possible to design explanations that are not only devoid of justificatory power but are still perceived by certain groups as having value (like how the non-AI group found explanatory value in AD's justification-less statements). In Section 6.2, we share mitigation strategies that could potentially counteract these negative effects.

We view our findings as a *call-to-action* to proactively investigate how we might prevent harmful manipulation of explanations by abusing certain heuristics common in people with and without AI backgrounds. More broadly, the implication of perceptual and interpretive differences has the potential to have further knock-on effects on pertinent aspects such as fairness, trust, and accountability of AI systems.

7 LIMITATIONS & FUTURE WORK

This research is a first step towards understanding how AI background—an important user characteristic—impacts user perceptions of AI explanations. As highlighted in Section 6.3, a focus on AI students helped to understand how enculturation into AI alters people's perceptions of AI explanations and makes visible the presence of other kinds of – often similar – interpretations within a heterogeneous set of non-AI group participants. In this initial effort, we scoped our study in the context of three types of AI explanations– while they are informative, they are not exhaustive of the different types of explanations. While we were able to explore the *what* and *how* of the differences between the two groups through our mixed-methods study, we postulated on the *why* using conceptual lenses of heuristics and appropriation. Future work could empirically explore these mechanisms behind the group differences and expand on our findings. We also did not explore how users with two or more different user characteristics (e.g., comparison with multiple facets of one's background) or more homogeneous and striated AI backgrounds (e.g., years of AI programming experience) perceive explanations in different ways, which restricts our ability to make claims about these areas. With these insights in mind, our findings are circumscribed by the scope of the study and should be interpreted as such. In the future, we want to move to a multi-stakeholder scenario and explore how AI background differences between homogeneous stakeholder groups (e.g., data scientists vs. business analysts) impact interpretations of AI explanations.

8 CONCLUSIONS

In this paper, we focus on the *who* of XAI by investigating how two different groups of *whos*—people with and without a background in AI—perceive different types of AI explanations. Through a mixed-methods user study, we demonstrate how one's AI background, or lack thereof, shapes their interpretations and perceptions. Our mixed-methods analysis provides different levels of insights. *What* people prefer is relatively clear—they prefer natural language-based justificatory rationales. While the *what* is somewhat straightforward, the *why* and *how* behind their preferences is nuanced. Different perceptions (e.g., unanticipated explanatory value) sometimes arise from different forms of appropriation (e.g., diagnosis vs. affirmatory intent). The same perception (unwarranted faith in numbers) sometimes arises from different types of heuristics (associating numerical vs. incomprehensible reasoning with intelligence) And, finally, both groups were very similar in their desire to engage with natural-language based explanations.

Explainability of AI systems is crucial to instil appropriate user trust and facilitate recourse. Disparities in AI backgrounds have the potential to exacerbate the challenges arising from the differences between how designers imagine users will appropriate explanations vs. how users actually interpret and use them. We provided concrete design implications to mitigate the risk of over-reliance of numbers and broader lessons about the need for re-imagining AI education. By focusing on the who and not just the what of XAI, our work takes a formative step in advancing a pluralistic human-centered XAI discourse to help bridge the creator-consumer gap.

ACKNOWLEDGMENTS

With our deepest gratitude, we thank our participants for generously investing their time in this project. We are grateful to past and present members of the Human-Centered AI Lab at Georgia Tech, especially Pradyumna Tambwekar, Shukan Shah, Chenghann Gan, Jiahong Sun, and Sruthi Sudhakar, whose continued input refined the conceptualizations presented here. We are indebted to Simone Stumpf, Ishtiaque Ahmed, Devleena Das, Intekhab Hossain, Brent Harrison, and Malte Jung for their generous feedback and time that helped scope the project appropriately. Special thanks to Rachel Urban for generously providing proofreading feedback. This project was partially supported by the National Science Foundation under Grant No. 1928586.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Alan Agresti. 2003. *Categorical data analysis*. Vol. 482. John Wiley & Sons.

[4] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[5] Junia Anacleto and Sidney Fels. 2013. Adoption and appropriation: A design process from HCI research at a Brazilian neurological hospital. In *IFIP Conference on Human-Computer Interaction*. Springer, 356–363.

[6] Jacob Andreas, Anca Dragan, and Dan Klein. 2017. Translating neuralese. *arXiv preprint arXiv:1704.06960* (2017).

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[8] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[9] David Bawden et al. 2008. Origins and concepts of digital literacy. *Digital literacies: Concepts, policies and practices* 30, 2008 (2008), 17–32.

[10] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior research methods* 43, 3 (2011), 800.

[11] Frank R Bentley, Nediyana Daskalova, and Brooke White. 2017. Comparing the reliability of Amazon Mechanical Turk and Survey Monkey to traditional market research surveys. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1092–1099.

[12] Richard A Berk and Justin Bleich. 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol'y* 12 (2013), 513.

[13] Timothy W Bickmore and Rosalind W Picard. 2004. Towards caring machines. In *CHI'04 extended abstracts on Human factors in computing systems*. 1489–1492.

[14] Wiebe E Bijker. 1997. *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change.* MIT press.

[15] 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*. ACM, 377.

[16] John Bowers. 2012. The logic of annotated portfolios: communicating the value of'research through design'. In *Proceedings of the Designing Interactive Systems Conference*. 68–77.

[17] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[18] Angelo Cafaro, Hannes Högni Vilhjálmsson, Timothy Bickmore, Dirk Heylen, Kamilla Rún Jóhannsdóttir, and Gunnar Steinn Valgarðsson. 2012. First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *International conference on intelligent virtual agents*. Springer, 67–80.

[19] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.

[20] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the" Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[21] Kathy Charmaz. 2014. *Constructing Grounded Theory (Introducing Qualitative Methods series) 2nd Edition.* Sage, London.

[22] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2016. Interpretable deep models for ICU outcome prediction. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 371.

[23] Jim Q Chen and Sang M Lee. 2003. An exploratory cognitive DSS for strategic decision making. *Decision support systems* 36, 2 (2003), 147–160.

[24] 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*. 1–12.

[25] EunJeong Cheon and Norman Makoto Su. 2017. Configuring the User: " Robots have Needs Too". In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 191–206.

[26] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[27] Mark Core, H. Chad Lane, Michael van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. 2006. Building Explainable Artificial Intelligence Systems. In *Proceedings of the 18th Innovative Applications of Artificial Intelligence Conference*. Boston, MA.

[28] Pat Croskerry. 2003. Cognitive forcing strategies in clinical decisionmaking. *Annals of emergency medicine* 41, 1 (2003), 110–120.

[29] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 351–360.

[30] Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 510–518.

[31] Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. 2018. Boolean Decision Rules via Column Generation. *Advances in Neural Information Processing Systems* 31 (2018), 4655–4665.

[32] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. 1989. User acceptance of computer technology: a comparison of two theoretical models. *Management science* 35, 8 (1989), 982–1003.

[33] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 251–258.

[34] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. 1591–1602.

[35] Andrea A DiSessa. 2001. *Changing minds: Computers, learning, and literacy*. Mit Press.

[36] Alan Dix. 2007. Designing for appropriation. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21*. 1–4.

[37] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE 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.

[38] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[39] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence, Ethics, and Society*.

[40] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[41] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. *arXiv preprint arXiv:2002.01092* (2020).

[42] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *Proceedings of the International Conference on Intelligence User Interfaces*.

[43] Malin Eiband, Daniel Buschek, and Heinrich Hussmann. 2021. How to support users in understanding intelligent systems? Structuring the discussion. In *26th International Conference on Intelligent User Interfaces*. 120–132.

[44] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312787

[45] Kimmo Eriksson. 2012. The nonsense math effect. *Judgment and Decision Making* 7, 6 (2012), 746–749. http://decisionsciencenews.com/sjdm/journal.sjdm.org/12/12810/jdm12810.pdf

[46] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/2702123.2702556

[47] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (Sep 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[48] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL) Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[49] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069* (2018).

[50] Barney Glaser and Anselm Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transactions, Chicago.

[51] Joseph K Goodman and Gabriele Paolacci. 2017. Crowdsourcing consumer research. *Journal of Consumer Research* 44, 1 (2017), 196–210.

[52] Charles Goodwin. 1994. Professional Vision. *American Anthropologist* 96, 3 (1994), 606–633.

[53] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[54] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[55] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[56] Karen Hao. 2019. AI is sending people to jail – and getting it wrong. *MIT Technology Review* (21 January 2019). Retrieved 26-August-2019 from https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/

[57] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48, 1 (2016), 400–407.

[58] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[59] Myles Hollander, Douglas A Wolfe, and Eric Chicken. 2013. *Nonparametric statistical methods*. Vol. 751. John Wiley & Sons.

[60] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).

[61] Chien-Ming Huang, Takamasa Iio, Satoru Satake, and Takayuki Kanda. 2014. Modeling and Controlling Friendliness for An Interactive Museum Robot.. In *Robotics: Science and Systems*. 12–16.

[62] Lilly Irani and M Six Silberman. 2014. From critical design to critical infrastructure: lessons from turkopticon. *interactions* 21, 4 (2014), 32–35.

[63] James Jaccard. 2001. *Interaction effects in logistic regression*. Vol. 135. SAGE Publications, Incorporated.

[64] Alejandro Jaimes, Daniel Gatica-Perez, Nicu Sebe, and Thomas S Huang. 2007. Guest Editors' Introduction: Human-Centered Computing–Toward a Human Revolution. *Computer* 40, 5 (2007), 30–34.

[65] Jakob D Jensen, Andy J King, LaShara A Davis, and Lisa M Guntzviller. 2010. Utilization of internet technology by low-income adults: the role of health literacy, health numeracy, and computer assistance. *Journal of aging and health* 22, 6 (2010), 804–826.

[66] Weina Jin, Sheelagh Carpendale, Ghassan Hamarneh, and Diane Gromala. 2019. Bridging ai developers and end users: an end-user-centred explainable ai taxonomy and visual vocabularies. *Proceedings of the IEEE Visualization, Vancouver, BC, Canada* (2019), 20–25.

[67] W Lewis Johnson. 1994. Agents that Learn to Explain Themselves.. In *AAAI*. 1257–1263.

[68] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[69] Gajendra Jung Katuwal and Robert Chen. 2016. Machine learning model interpretability for precision medicine. *arXiv preprint arXiv:1610.09045* (2016).

[70] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[71] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[72] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293. https://doi.org/10.1093/qje/qjx032

[73] Rob Kling and Susan Leigh Star. 1998. Human centered systems in the perspective of organizational and social informatics. *ACM SIGCAS Computers and Society* 28, 1 (1998), 22–29.

[74] Tomoko Koda and Pattie Maes. 1996. Agents with faces: The effect of personification. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*. IEEE, 189–194.

[75] Alina Krischkowsky. 2018. *Technology in Uses: Negotiating the in-between of designer | user and design | use.* Ph.D. Dissertation. University of Salzburg, Salzburg, Austria.

[76] 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, 3–10.

[77] Minae Kwon, Sandy H Huang, and Anca D Dragan. 2018. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 87–95.

[78] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.

[79] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010. Receptionist or information kiosk: how do people talk with a robot?. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 31–40.

[80] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 203–210.

[81] Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What can you do? Studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 acm conference on designing interactive systems*. 264–275.

[82] 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*. 1–15.

[83] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. 2018. All work and no play? Conversations with a Question-and-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[84] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[85] Tung-Ching Lin, Sheng Wu, Jack Shih-Chieh Hsu, and Yi-Ching Chou. 2012. The integration of value-based adoption and expectation–confirmation models: An example of IPTV continuance intention. *Decision Support Systems* 54, 1 (2012), 63–75.

[86] Peter Lipton. 2001. What good is an explanation? In *Explanation*. Springer, 43–59.

[87] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[88] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods* 49, 2 (2017), 433–442.

[89] Tyler J. Loftus, Patrick J. Tighe, Amanda C. Filiberto, Philip A. Efron, Scott C. Brakenridge, Alicia M. Mohr, Parisa Rashidi, Jr Upchurch, Gilbert R., and Azra Bihorac. 2020. Artificial Intelligence and Surgical Decision-making. *JAMA Surgery* 155, 2 (02 2020), 148–158. https://doi.org/10.1001/jamasurg.2019.4917 arXiv:https://jamanetwork.com/journals/jamasurgery/articlepdf/2756311/jamasurgery_loftus_2019_rv_190007.pdf

[90] Tania Lombrozo. 2011. The instrumental value of explanations. *Philosophy Compass* 6, 8 (2011), 539–551.

[91] Tania Lombrozo. 2012. Explanation and abductive inference. *Oxford handbook of thinking and reasoning* (2012), 260–276.

[92] Tania Lombrozo. 2016. Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences* 20, 10 (2016), 748 – 759. https://doi.org/10.1016/j.tics.2016.08.001

[93] Tania Lombrozo and Nicholas Z Gwynne. 2014. Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience* 8 (2014), 700.

[94] Tania Lombrozo and Daniel Wilkenfeld. 2019. Mechanistic versus functional understanding. *Varieties of understanding: New perspectives from philosophy, psychology, and theology* (2019), 209.

[95] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. 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, 1–16. https://doi.org/10.1145/3313831.3376727

[96] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.

[97] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[98] Donald MacKenzie. 2018. Material Signals: A Historical Sociology of High-Frequency Trading. *Amer. J. Sociology* 123, 6 (2018), 1635–1683. https://doi.org/10.1086/697318

[99] Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* 42, 2 (1980), 109–127.

[100] Nora McDonald and Shimei Pan. 2020. Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–19.

[101] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 115 (Oct. 2020), 25 pages. https://doi.org/10.1145/3415186

[102] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[103] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[104] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. arXiv:1811.11839 [cs.HC]

[105] Johanna D Moore and William R Swartout. 1988. *Explanation in expert systemss: A survey*. Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

[106] Michael Muller, Melanie Feinberg, Timothy George, Steven J. Jackson, Bonnie E. John, Mary Beth Kery, and Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299018

[107] Michael Muller, Katja Neureiter, Nervo Verdezoto, Alina Krischkowsky, Anna Maria Al Zubaidi-Polli, and Manfred Tscheligi. 2016. Collaborative appropriation: How couples, teams, groups and communities adapt and adopt technologies. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 473–480.

[108] John Murawski. 2019. Mortgage Providers Look to AI to Process Home Loans Faster. *Wall Street Journal* (18 March 2019). Retrieved 16-September-2020 from https://www.wsj.com/articles/mortgage-providers-look-to-ai-to-process-home-loans-faster-11552899212

[109] Chelsea M Myers, Anushay Furqan, and Jichen Zhu. 2019. The impact of user characteristics and preferences on performance with an unfamiliar voice user interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[110] Clifford Nass, BJ Fogg, and Youngme Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6 (1996), 669–678.

[111] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[112] Clifford Nass, Jonathan Steuer, Lisa Henriksen, and D Christopher Dryer. 1994. Machines, social attributions, and ethopoeia: Performance assessments of computers subsequent to" self-" or" other-" evaluations. *International Journal of Human-Computer Studies* 40, 3 (1994), 543–559.

[113] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[114] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, 39–48.

[115] Samir Passi and Steven J. Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM Press, 2436–2447.

[116] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (nov 2018), 1–28.

[117] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (2020), 2053951720939605. https://doi.org/10.1177/2053951720939605

[118] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[119] Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/2702123.2702174

[120] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[121] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 673–705.

[122] Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review* 2, 1 (31 3 2020). https://doi.org/10.1162/99608f92.6ed64b30 https://hdsr.mitpress.mit.edu/pub/7z10o269.

[123] Antti Salovaara. 2008. Inventing new uses for tools: a cognitive foundation for studies on appropriation. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments* (2008).

[124] Antti Salovaara and Sakari Tamminen. 2009. Acceptance or appropriation? A design-oriented critique of technology acceptance models. In *Future interaction design II*. Springer, 157–173.

[125] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[126] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58.

[127] Steven J Sherman and Eric Corty. 1984. Cognitive heuristics. (1984).

[128] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.

[129] Edward Shortliffe. 2012. *Computer-based medical consultations: MYCIN*. Vol. 2. Elsevier.

[130] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[131] Kwonsang Sohn and Ohbyung Kwon. 2020. Technology acceptance theories and factors influencing artificial intelligence-based intelligent products. *Telematics and Informatics* 47 (2020), 101324.

[132] Jon Sprouse. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods* 43, 1 (2011), 155–167.

[133] Sonja Stange and Stefan Kopp. 2020. Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 619–627.

[134] Anselm Strauss and Juliet M. Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Techniques and Procedures*. Sage, New York.

[135] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325.

[136] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 303–310.

[137] Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry* 1, 4 (1990), 285–293.

[138] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).

[139] Manfred Tscheligi, Alina Krischkowsky, Katja Neureiter, Kori Inkpen, Michael Muller, and Gunnar Stevens. 2014. Potentials of the" Unexpected" Technology Appropriation Practices and Communication Needs. In *Proceedings of the 18th international conference on supporting group work*. 313–316.

[140] UCLA. 2014. Ordinal Logistic Regression | R Data Analysis Examples. https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

[141] Michael van Lent, , Paul Carpenter, Ryan McAlinden, and Paul Brobst. 2005. Increasing Replayability with Deliberative and Reactive Planning. In *1st Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*. Maria Del Rey, California, 135–140.

[142] Michael Van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 900–907.

[143] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530.

[144] W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S* (fourth ed.). Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4 ISBN 0-387-95457-0.

[145] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

[146] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[147] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[148] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.

[149] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.

[150] Dennis Wei, Sanjeeb Dash, Tian Gao, and Oktay Gunluk. 2019. Generalized linear rule models. In *International Conference on Machine Learning*. PMLR, 6687–6696.

[151] Karl E. Weick. 1995. *Sensemaking in Organizations*. SAGE Publications, Thousand Oaks, California.

[152] Daniel A Wilkenfeld and Tania Lombrozo. 2015. Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education* 24, 9-10 (2015), 1059–1077.

[153] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[154] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[155] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.

[156] Achim Zeileis, Mark A Wiel, Kurt Hornik, and Torsten Hothorn. 2008. Implementing a class of permutation tests: the coin package. *Journal of statistical software* 28, 8 (2008), 1–23.

A APPENDIX

A.1 Best practices for data integrity and participant engagement

Data quality and engagement with user study participants are integral to research. Below we share how previous guidelines for fair and equitable work treatment for MTurk workers [62] helped us decide the payment structure, tips on engagement with participants, design motivations behind the task environment, and how we deployed and reviewed the task catering for a global audience. While these insights are transferable to other contexts, most of these practices are geared towards MTurk participants given one has less control on the platform.

1. *Payment*: our study was not a micro-task that is traditionally deployed in MTurk. As a result, we calibrated our payment to reflect the task duration and expected effort. We tried our best to structure our study based on previous guidelines for fair and equitable work treatment for MTurk workers [62]. We strived to pay equal to or more than a minimum wage [at the time of deployment, the local minimum wage was \$8.5/hour]. We paid \$10 for a task budgeted for 45 minutes, making the hourly pay \$13.3. However, almost all participants took 30 mins to complete the task on average, making the effective hourly rate around \$20/hour.

   As a policy, we disbursed payments within 48 hours of task completion. This robust turnaround time helped our reputation on Turkopticon, a forum for MTurk workers to engage in peer-to-peer assistance on job information and hold employers accountable for fair treatment [62]. Moreover, one researcher regularly engaged with workers on Turkopticon, answering questions and returning compliments. This engagement built rapport throughout the study. As a platform, TurkPrime allows internal messages between MTurk workers and employers by using a proxy userID and protecting the privacy of the worker. This feature allowed participants to communicate if they had internet issues or were running out of time. In such cases, we sent them a one-time link for completion. The same applied for participants in the AI background group who were not bound by AMT rules.

   Every payment of a HIT had a thank you message attached to it. Every rejection had custom justifications backed by evidence. The research team created message templates based on major issues, which allowed for a quick turnaround time even with a custom message. For participants who failed to do the task despite best efforts, we paid them for their time even if we could not use their data. This equitable policy also made our HITs one of the most sought after in the marketplace.
2. *Task environment and setup:*
   (a) *Task orientation*: Pilot testing showed that participants preferred a multi-modal (e.g., video) orientation compared to a textual description of it. Therefore, we provided both modalities.
   (b) *Task engagement*: Participant had to successfully pass attention checks and/or explicitly acknowledge they understood the instructions in a given module. On the backend, we had timers to evaluate if a participant spent a reasonable amount of time on a particular section. For instance, if the video was 2 minutes long and a participant clicked through that section in 30 seconds, it triggered a review of their responses. These steps augment the quality of the experimental data. Moreover, for tasks that had to be rejected, these metrics also served as justificatory evidence.
   (c) *Design of the robots*: To mitigate effects of preconceived notions, we did not use any descriptive names for the robots; instead, we introduced the robots as “Robot A” [= the Rationale-Generation robot], “Robot B” [= the Action-Declaring robot], and “Robot C” [= the Numerical robot] (details on each of their attributes are below/above). To reduce any preferential treatment of robots based on their appearances, we need to standardize their appearances without sacrificing their distinctness. That is, the robots needed to look similar,

but not to the point of indistinguishability. This insight came from pilot testing which indicated that if we made the robots identical in appearance but different in color, the cognitive load for recall was too high. Therefore, we iterated and struck a balance where all robots had wheels with a "car-like" structure (see Fig. 1) while they differed in color and shape.

(3) *Deployment and Review*: Across both groups, we manually reviewed every response in the survey, especially the qualitative justifications provided by the participants. We deployed 10-15 tasks per day to allow for manual reviews. To facilitate outreach of our task to all time zones, using an automated scheduling system, we released 3 tasks every 3 hours over a 24-hour cycle. This improved the potential for global participation in our task.
    Spamming is a serious issue when it comes to survey data. Here, the qualitative responses served a secondary screening purpose. Participants with good-faith efforts always had reasonable qualitative justifications. Those who had spamming intentions shared non-sensical and even comical qualitative responses; e.g., Movie titles and plots, snippets from Wikipedia, etc.

While all of these steps required considerable time, and effort, they paid off in the high data quality we received for a task lasting 30 minutes on average.

A.2 Participant Screening

The following aspects of the selection criteria facilitated formation of the two user groups:

The AI knowledge questionnaire: We developed the knowledge questionnaire iteratively using a participatory process involving Teaching Assistants for the class along with Graduate students familiar with the area. There are five (5) multiple-choice questions in total equalling 5 points. The first two questions are programming questions: the first being a question asking for the output of a simple print-statement, the second being one that asks for the output of a for-loop. The remaining three covered concepts in AI such as Markov Decision Processes, Reinforcement Learning, and Unsupervised Learning. By the time of deployment, the AI students had already gone through lectures covering the AI topics in the questionnaire. All the questions are inspired by or directly taken from past exam questions on various topics. For further details, please refer to section A.2.1 for the AI knowledge questionnaire.

To calibrate the relative difficulty of the knowledge test, we used a collaborative and iterative process until consensus between the researchers and the teaching staff was reached. We expected that most students with satisfactory prerequisites (that contain fundamentals of programming) and current knowledge from the class should at least get 4 out of the 5 correct. This calibration appears to have been a reasonable one since all students naturally passed these thresholds. On the other hand, in order to be assigned to the non-AI background group, participants had to score less than or equal to 1 (out of 5). We expected that some participants, without any AI background, might be able to guess the output of the “print” statement question. However, it was unlikely that someone without basic programming understanding would be able to answer the “for-loop” output question. Therefore, if someone correctly answers more than one, their AI knowledge background is not the type we needed for members of the non-AI background group.

AI background measurement: The knowledge test was followed by two 5-point Likert-scale questions measuring the AI background for computer programming and AI concepts. The range of self-reported knowledge goes from "No knowledge"[= 1] to "A lot of knowledge" [= 5]. Each level of knowledge has a sentence clarifying the meaning behind the label. For illustrative purposes, here is an example from the AI scale: "No knowledge: I might be *aware* of AI, but have *no knowledge* about it." Both scales had similar construction and wording. For further details, please refer to section A.2.1 for the AI background knowledge Likert-scales.

AI class: Finally, participants answer if they have ever taken any classes on Artificial Intelligence.

A.2.1 ***Screening questionnaire***.

Here we share the survey instruments used to screen participants. *Knowledge test questionnaire*

(1) What would be the output of the following python program?

```
name = "Peter"
print("Hello " + name)
```

(a) Peter
(b) Hello Peter
(c) Hello + Peter
(d) "Hello" + name

(2) What would be the output of the following python program?

```
numbers = [2, 4]
for i in range(len(numbers)):
    print(numbers[i] + i)
```

(a) 2
5

(b) 2
5
8

(c) 2
4

(d) 2
4
10

(3) Which of the following is an unsupervised learning task?
  (a) Distinguishing pictures containing cats from pictures not containing cats
  (b) Flagging text messages as appropriate or inappropriate
  (c) Divide data points into different clusters without any labels available
  (d) Predict the value of a house after training on a dataset with house features and values

(4) What is the general goal of reinforcement learning?
  (a) Maximize potential or expected punishment
  (b) Maximize potential or expected reward
  (c) Get to the goal as soon as possible
  (d) Avoid the most obstacles in any given state

(5) In MDPs, the Markov assumption is that:
  (a) The current state is independent of all other states
  (b) The current state depends only on the history of previous states and actions
  (c) The current state depends on the full sequence of states and actions (past and future)
  (d) The current state only depends on the immediate previous state and action

Computer Programming Background Knowledge.

When it comes to computer programming or coding, I believe I have

(1) No knowledge: I might be *aware* of computer programs, but have *never* coded before
(2) A little knowledge: I know *basic* concepts in programming, but have *never applied* it
(3) Some knowledge: I have *applied* programming concepts by coding at least *once* before
(4) Moderate knowledge: I *apply* programming concepts *somewhat frequently* for my work, class, or leisure
(5) A lot of knowledge: I *apply* programming concepts *very frequently* or create cutting edge software

AI Background Knowledge.

When it comes to Artificial Intelligence (AI), I believe I have

(1) No knowledge: I might be *aware* of AI, but have *no knowledge* about it
(2) A little knowledge: I know *basic* concepts in AI, but have *never applied* it
(3) Some knowledge: I have *applied* AI concepts by coding at least *once* before
(4) Moderate knowledge: I *apply* AI concepts *somewhat frequently* for my work, class, or leisure
(5) A lot of knowledge: I *apply* AI concepts *very frequently* or create cutting edge software

AI class.
Have you ever taken or are currently taking any classes on Artificial Intelligence?

- Yes
- No

A.3 OLR Summary Tables

Table 3. Summary of OLR with Ranking as Response and Robot Type as Predictor

	Value	Std. Error	*t*- value	*p*- value	Odds Ratio
AD Robot	-1.711	0.104	-16.473	**< 0.001**	**0.181**
NR_Robot	-2.691	0.115	-23.393	**< 0.001**	**0.068**
1|2	-2.366	0.093	-25.567	< 0.001	0.094
2|3	-0.598	0.077	-7.799	< 0.001	0.5501

Note: RG_Robot is the reference level.

Table 4. OLR Summary with Robot Type

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.7105	0.1038	16.4728	0	5.5319
Numerical-Reasoning Robot	-0.9807	0.1001	-9.7980	0	0.3751
1|2	-0.6550	0.0695	-9.4282	0	0.5195
2|3	1.1129	0.0734	15.1592	0	3.0433

Note: Action-Declaring Robot is the reference level.

Table 5. OLR Summary - Ref. Levels: Rationale-Generation Robot and AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-2.0246	0.1302	-15.5462	0.0000	0.1320
Numerical-Reasoning Robot	-2.5072	0.1391	-18.0221	0.0000	0.0815
Non-AI Group	-0.1563	0.1582	-0.9879	0.3232	0.8553
Action-Declaring Robot:Non-AI Group	0.8422	0.2093	4.0243	0.0001	2.3214
Numerical-Reasoning Robot:Non-AI Group	-0.6088	0.2264	-2.6892	0.0072	0.5440
1|2	-2.4515	0.1104	-22.2009	0.0000	0.0862
2|3	-0.6507	0.0965	-6.7413	0.0000	0.5217

Table 6. OLR Summary - Ref. Levels: Rationale-Generation Robot and Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-1.1824	0.1674	-7.0614	0.0000	0.3065
Numerical-Reasoning Robot	-3.1160	0.1893	-16.4636	0.0000	0.0443
AI Group	0.1564	0.1582	0.9881	0.3231	1.1692
Action-Declaring Robot:AI Group	-0.8422	0.2093	-4.0245	0.0001	0.4308
Numerical-Reasoning Robot:AI Group	0.6087	0.2264	2.6889	0.0072	1.8381
1|2	-2.2952	0.1362	-16.8540	0.0000	0.1007
2|3	-0.4944	0.1258	-3.9305	0.0001	0.6099

Table 7. OLR Summary - Ref. Levels: Action-Declaring Robot and AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	2.0246	0.1302	15.5462	0e+00	7.5732
Numerical-Reasoning Robot	-0.4826	0.1224	-3.9421	1e-04	0.6172
Non-AI Group	0.6859	0.1368	5.0127	0e+00	1.9855
Rationale-Generation Robot:Non-AI Group	-0.8422	0.2093	-4.0243	1e-04	0.4308
Numerical-Reasoning Robot:Non-AI Group	-1.4510	0.2125	-6.8278	0e+00	0.2343
1|2	-0.4269	0.0836	-5.1091	0e+00	0.6526
2|3	1.3739	0.0904	15.2045	0e+00	3.9507

Table 8. OLR Summary - Ref. Levels: Action-Declaring Robot and Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.1824	0.1674	7.0613	0e+00	3.2622
Numerical-Reasoning Robot	-1.9336	0.1750	-11.0468	0e+00	0.1446
AI Group	-0.6859	0.1368	-5.0127	0e+00	0.5037
Rationale-Generation Robot:AI Group	0.8422	0.2093	4.0245	1e-04	2.3215
Numerical-Reasoning Robot:AI Group	1.4510	0.2125	6.8278	0e+00	4.2672
1|2	-1.1127	0.1147	-9.6973	0e+00	0.3287
2|3	0.6880	0.1123	6.1249	0e+00	1.9898

Table 9. OLR Summary - Ref. Levels: Numerical-Reasoning Robot and AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	2.5072	0.1391	18.0222	0.0000	12.2708
Action-Declaring Robot	0.4826	0.1224	3.9422	0.0001	1.6203
Non-AI Group	-0.7651	0.1620	-4.7240	0.0000	0.4653
Rationale-Generation Robot:Non-AI Group	0.6088	0.2264	2.6891	0.0072	1.8382
Action-Declaring Robot:Non-AI Group	1.4510	0.2125	6.8278	0.0000	4.2672
1|2	0.0557	0.0923	0.6037	0.5460	1.0573
2|3	1.8565	0.1033	17.9798	0.0000	6.4012

Table 10. OLR Summary - Ref. Levels: Numerical-Reasoning Robot and Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	3.1160	0.1893	16.4637	0.0000	22.5559
Action-Declaring Robot	1.9336	0.1750	11.0468	0.0000	6.9141
AI Group	0.7651	0.1620	4.7240	0.0000	2.1492
Rationale-Generation Robot:AI Group	-0.6088	0.2264	-2.6891	0.0072	0.5440
Action-Declaring Robot:AI Group	-1.4510	0.2125	-6.8279	0.0000	0.2343
1|2	0.8208	0.1336	6.1447	0.0000	2.2724
2|3	2.6216	0.1444	18.1523	0.0000	13.7575

Table 11. OLR Summary - Confidence - Ref. Levels: Type Rationale-Generation Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-1.3753	0.2740	-5.0185	0.0000	0.2528
Numerical-Reasoning Robot	-1.2607	0.2815	-4.4783	0.0000	0.2835
Non-AI Group	-0.0290	0.3297	-0.0879	0.9299	0.9714
Action-Declaring Robot:Non-AI Group	0.7232	0.4545	1.5913	0.1115	2.0610
Numerical-Reasoning Robot:Non-AI Group	-0.6906	0.4668	-1.4795	0.1390	0.5013
1|2	-1.6705	0.2164	-7.7194	0.0000	0.1882
2|3	-0.1251	0.1996	-0.6268	0.5308	0.8824

Table 12. OLR Summary - Confidence - Ref. Levels: Type Rationale-Generation Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-0.6521	0.3654	-1.7845	0.0743	0.5210
Numerical-Reasoning Robot	-1.9513	0.3800	-5.1343	0.0000	0.1421
AI Group	0.0290	0.3297	0.0879	0.9299	1.0294
Action-Declaring Robot:AI Group	-0.7232	0.4545	-1.5913	0.1115	0.4852
Numerical-Reasoning Robot:AI Group	0.6906	0.4668	1.4795	0.1390	1.9949
1|2	-1.6415	0.2777	-5.9112	0.0000	0.1937
2|3	-0.0961	0.2651	-0.3626	0.7169	0.9083

Table 13. OLR Summary - Confidence - Ref. Levels: Type Action-Declaring Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.3753	0.2740	5.0185	0.0000	3.9561
Numerical-Reasoning Robot	0.1146	0.2678	0.4279	0.6687	1.1214
Non-AI Group	0.6942	0.3127	2.2203	0.0264	2.0021
Rationale-Generation Robot:Non-AI Group	-0.7232	0.4545	-1.5913	0.1115	0.4852
Numerical-Reasoning Robot:Non-AI Group	-1.4138	0.4561	-3.0995	0.0019	0.2432
1|2	-0.2952	0.1873	-1.5759	0.1150	0.7444
2|3	1.2501	0.1979	6.3175	0.0000	3.4908

Table 14. OLR Summary - Confidence - Ref. Levels: Type Action-Declaring Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	0.6521	0.3654	1.7845	0.0743	1.9196
Numerical-Reasoning Robot	-1.2992	0.3688	-3.5230	0.0004	0.2727
AI Group	-0.6942	0.3127	-2.2203	0.0264	0.4995
Rationale-Generation Robot:AI Group	0.7232	0.4545	1.5913	0.1115	2.0609
Numerical-Reasoning Robot:AI Group	1.4138	0.4561	3.0995	0.0019	4.1115
1|2	-0.9894	0.2603	-3.8015	0.0001	0.3718
2|3	0.5560	0.2566	2.1668	0.0303	1.7436

Table 15. OLR Summary - Confidence - Ref. Levels: Type Numerical-Reasoning Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.2607	0.2815	4.4783	0.0000	3.5278
Action-Declaring Robot	-0.1146	0.2678	-0.4279	0.6687	0.8918
Non-AI Group	-0.7196	0.3304	-2.1778	0.0294	0.4870
Rationale-Generation Robot:Non-AI Group	0.6906	0.4668	1.4795	0.1390	1.9949
Action-Declaring Robot:Non-AI Group	1.4138	0.4561	3.0995	0.0019	4.1115
1|2	-0.4098	0.1999	-2.0501	0.0404	0.6638
2|3	1.1356	0.2076	5.4701	0.0000	3.1130

Table 16. OLR Summary - Confidence - Ref. Levels: Type Numerical-Reasoning Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.9513	0.3800	5.1343	0.0000	7.0377
Action-Declaring Robot	1.2992	0.3688	3.5230	0.0004	3.6664
AI Group	0.7196	0.3304	2.1778	0.0294	2.0536
Rationale-Generation Robot:AI Group	-0.6906	0.4668	-1.4795	0.1390	0.5013
Action-Declaring Robot:AI Group	-1.4138	0.4561	-3.0995	0.0019	0.2432
1|2	0.3098	0.2666	1.1620	0.2452	1.3631
2|3	1.8552	0.2812	6.5966	0.0000	6.3927

Table 17. OLR Summary - Friendliness - Ref. Levels: Type Rationale-Generation Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-5.8425	0.6241	-9.3621	0.0000	0.0029
Numerical-Reasoning Robot	-9.1613	0.6886	-13.3037	0.0000	0.0001
Non-AI Group	-1.2732	0.6531	-1.9494	0.0512	0.2799
Action-Declaring Robot:Non-AI Group	2.3121	0.7844	2.9477	0.0032	10.0953
Numerical-Reasoning Robot:Non-AI Group	0.4423	0.8811	0.5021	0.6156	1.5564
1|2	-7.4844	0.6292	-11.8956	0.0000	0.0006
2|3	-3.1140	0.5110	-6.0945	0.0000	0.0444

Table 18. OLR Summary - Friendliness - Ref. Levels: Type Rationale-Generation Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-3.5301	0.5314	-6.6431	0.0000	0.0293
Numerical-Reasoning Robot	-8.7179	0.7549	-11.5482	0.0000	0.0002
AI Group	1.2734	0.6531	1.9497	0.0512	3.5730
Action-Declaring Robot:AI Group	-2.3128	0.7844	-2.9485	0.0032	0.0990
Numerical-Reasoning Robot:AI Group	-0.4435	0.8809	-0.5035	0.6146	0.6418
1|2	-6.2112	0.5476	-11.3434	0.0000	0.0020
2|3	-1.8407	0.4068	-4.5247	0.0000	0.1587

Table 19. OLR Summary - Friendliness - Ref. Levels: Type Action-Declaring Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	5.8429	0.6241	9.3622	0.0000	344.7825
Numerical-Reasoning Robot	-3.3185	0.3827	-8.6712	0.0000	0.0362
Non-AI Group	1.0394	0.4342	2.3940	0.0167	2.8274
Rationale-Generation Robot:Non-AI Group	-2.3129	0.7844	-2.9486	0.0032	0.0990
Numerical-Reasoning Robot:Non-AI Group	-1.8692	0.7336	-2.5481	0.0108	0.1542
1|2	-1.6417	0.2599	-6.3158	0.0000	0.1936
2|3	2.7288	0.3585	7.6120	0.0000	15.3144

Table 20. OLR Summary - Friendliness - Ref. Levels: Type Action-Declaring Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	3.5301	0.5314	6.6431	0.0000	34.1267
Numerical-Reasoning Robot	-5.1877	0.6660	-7.7898	0.0000	0.0056
AI Group	-1.0394	0.4341	-2.3941	0.0167	0.3537
Rationale-Generation Robot:AI Group	2.3125	0.7844	2.9483	0.0032	10.1000
Numerical-Reasoning Robot:AI Group	1.8692	0.7336	2.5481	0.0108	6.4832
1|2	-2.6811	0.4164	-6.4392	0.0000	0.0685
2|3	1.6894	0.3421	4.9375	0.0000	5.4160

Table 21. OLR Summary - Friendliness - Ref. Levels: Type Numerical-Reasoning Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	9.1614	0.6887	13.3033	0.0000	9522.2302
Action-Declaring Robot	3.3185	0.3827	8.6712	0.0000	27.6181
Non-AI Group	-0.8299	0.5912	-1.4038	0.1604	0.4361
Rationale-Generation Robot:Non-AI Group	-0.4436	0.8809	-0.5036	0.6146	0.6417
Action-Declaring Robot:Non-AI Group	1.8693	0.7336	2.5481	0.0108	6.4835
1|2	1.6768	0.2812	5.9620	0.0000	5.3482
2|3	6.0473	0.4618	13.0936	0.0000	422.9598

Table 22. OLR Summary - Friendliness - Ref. Levels: Type Numerical-Reasoning Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	8.7179	0.7549	11.5480	0.0000	6111.4635
Action-Declaring Robot	5.1879	0.6660	7.7897	0.0000	179.0894
AI Group	0.8301	0.5912	1.4041	0.1603	2.2934
Rationale-Generation Robot:AI Group	0.4434	0.8809	0.5033	0.6148	1.5579
Action-Declaring Robot:AI Group	-1.8695	0.7336	-2.5483	0.0108	0.1542
1|2	2.5068	0.5200	4.8208	0.0000	12.2655
2|3	6.8773	0.6364	10.8059	0.0000	970.0040

Table 23. OLR Summary - Intelligence - Ref. Levels: Type Rationale-Generation Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-1.8632	0.2799	-6.6555	0.0000	0.1552
Numerical-Reasoning Robot	-0.9546	0.2801	-3.4088	0.0007	0.3850
Non-AI Group	-0.3321	0.3304	-1.0053	0.3147	0.7174
Action-Declaring Robot:Non-AI Group	1.0047	0.4542	2.2120	0.0270	2.7310
Numerical-Reasoning Robot:Non-AI Group	-0.1047	0.4626	-0.2262	0.8210	0.9006
1|2	-1.7485	0.2186	-7.9976	0.0000	0.1740
2|3	-0.2121	0.2011	-1.0547	0.2916	0.8089

Table 24. OLR Summary - Intelligence - Ref. Levels: Type Rationale-Generation Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-0.8585	0.3631	-2.3643	0.0181	0.4238
Numerical-Reasoning Robot	-1.0593	0.3716	-2.8510	0.0044	0.3467
AI Group	0.3321	0.3304	1.0053	0.3148	1.3939
Action-Declaring Robot:AI Group	-1.0046	0.4542	-2.2119	0.0270	0.3662
Numerical-Reasoning Robot:AI Group	0.1047	0.4626	0.2263	0.8210	1.1104
1|2	-1.4164	0.2748	-5.1533	0.0000	0.2426
2|3	0.1200	0.2651	0.4528	0.6507	1.1275

Table 25. OLR Summary - Intelligence - Ref. Levels: Type Action-Declaring Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.8631	0.2799	6.6555	0.0000	6.4440
Numerical-Reasoning Robot	0.9085	0.2704	3.3604	0.0008	2.4806
Non-AI Group	0.6725	0.3110	2.1623	0.0306	1.9592
Rationale-Generation Robot:Non-AI Group	-1.0047	0.4542	-2.2120	0.0270	0.3662
Numerical-Reasoning Robot:Non-AI Group	-1.1093	0.4502	-2.4641	0.0137	0.3298
1|2	0.1147	0.1878	0.6106	0.5415	1.1215
2|3	1.6511	0.2051	8.0509	0.0000	5.2125

Table 26. OLR Summary - Intelligence - Ref. Levels: Type Action-Declaring Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	0.8585	0.3631	2.3642	0.0181	2.3596
Numerical-Reasoning Robot	-0.2008	0.3593	-0.5589	0.5762	0.8181
AI Group	-0.6725	0.3110	-2.1623	0.0306	0.5104
Rationale-Generation Robot:AI Group	1.0047	0.4542	2.2120	0.0270	2.7310
Numerical-Reasoning Robot:AI Group	1.1093	0.4502	2.4641	0.0137	3.0323
1|2	-0.5579	0.2530	-2.2047	0.0275	0.5724
2|3	0.9785	0.2566	3.8132	0.0001	2.6605

Table 27. OLR Summary - Intelligence - Ref. Levels: Type Numerical-Reasoning Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	0.9546	0.2801	3.4088	0.0007	2.5977
Action-Declaring Robot	-0.9085	0.2704	-3.3604	0.0008	0.4031
Non-AI Group	-0.4368	0.3244	-1.3467	0.1781	0.6461
Rationale-Generation Robot:Non-AI Group	0.1047	0.4626	0.2263	0.8210	1.1104
Action-Declaring Robot:Non-AI Group	1.1093	0.4502	2.4641	0.0137	3.0323
1|2	-0.7939	0.2021	-3.9274	0.0001	0.4521
2|3	0.7425	0.2016	3.6836	0.0002	2.1013

Table 28. OLR Summary - Intelligence - Ref. Levels: Type Numerical-Reasoning Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.0593	0.3716	2.8510	0.0044	2.8844
Action-Declaring Robot	0.2008	0.3593	0.5589	0.5763	1.2224
AI Group	0.4368	0.3244	1.3467	0.1781	1.5477
Rationale-Generation Robot:AI Group	-0.1047	0.4626	-0.2263	0.8210	0.9006
Action-Declaring Robot:AI Group	-1.1093	0.4502	-2.4641	0.0137	0.3298
1|2	-0.3571	0.2628	-1.3587	0.1742	0.6997
2|3	1.1793	0.2693	4.3787	0.0000	3.2522

Table 29. OLR Summary - Potential - Ref. Levels: Type Rationale-Generation Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-1.4669	0.2736	-5.3619	0.0000	0.2306
Numerical-Reasoning Robot	-1.4049	0.2855	-4.9210	0.0000	0.2454
Non-AI Group	0.3139	0.3403	0.9226	0.3562	1.3688
Action-Declaring Robot:Non-AI Group	0.3615	0.4584	0.7888	0.4303	1.4355
Numerical-Reasoning Robot:Non-AI Group	-1.2945	0.4854	-2.6667	0.0077	0.2740
1|2	-1.7680	0.2168	-8.1558	0.0000	0.1707
2|3	-0.1478	0.1969	-0.7505	0.4530	0.8626

Table 30. OLR Summary - Potential - Ref. Levels: Type Rationale-Generation Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-1.1053	0.3729	-2.9642	0.0030	0.3311
Numerical-Reasoning Robot	-2.6994	0.4039	-6.6836	0.0000	0.0672
AI Group	-0.3139	0.3403	-0.9224	0.3563	0.7306
Action-Declaring Robot:AI Group	-0.3616	0.4584	-0.7889	0.4302	0.6966
Numerical-Reasoning Robot:AI Group	1.2944	0.4854	2.6666	0.0077	3.6489
1|2	-2.0820	0.2966	-7.0203	0.0000	0.1247
2|3	-0.4617	0.2792	-1.6534	0.0983	0.6302

Table 31. OLR Summary - Potential - Ref. Levels: Type Action-Declaring Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.4668	0.2736	5.3616	0.0000	4.3354
Numerical-Reasoning Robot	0.0620	0.2728	0.2271	0.8203	1.0639
Non-AI Group	0.6755	0.3077	2.1950	0.0282	1.9649
Rationale-Generation Robot:Non-AI Group	-0.3614	0.4584	-0.7885	0.4304	0.6967
Numerical-Reasoning Robot:Non-AI Group	-1.6561	0.4643	-3.5668	0.0004	0.1909
1|2	-0.3012	0.1880	-1.6022	0.1091	0.7399
2|3	1.3191	0.2005	6.5806	0.0000	3.7401

Table 32. OLR Summary - Potential - Ref. Levels: Type Action-Declaring Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.1053	0.3729	2.9641	0.0030	3.0201
Numerical-Reasoning Robot	-1.5941	0.3754	-4.2465	0.0000	0.2031
AI Group	-0.6755	0.3077	-2.1951	0.0282	0.5089
Rationale-Generation Robot:AI Group	0.3616	0.4584	0.7889	0.4302	1.4356
Numerical-Reasoning Robot:AI Group	1.6560	0.4643	3.5667	0.0004	5.2385
1|2	-0.9767	0.2543	-3.8407	0.0001	0.3766
2|3	0.6436	0.2510	2.5639	0.0104	1.9034

Table 33. OLR Summary - Potential - Ref. Levels: Type Numerical-Reasoning Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.4049	0.2855	4.9210	0.0000	4.0752
Action-Declaring Robot	-0.0620	0.2728	-0.2271	0.8203	0.9399
Non-AI Group	-0.9806	0.3456	-2.8371	0.0046	0.3751
Rationale-Generation Robot:Non-AI Group	1.2945	0.4854	2.6667	0.0077	3.6491
Action-Declaring Robot:Non-AI Group	1.6560	0.4643	3.5667	0.0004	5.2386
1|2	-0.3631	0.2060	-1.7628	0.0779	0.6955
2|3	1.2571	0.2162	5.8156	0.0000	3.5154

Table 34. OLR Summary - Potential - Ref. Levels: Type Numerical-Reasoning Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	2.6994	0.4039	6.6837	0.0000	14.8709
Action-Declaring Robot	1.5941	0.3754	4.2466	0.0000	4.9238
AI Group	0.9806	0.3456	2.8371	0.0046	2.6659
Rationale-Generation Robot:AI Group	-1.2945	0.4854	-2.6667	0.0077	0.2740
Action-Declaring Robot:AI Group	-1.6560	0.4643	-3.5667	0.0004	0.1909
1|2	0.6174	0.2798	2.2063	0.0274	1.8541
2|3	2.2377	0.2983	7.5020	0.0000	9.3718

Table 35. OLR Summary - Understandability - Ref. Levels: Type Rationale-Generation Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-2.6153	0.3274	-7.9873	0.0000	0.0731
Numerical-Reasoning Robot	-4.2054	0.3661	-11.4866	0.0000	0.0149
Non-AI Group	-0.4175	0.3664	-1.1393	0.2546	0.6587
Action-Declaring Robot:Non-AI Group	1.5455	0.4949	3.1227	0.0018	4.6903
Numerical-Reasoning Robot:Non-AI Group	-1.7585	0.7309	-2.4059	0.0161	0.1723
1|2	-3.5627	0.3026	-11.7746	0.0000	0.0284
2|3	-1.0644	0.2358	-4.5135	0.0000	0.3449

Table 36. OLR Summary - Understandability - Ref. Levels: Type Rationale-Generation Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Action-Declaring Robot	-1.0698	0.3807	-2.8100	0.0050	0.3431
Numerical-Reasoning Robot	-5.9638	0.6824	-8.7391	0.0000	0.0026
AI Group	0.4175	0.3664	1.1393	0.2546	1.5182
Action-Declaring Robot:AI Group	-1.5455	0.4949	-3.1227	0.0018	0.2132
Numerical-Reasoning Robot:AI Group	1.7584	0.7309	2.4058	0.0161	5.8032
1|2	-3.1452	0.3361	-9.3583	0.0000	0.0431
2|3	-0.6469	0.2808	-2.3041	0.0212	0.5237

Table 37. OLR Summary - Understandability - Ref. Levels: Type Action-Declaring Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	2.6151	0.3274	7.9869	0.0000	13.6684
Numerical-Reasoning Robot	-1.5903	0.2993	-5.3142	0.0000	0.2039
Non-AI Group	1.1277	0.3316	3.4009	0.0007	3.0884
Rationale-Generation Robot:Non-AI Group	-1.5454	0.4949	-3.1226	0.0018	0.2132
Numerical-Reasoning Robot:Non-AI Group	-3.3039	0.7154	-4.6183	0.0000	0.0367
1|2	-0.9474	0.2107	-4.4964	0.0000	0.3877
2|3	1.5506	0.2308	6.7170	0.0000	4.7143

Table 38. OLR Summary - Understandability - Ref. Levels: Type Action-Declaring Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	1.0698	0.3807	2.8099	0.0050	2.9147
Numerical-Reasoning Robot	-4.8941	0.6646	-7.3645	0.0000	0.0075
AI Group	-1.1281	0.3316	-3.4020	0.0007	0.3237
Rationale-Generation Robot:AI Group	1.5456	0.4949	3.1229	0.0018	4.6909
Numerical-Reasoning Robot:AI Group	3.3039	0.7153	4.6190	0.0000	27.2187
1|2	-2.0755	0.2980	-6.9645	0.0000	0.1255
2|3	0.4228	0.2591	1.6321	0.1027	1.5263

Table 39. OLR Summary - Understandability - Ref. Levels: Type Numerical-Reasoning Robot and Group AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	4.2054	0.3661	11.4867	0.0000	67.0472
Action-Declaring Robot	1.5901	0.2993	5.3136	0.0000	4.9043
Non-AI Group	-2.1759	0.6325	-3.4403	0.0006	0.1135
Rationale-Generation Robot:Non-AI Group	1.7584	0.7309	2.4058	0.0161	5.8031
Action-Declaring Robot:Non-AI Group	3.3039	0.7153	4.6189	0.0000	27.2188
1|2	0.6427	0.2166	2.9667	0.0030	1.9015
2|3	3.1410	0.2846	11.0345	0.0000	23.1261

Table 40. OLR Summary - Understandability - Ref. Levels: Type Numerical-Reasoning Robot and Group Non-AI Group

	Value	Std. Error	t value	p value	Odds Ratio
Rationale-Generation Robot	5.9638	0.6824	8.7390	0.0000	389.0826
Action-Declaring Robot	4.8940	0.6646	7.3643	0.0000	133.4890
AI Group	2.1759	0.6325	3.4403	0.0006	8.8105
Rationale-Generation Robot:AI Group	-1.7584	0.7309	-2.4058	0.0161	0.1723
Action-Declaring Robot:AI Group	-3.3039	0.7153	-4.6189	0.0000	0.0367
1|2	2.8186	0.5943	4.7431	0.0000	16.7534
2|3	5.3169	0.6257	8.4981	0.0000	203.7525