"Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?

Markus Langer
Universität des Saarlandes
Germany
markus.langer@uni-saarland.de

Tim Hunsicker
Universität des Saarlandes
Germany
tim.hunsicker@uni-saarland.de

Tina Feldkamp
Universität des Saarlandes
Germany
tina.feldkamp@uni-saarland.de

Cornelius J. König
Universität des Saarlandes
Germany
ckoenig@mx.uni-saarland.de

Nina Grgić-Hlača
Max Planck Institute for Software Systems, Max Planck Institute for Research on Collective Goods
Germany
nghlaca@mpi-sws.org

ABSTRACT

In the media, in policy-making, but also in research articles, algorithmic decision-making (ADM) systems are referred to as algorithms, artificial intelligence, and computer programs, amongst other terms. We hypothesize that such terminological differences can affect people's perceptions of properties of ADM systems, people's evaluations of systems in application contexts, and the replicability of research as findings may be influenced by terminological differences. In two studies (N = 397, N = 622), we show that terminology does indeed affect laypeople's perceptions of system properties (e.g., perceived complexity) and evaluations of systems (e.g., trust). Our findings highlight the need to be mindful when choosing terms to describe ADM systems, because terminology can have unintended consequences, and may impact the robustness and replicability of HCI research. Additionally, our findings indicate that terminology can be used strategically (e.g., in communication about ADM systems) to influence people's perceptions and evaluations of these systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**;

KEYWORDS

human-centered AI, terminology, research methodology, replicability

ACM Reference Format:
Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3491102.3517527

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517527

1 INTRODUCTION

When the public discusses algorithmic decision-making systems (ADM systems) – systems that either automate decision-making or support human decision-making – when journalists report about such systems, and when policy-makers develop regulations about such systems, there is a variety of terms used to refer to them. For instance, newspaper articles refer to such systems as intelligent systems [42], as algorithms [13], or robotic systems [22]. Likewise, there is large variety in terminology used to refer to ADM systems in policy-making documents. For instance, within the European Commission's "Ethics Guidelines for Trustworthy AI" [50], the authors refer to ADM systems as algorithms, artificial intelligence, AI technologies, AI systems, and robots whereas the General Data Protection Regulation (GDPR) refers to ADM systems as automated means.

Similar variation in the terminology used to refer to ADM systems also occurs in research investigating interactions between humans and ADM systems. In such research, researchers develop materials where they describe the respective system to their participants. For instance, researchers might be interested in how trustworthy their participants perceive a system to be [39] or may investigate whether participants accept the respective system [34]. In such studies, research has used the terms algorithm [39], automated system [33], artificial intelligence [40], computer program [28], machine learning [26], sophisticated statistical model [17], or robot [51] – all to refer to a system that either automates decision-making or that supports human decision-making in a variety of application contexts (e.g., for systems that support hiring decisions [34], medical decisions [33], or bail decisions at court [28]).

Whereas all those terms reflect a similar idea – a system that interacts with humans – they might induce very different mental pictures, expectations, and thoughts associated with the ADM system in question. More generally, presenting participants a system

using the term "automated system" versus "algorithm" versus "artificial intelligence" may affect how people perceive and evaluate these systems. On the one hand, this may affect the robustness and replicability of HCI research as findings may vary between studies only because of terminological differences. For instance, people's acceptance of an ADM system in medicine might differ depending on whether the system is described as an "algorithm" or as a "computer program". On the other hand, communicating about ADM systems (e.g., in policy-making) using the term "automated system" versus "artificial intelligence" might alter what people expect when they hear the respective term. For instance, an "automated system" might sound less advanced compared to using "artificial intelligence" and this could affect initial perceptions of trustworthiness with respect to the system in question because "artificial intelligence" is associated with a system with more potential than an "automated system".

In this paper, we propose that terminology crucially affects the ways in which people perceive and evaluate ADM systems. More precisely, we argue that the choice of the term used to refer to ADM systems will affect people's perceptions about the properties of the system (e.g., perceived complexity) as well as people's evaluation of the system (e.g., trust evaluations) in application contexts. We conducted two experimental, between-subject studies to test whether terminology matters, and if different terminology can cause different effects in communication about ADM systems. In the first study, we varied ten terms that research has used to refer to ADM systems to explore how this affects people's perceptions of properties of the respective systems. Additionally, we examined terminological effects on people's evaluation of whether systems or humans are better able to conduct a set of different tasks (e.g., medical diagnoses, criminal recidivism prediction). In the second study, we used vignettes of a well-known study in HCI by Lee [39] and varied the term used within those vignettes to test if evaluations of fairness and trust in application contexts differ depending on the terminology used to refer to ADM systems.

Contributions. In this paper, we contribute to research on HCI by showing that terminological differences affect

- Human perceptions of properties of ADM systems (e.g., perceived complexity)
- Human evaluations of systems (e.g., trust)

We thereby highlight the importance of terminology in communication about ADM systems. On the one hand, variation in terminology can have unintended negative effects on the robustness and replicability of HCI research. On the other hand, terminology can be used strategically to steer human perceptions and evaluations of such systems.

2 RELATED WORK

2.1 Why terminology may matter

Studies throughout disciplines have shown the importance of terminology as it can affect human perceptions, emotions, and behavior [19, 55, 58]. We propose that terminological differences used to refer to ADM systems in HCI will affect people's perceptions and evaluations of ADM systems. Specifically, research has used a variety of terms to refer to ADM systems [30, 36, 57] which applies to the description of ADM systems within papers but more crucially to the communication about ADM systems when presenting them to research participants. For instance, Wang et al. [67] told their participants that an "algorithm" processes their MTurk work history, decides who will get a promotion (i.e., become a master worker), and then asked participants to evaluate fairness of the algorithm-based decision. In a school admission scenario, Marcinkowski et al. [45] told participants that an "AI technology" analyzes applicant data and recommends applicant approval or rejection. They also asked for participants' evaluations of the fairness of the AI technology's decision. In a work scheduling setting, Uhde et al. [62] told participants that a "system" decides who gets vacation and asked them to report how they perceive and evaluate system-based decisions for scheduling. Even in single papers presenting multiple studies, terminology to refer to ADM systems might vary. For example, Longoni et al. [43] present multiple studies on the acceptance of AI in healthcare (e.g., in skin cancer screening). In their studies they described to participants that the respective ADM system is a "computer [that] uses an algorithm", "a computer that is capable of artificial intelligence", "a computer program" or "a well-trained algorithm" that provides outputs that help to make medical decisions. As another example, Binns et al. [6] asked participants about their evaluations of situations where a "computer system" or a "predictive model" is used to decide whether a person should receive a promotion.

Terminology effects might be especially influential in previous studies because participants often received limited information regarding the system in question. In fact, Langer and Landers [36] reviewed research on people's perceptions and evaluations of automated systems in different decision-making situations (e.g., management, healthcare). In many of the studies they reviewed, the term to refer to the system was the main experimental manipulation as it was this term that informed people about the fact that there is an ADM system automating decisions or supporting decision-making. For instance, Nagtegaal [47] told participants that a "computer, using an automated algorithm", decides about travel reimbursement or evaluates employee performance. In Langer et al. [34], the only information their participants, who had to record responses to job interview questions, received was that a "computer will automatically analyze the audio recordings and evaluate [their] answers". In both these examples, the focus seems to be on the automation of a decision by an ADM system without further specifying this system. In further examples, Shaffer et al. [56] described to participants who had to rate the expertise of doctors that a "doctor [...] indicates she is going to use a decision aid [computer program]" and Dietvorst et al. [17] described to their participants who had the option to use outputs by a model as additional information to forecast student performance that "the admissions office had created a statistical model that was designed to forecast student performance" and provide the additional information that this model is "sophisticated". In both these examples, ADM systems were introduced to support decision-making but there was no further information about underlying technology or about, for instance, how the system produces its outputs. In other work (e.g. [51]), terminology such as robot may have been chosen deliberately to describe an embodied ADM system and to additionally anthropomorphize the system by describing it as a humanoid robot. Importantly, in these and in many more studies investigating people's reactions to (partly)

automated decision-making [36], there was limited additional information regarding the functionalities or performance of the system, limited information regarding how the system works, and especially a limited rationale regarding why respective authors chose a specific term to describe an ADM system to participants. Without additional information about how a respective system works, or about functionalities of a system, people need to rely on salient aspects within study information to form their perceptions and evaluations of the respective situation [1, 4]. This kind of salient information can be the term used to refer to ADM systems.

To decide which terms to investigate, we drew on Langer and Landers' [36] review that provides an overview on the terms research has used to describe ADM systems. Additionally, we added two more terms that have been used to refer to ADM systems in studies not included in Langer and Landers' review. We added the term "technical system" [46] as a term that is very generic, as well as the term "sophisticated statistical model" [17] as a term that is very specific. Table 1 presents the final set of terms we decided to investigate as well as sample sources that have used these terms in their studies.

Table 1: Terminological differences to refer to ADM systems with the 10 terms used in Study 1 and exemplary studies that have used these terms.

Term	Exemplary Study
Algorithm	Lee [39]
Automated system	Keel et al. [33]
Artificial intelligence	Marcinkowski et al. [45]
Computer	Langer et al. [34]
Computer program	Grgić-Hlača et al. [27]
Decision support system	Shibl et al. [60]
Machine learning	Gonzalez et al. [26]
Technical system	Montague et al. [46]
Robot	Ötting and Maier [51]
Sophisticated statistical model	Dietvorst et al. [17]

2.2 Consequences of terminological differences

In this paper, we empirically investigate two broad consequences of terminological differences when referencing ADM systems. First, we explore consequences for perceptions of properties of ADM systems. For this, we investigate what kind of properties people associate with different references to ADM systems, irrespective of the context in which the ADM system is used. In other words, to shed light on the properties associated with the respective term to describe ADM systems, we chose to only vary the term and to not give any additional information (e.g., on system functionalities or the application context). Understanding how terminology affects perceptions of properties associated with the entity is important as this might provide us with insights regarding what basic properties are associated with different terms, which might allow conclusions regarding more downstream consequences (e.g., acceptance of systems).

Second we explore consequences for evaluations of ADM systems in application contexts. This means we explore whether different terminology to describe an ADM system in an application context can differently affect people's evaluations of the respective system. This is important because it allows insights regarding whether and to what extent using different terms to describe ADM systems in application contexts may affect people's evaluations of ADM systems (e.g., regarding trust, fairness).

2.2.1 Consequences of terminological differences for perceptions of the properties of ADM systems. We chose to assess six properties associated with ADM systems: tangibility, complexity, controllability, familiarity, anthropomorphism, and machine competence. We chose these properties because they can be evaluated without putting ADM systems in an application context and because research has shown them to be related to more downstream consequences such as acceptance of systems, or human behavior in the interaction with systems.

Tangibility. Tangibility is associated with people having a shape in mind when they think about a term and whether a term is associated with an entity humans can touch [25]. People may interact differently with agents having a physical appearance compared to disembodied agents, perceive them as more socially present [38, 41], and may have different expectations regarding relationship building with more tangible entities [25]. With respect to the terms we use, we imagine that terms such as "computer" or "robot" are more likely perceived as tangible compared to terms such as "algorithm" or "artificial intelligence" since the former have a shape while the latter reflect disembodied manifestations of ADM systems.

Complexity. In this paper, high complexity would mean people believe the entity described by the term is hard to understand, including its functionalities and its design-process, and for which it is hard to comprehend how it works [24, 47]. Perceived complexity can be associated with the acceptance of systems [47] and with beliefs about system quality [20]. Regarding the terms, "computer program" might be perceived to be less complex compared to "artificial intelligence" because even though people might not understand how computer programs work, artificial intelligence may be associated with more complex technologies.

Controllability. Controllability is associated with whether people believe humans can control the behavior of the entity described by the term. Perceived controllability relates to the acceptance of systems [63, 64]. With respect to the different terms, "computer" might be associated with an entity that is more controllable compared to "robot" because people have already operated the former and might believe that the latter is acting more autonomously [52].

Familiarity. If a term is associated with something that is familiar, people have already heard of the term, have had experience with using the entity associated with this term, and believe that the entity is something that is a part of everyday life. Familiarity is, for instance, associated with better acceptance of systems [14, 63]. We, for instance, imagine that "computer program" is perceived to be more familiar than "machine learning" since computer programs are something people use every day, whereas machine learning reflects a more specific concept where only experts would say that it is familiar to them.

Anthropomorphism. Anthropomorphism refers to whether people perceive the term describing an entity as possessing human-like characteristics [15, 21]. For instance, anthropomorphism can be associated with believing that an entity has intentions or makes

autonomous decisions. Anthropomorphism is an important variable in agent design where virtual agents can be designed more or less anthropomorphic in order to affect human-agent interaction patterns [3, 12, 15]. Regarding different terms, it is possible that people perceive "technical system" to be less associated with human-like characteristics compared to "robot" or "artificial intelligence" that are often presented as having or evolving these characteristics in popular media.

Machine competence. Under machine competence, we understand whether a term is associated with an entity that has great capabilities and strong potential regarding its successful application in different contexts [25]. Machine competence is usually associated with high expectations regarding the performance of ADM systems and may thus determine whether people use a respective system [25, 29, 37]. Regarding the different terms, the capabilities that people ascribe to "artificial intelligence" might be stronger compared to capabilities associated with "decision support system" because artificial intelligence may sound like something with broader application possibilities than decision support systems.

Considering the different perceptions different terminologies regarding ADM systems can invoke, we propose the following research question:

Research Question 1: Does varying the terminology regarding ADM systems affect people's perceptions of the properties of ADM systems? [1]

2.2.2 Consequences for evaluations of ADM systems in application contexts. Up to this point, we have focused on perceptions or properties associated with ADM systems without considering the application context in which these systems may operate. Consequences of terminological differences become even more important when considering the evaluation of systems in specific application contexts. Specifically, varying the term used to refer to ADM systems may affect whether people positively or negatively evaluate the use of said system in a respective context, and may lead to a lack of acceptance or disuse just due to terminological differences and not actual differences in system-design or functionalities [32].

To investigate whether terminological differences affect people's evaluations of ADM systems in application contexts we a) examine whether terminological differences affect evaluations regarding the ability of systems to conduct a set of different tasks, and b) investigate whether terminological differences affect evaluations of fairness and trust in systems as well as robustness and replicability of research by replicating a well-known study on evaluations of ADM systems in application contexts (i.e., [39]).

Regarding a), we thus chose a set of different tasks that are associated with the use of ADM systems (e.g., making shopping recommendations, evaluating applicant documents, providing therapy recommendations in medicine) to explore whether the term used to refer to ADM systems affects whether people evaluate a system to be able to perform a respective task. We chose a set of tasks that reflects a variety of application contexts as well as different tasks in single application contexts (e.g., in medicine). Since recent work shows emerging interest in understanding the tasks where people believe systems to perform better or at least equally well as human beings (see e.g., [8, 16, 36, 39]), we wanted to investigate whether the evaluation of the performance of systems in such tasks also depends on the terminology to describe the system.

Research Question 2: Does varying the term to refer to ADM systems affect people's evaluation regarding the performance of systems in various tasks?

Regarding b), instead of devising a novel study paradigm, we chose to replicate Lee's [39] well-known study on evaluations of fairness and trust in different application contexts and varied the terminology she used to refer to the respective ADM system described in her study. She presented participants with textual vignettes that described one of four application contexts (work assignment, work scheduling, hiring, and work evaluation) where an ADM system described with the term "algorithm" provided decisions that affect human decision-recipients. She found that for tasks that afford human skills (hiring, work evaluation) people evaluated the algorithm to be less fair and participants trusted the algorithm less in these application contexts compared to human decisions. In contrast, she found less, and non-significant, differences between the human manager and the algorithm for tasks that afford mechanical skills (work assignment, work scheduling).

We propose that using a different term than "algorithm" might affect the results of her study and consequently the conclusions we can draw from the study. Specifically, instead of "algorithm", it is equally possible to refer to the system that produces a decision as an "automated system" which may affect people's evaluations of the respective system. For example, if people evaluate algorithms to be more capable of conducting a specific task compared to automated systems, this could lead to different levels of trust. Similarly, if people evaluate automated systems to be more consistent in decision-making than algorithms, this could affect fairness perceptions. If we find that terminological differences indeed affect the conclusions we draw from the respective study (e.g., for certain terms we find stronger, significant effects, whereas for others we find smaller, non-significant ones), we might need to infer that parts of variability in findings from previous research were due to differing terminology to refer to ADM systems [36]. Additionally, finding that terminological differences can affect the conclusions we draw from research would indicate that it is necessary to be more mindful when choosing the terminology to describe ADM systems to participants in studies.

In addition to the evaluation of trust and fairness that Lee [39] investigated in her study, we chose to also capture perceived procedural justice [11] as a related concept. Furthermore, since Lee [39] investigated different application contexts in her study, we took the opportunity to investigate whether terminological differences affect evaluations of systems differently depending on the application context. If this would be the case, we would find an interaction effect in our study results, indicating that the effect of different terminology may also depend on the task for which a respective ADM system is used. In other words, terminological effects may be stronger for one task than for another. Overall, we thus propose the following research questions:

[1] Before data collection for the respective studies started, we preregistered the research questions, dependent variables that we wanted to capture, experimental manipulations, data exclusion plan, data analysis plan, and planned number of participants to include in the studies. The respective blinded preregistrations are available under https://aspredicted.org/LDC_GSM and https://aspredicted.org/NTE_WND

Research Question 3: Does varying the term to refer to ADM systems affect people's evaluation of ADM systems (in our case evaluations of trust, fairness, and procedural justice)?

Research Question 4: Will the term used to refer to ADM systems and the task for which ADM systems is used interact to affect evaluations of ADM systems?

3 METHODOLOGY

3.1 Study 1

Study 1 investigated people's perceptions of properties associated with different terms to refer to ADM systems. Additionally, Study 1 shed light on people's perceptions regarding whether the term to refer to the system affects the perceived ability of systems to perform several different tasks.

3.1.1 Sample. We gathered data on Prolific. The only inclusion criterion was that participants were native English-speakers and 18 years or older. Completing the study took on average 8 (SD = 2) minutes and participants received 1.27 British Pounds as payment. We gathered data from N = 417 participants. We excluded 6 participants because they did not recall the correct term that was used in their version of the study as well as 14 participants because they failed the included attention check item. The final sample consisted of N = 397 participants (65% female; 35% male), with a mean age of 35 years (SD = 13); 69% of participants were employed. 21% of participants were students. Furthermore, 7% of participants reported their highest education level as "attended high school", 19% reported that they have a high school degree, 18% reported a 2-year community/technical/professional/trade college degree, 35% reported a 4-year college or university degree, and 20% reported a graduate degree or PhD. A majority of participants was from the United Kingdom (78%), with the rest of participants coming from South Africa (5%), Australia (4%), the US (3%), Ireland (2%), and small numbers of participants from various other countries. Since we imagined that participants' interest in and prior exposure to different technologies may affect how they perceive the different terms, we measured participants' affinity for technology [23] as a possible control variable. The mean value of participants' affinity for technology was M = 3.02 (SD = 1.03) (See Figure 1). This indicates a mean value that would correspond to the label "slightly disagree" in the response options to the affinity for technology scale but also some between-participant variation regarding affinity for technology.

3.1.2 Procedure. Study 1 was conducted online and followed a randomized experimental design with 10 between-subject conditions. This means that a single participant was presented with exactly one of the terms to refer to ADM systems presented in Table 1. Following, each time we use "the term" we use it as a placeholder for the experimentally manipulated terms.

Participants accessed the study through a link that directed them to the first page in the online questionnaire tool (we used SoSci Survey). After providing informed consent, participants were randomly assigned to their respective experimental condition. First, participants received a set of items that asked for their perceptions of properties associated with the entity described by the respective term (see a screenshot in Appendix Figure 8). Participants responded to

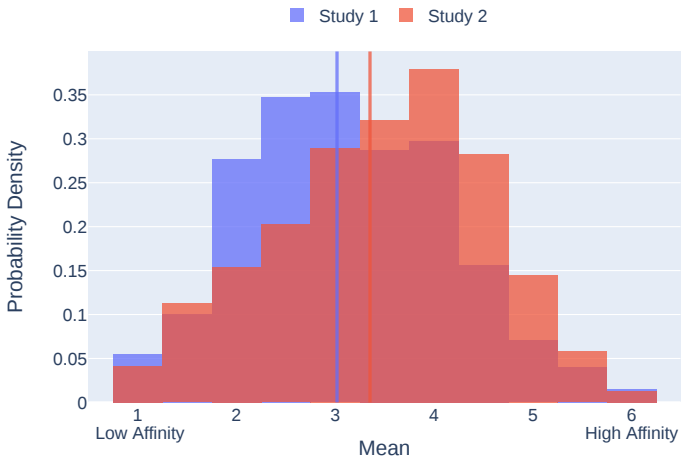

Figure 1: Distribution of participants' responses to the affinity for technology scale in Study 1 (blue) and Study 2 (red), and the corresponding mean values.

items assessing tangibility, complexity, controllability, familiarity, anthropomorphism, and machine competence. Second, participants were asked to evaluate how well they believe ADM systems will perform different tasks in comparison to humans. Specifically, we used thirteen tasks commonly associated with the use of systems that covered a range of different application settings (see Section 3.1.3; see a screenshot in Appendix Figure 9). Third, participants responded to a scale assessing their affinity for technology [23] which served as a control measure to investigate whether general affinity for technology affects participants' perceptions of respective terms. Fourth, participants reported demographic information (gender, age, education level, whether they are students, and whether they are employed). [2]

To ensure data quality, participants responded to two attention check items: the first one asked them to respond "strongly disagree" to the respective attention check item, the second one asked them to report which of the ten terms they were presented with during the study.

3.1.3 Measures. Unless otherwise stated, participants responded to the items measuring the dependent variables on a scale from 1 (strongly disagree) to 5 (strongly agree) – "the term" was replaced with one of the terms reflecting our experimental manipulation. All items for Study 1 can be found in the Appendix (see Appendix Table 3). As a measure of scale reliability, for all scales with more than two items, we report Cronbach's α; for two item scales, we report the Spearman-Brown correlation as suggested by Eisinga et al. [18].

[2]Participants also responded to the negative attitudes towards robot scale [49], where we replaced the term robot with the respective term to refer to ADM systems. Participants also responded to the Godspeed scale [5]. Furthermore, they were asked to report "What is 'the term' for you?", and we included an item asking for their knowledge regarding the respective term. Finally, participants were asked to respond to the question "Could you give us an example of 'the term' that you have heard of or already used for work or in your free time?" Results for these measures can be made available upon request.

Tangibility was measured with two self-developed items and assessed whether people have a clear picture or shape in mind when thinking about the respective term [25]. A sample item was "When I think of 'the term', I have a clear picture in mind" (Spearman-Brown correlation = .67).

Complexity was measured with three self-developed items and assessed whether people believe that the term reflects something complex and non-comprehensible [20]. A sample item was "'the term' is complex" (Cronbach's α = .70). [3]

Controllability was measured with two self-developed items that assessed whether people believe that the term reflects something that is controllable for humans or whether the term reflects something that acts autonomously [52]. A sample item was "'the term' is controllable by humans" (Spearman-Brown correlation = 0.78). [4]

Familiarity was measured with three self-developed items that should reflect familiarity as described by Luhmann [44]. A sample item was "'the term' is something I encounter in everyday life" (Cronbach's α = .68).

Anthropomorphism was measured with eight items taken from Shank and DeSanti [59]. A sample item was "'the term' has intentions" (Cronbach's α = 0.82).

Machine competence was measured with six self-developed items which assessed perceptions of high capabilities associated with the term [25]. A sample item was "'the term' has great potential in terms of what it can be used for" (Cronbach's α = 0.77).

Perceived ability to perform several tasks in comparison to humans was measured for thirteen tasks: shopping recommendations, evaluating applicant documents, scheduling work, predicting criminal recidivism, making medical diagnoses, evaluating X-rays and MRIs, predicting the weather, evaluating job interviews, therapy recommendations in medicine, diagnosing mental illness, identifying faces, assessing dangerous situations while driving, and predicting the spread of infectious diseases. For instance, participants read: "'the term' can make shopping recommendations" and were then asked to rate this statement on a scale from 1 (worse than a human) to 5 (better than a human) with the middle category 3 (as good as a human).

The control variable affinity for technology was measured with four items taken from Franke et al. [23]. For this measure, we used the original response scale from 1 (completely disagree) to 6 (completely agree) (Cronbach's α = .83).

3.2 Study 2

To investigate effects of terminological differences on evaluations of ADM systems (e.g., trust), and to explore whether terminological differences affect robustness and replicability of research, Study 2 followed the methodology of Lee [39] and thus partly replicated her study that examined human evaluations of ADM system-based decisions in different application scenarios.

3.2.1 Sample. We again gathered data on Prolific. The inclusion criteria were that participants were native English-speakers and 18 years or older, and that they had participated in at least 10 studies on Prolific and had a 100% approval rate. Completing the study took on average 4 minutes (SD = 1) and participants received 0.67 British Pound as payment. We gathered data from N = 722 participants. We excluded 24 participants because they did not recall the correct term that was used in their version of the study indicating that they were not attentive. Furthermore, we excluded 76 participants because they failed the included attention check item. The final sample consisted of N = 622 participants (62% female; 38% male), with a mean age of 36 years (SD = 13); 71% of participants were employed. 17% of participants self-reported to be students. Furthermore, 7% of participants reported their highest education level as "attended high school", 20% reported that they have a high school degree, 15% reported a 2-year community/technical/professional/trade college degree, 42% reported a 4-year college or university degree, and 16% reported a graduate degree or PhD. A majority of participants was from the United Kingdom (77%), with the rest of participants coming from South Africa (4%), the US (4%), Canada (3%) Ireland (3%), and small numbers of participants from various other countries. Interest in technology as well as prior exposure to technology could affect the evaluation of the terms in application context, we thus again measured participants' affinity for technology. The mean value of participants' affinity for technology was M = 3.35 (SD = 1.08) (See Figure 1). Similar to Study 1, this indicates a mean value that would correspond to the label "slightly disagree" in the response options to the affinity for technology scale but some between-participant variation.

3.2.2 Reducing the number of terms to include in Study 2. To reduce the complexity of the study, we wanted to use fewer terms in Study 2. To determine which terms to keep, we used Google's Universal Sentence Encoder (USE) [10] to estimate the semantic similarity between the 10 terms used in Study 1. Specifically, we estimated which terms are semantically most similar and which ones more different, and used this information to determine which terms to include in Study 2. Google's USE has been trained on unsupervised training data from web sources such as Wikipedia and discussion forums, and supervised data from the Stanford Natural Language Inference corpus [7], and was shown to perform well on the Semantic Textual Similarity Benchmark [9]. We utilized USE to encode our terms into 512-dimensional embedding vectors and, as suggested by Cer et al. [10], we calculated the angular similarity between these vectors in order to estimate the semantic similarity between the terms. We applied hierarchical clustering on the resulting distance matrix [5], using the UPGMA algorithm implemented in SciPy [66]. The resulting clusters are shown in Figure 2.

Based on the results for the semantic similarity analysis, we argue that there are four high-level clusters. The first cluster consisted of the term sophisticated statistical model. The second cluster

[3] In our survey, we had included five items to capture complexity which can be found in Appendix Table 3. Two of the items led to a low Cronbach's α. Following Allen et al. [2] and Peterson [53], we removed these items from the scale. Note that exclusion or inclusion of these items did not substantially change our interpretations for complexity. In both cases, we would find that different terminology affects perceived complexity.

[4] In our survey, we had included three items to capture controllability which can be found in Appendix Table 3. One of the items led to a low Cronbach's α. Following Allen et al. [2] and Peterson [53], we removed this item from the scale. Note that exclusion or inclusion of this item did not substantially change our interpretations for controllability. In both cases, we would find that different terminology affects perceived controllability.

[5] The distance between two terms encoded into 512-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ is calculated as $\text{dist}(\mathbf{u}, \mathbf{v}) = \arccos\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}\right)/\pi$, where $\mathbf{u}\cdot\mathbf{v}$ is the dot product of $\mathbf{u}$ and $\mathbf{v}$, and $\|*\|$ is the Euclidean norm of its argument $*$.

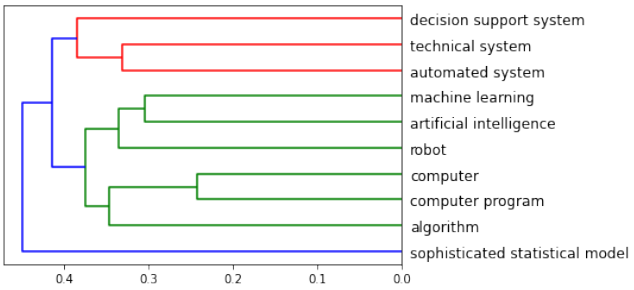

Figure 2: Clusters based on the semantic similarity between terms. The semantic similarity was estimated using Google's USE [10].

included algorithm, computer program, and computer. The third cluster included robot, artificial intelligence, and machine learning. The fourth cluster included automated system, technical system, and decision support system. For Study 2, we decided to use one term from each cluster which is why we chose: sophisticated statistical model, computer program, artificial intelligence, and automated system. Furthermore, we included the term that Lee [39] also used in her study: algorithm. Since all of these terms may reflect disembodied manifestations of ADM systems, we decided to also include one of the terms that previous work has used to describe an embodied ADM system [51]: robot.

3.2.3 Procedure. Study 2 was conducted online and followed a randomized 7 (condition term: six different terms plus human condition) x 2 (condition task: work assignment versus work evaluation) experimental between-subject design. To be clear, each participant was thus presented with exactly one term in one task. As described in Section 3.2.2, we focused only on a subset of terms in comparison to Study 1: algorithm, automated system, artificial intelligence, computer program, robot, and sophisticated statistical model. In line with Lee [39], we also included a human condition where participants read "a manager" instead of a term referring to a system. Our second experimental factor was the application context where we had the conditions work assignment and work evaluation. We decided to include these as they were also used in Lee [39] and because participants in Lee's study perceived system decisions to be much fairer and participants trusting these decisions to a larger extent for work assignment compared to work evaluation contexts.

After providing informed consent, participants received initial information on the experimental setting. Specifically, participants read that "In the situation below, 'the term' makes a decision autonomously without human intervention." After this information participants were introduced to the respective decision situation reflecting their experimental condition. Specifically, we used the textual vignettes developed by Lee [39] verbatim with two changes. First, we replaced the term "algorithm" that she used in these vignettes in her study with the respective term of the given experimental condition. Additionally, we standardized the name of the person in the textual vignette to be "Chris" in every condition. The vignettes for the tasks were the following (see also screenshots in Appendix Figures 10 and 11):

- Work assignment: "In the following situation, 'the term' makes a decision autonomously without human intervention. In a manufacturing factory, 'the term' assigns their employees to check and update certain components of the machinery to prevent any critical operation failures. The component assignment is based on data that show how often different components have worn out and broken down in the past. Chris works in the manufacturing factory. 'The term' assigns him to check a specific component of the machinery and he does the maintenance work on it.
- Work evaluation: "In the following situation, 'the term' makes a decision autonomously without human intervention. In a customer service center, 'the term' evaluates employees by analyzing the content and tone of their calls with customers. Chris works at the customer service center. Based on past call recordings, 'the term' evaluates his performance."

After reading their respective vignette, participants were asked to respond to the fairness and trust item as well as to the procedural justice items. Afterwards, participants responded to the affinity for technology items and to the demographic questions (gender, age, education level, whether they are studying, and whether they are employed).

Throughout Study 2, participants responded to two attention check items. The first one asked them to check the response option "to a large extent". The second one asked them to report which of the terms they were presented with during the study.

3.2.4 Measures. All items for Study 2 can be found in the Appendix (see Appendix Table 4). Fairness was measured with one item taken from Lee [39] that differed slightly with respect to the decision situation. In the case of work assignment this item was: "How fair or unfair is it for Chris that 'the term' assigns him to check a specific component of the machinery and he does the maintenance work on it?" In the case of work evaluation this item was: "How fair or unfair is it for Chris that 'the term' evaluates his performance?" Participants responded to this item on the same scale used by Lee [39] ranging from 1 (very unfair) to 7 (very fair).

Trust was measured with one item taken from Lee [39] that differed slightly with respect to the decision situation. In the case of work assignment this item was: "How much do you trust that 'the term' makes a good-quality work assignment?" In the case of work evaluation this item was: "How much do you trust that 'the term' makes a good-quality work evaluation?" Participants responded to this item on the same scale used by Lee [39] with a scale from 1 (do not trust at all) to 7 (extremely trust).

Procedural Justice was measured with seven items taken from Colquitt [11] (Cronbach's α = .72). A sample item was "Have those procedures been free of bias?" Participants responded to these items on the original scale from 1 (to a very small extent) to 5 (to a very large extent). Note that this scale was not captured in Lee's study.

We again measured affinity for technology by Franke et al. [23] as a possible control variable with the same items as in Study 1 (Cronbach's α = .84). [6]

[6]Participants were also asked to respond to the question "In your own words, please briefly explain what you think 'the term' is.", and to report their knowledge or the respective term with one item. Results can be made available upon request.

4 RESULTS

4.1 Study 1 Results

4.1.1 Perceptions regarding the properties of ADM systems. Research Question 1 asked whether varying the terminology regarding ADM systems affects perceptions about the properties of ADM systems. To analyze our data, we utilized linear regressions. For each of the six properties (i.e., tangibility, complexity, controllability, familiarity, anthropomorphism, machine complexity), we used separate linear regression models, each including one of the properties as the dependent variable. As independent variables, we used the 10 terms, dummy-coded with the term artificial intelligence as reference group. [7] To be clear, this means we included nine dummy-coded variables into the regression models, where in the first dummy-coded variable the term algorithm was coded with 1, and all other terms were coded with 0, in the second dummy-coded variable the term automated system was coded with 1 and all other terms with 0. In the end, each term is represented in one variable with the coding 1, except for the term artificial intelligence which always received the coding 0 to remain the reference group to which all other groups will be compared. The results regarding how the respective terminology affected the perceived properties of ADM systems are presented in Figure 3 and can be interpreted in comparison to the reference group artificial intelligence (e.g., how do familiarity perceptions differ between the term artificial intelligence and the term computer program). We additionally entered education level, gender, as well as mean-centered versions of the variables age and affinity for technology as control variables in the regression in order to test whether they affect our results. Education level, age, and gender only showed minor effects on the results which is why we did not include these variables in our final models. However, we included affinity for technology because it was consistently correlated with participants' perceptions of the properties of ADM systems. Participants with a higher affinity for technology perceived ADM systems to be more tangible, less complex, more controllable, found them to be more familiar, were less likely to anthropomorphize systems, and ascribed higher machine competence. [8] (This is reflected in Figure 3, where the last row of each graph presents the regression weight for affinity for technology; this means if the dot for affinity for technology is right to the zero line, affinity for technology was positively associated with the respective property, if it is left to the zero line, it was negatively associated with the respective property.) The final set of variables in our models thus included the nine dummy-coded variables that reflect the comparison of the respective terms to the reference group artificial intelligence as well as the control variable affinity for technology.

Results for **tangibility** showed that computers and robots were perceived as more tangible than artificial intelligence. In contrast, decision support systems, machine learning, sophisticated statistical models, algorithms, and technical system were perceived as comparably less tangible. For **complexity**, results indicated that the term artificial intelligence is perceived to be associated with an entity that is more complex than automated systems, decision support systems, and computers. Regarding **controllability**, results showed that computers, robots and computer programs were perceived as more controllable than artificial intelligence. Results for **familiarity** revealed that people perceived computers and computer programs to be especially more familiar than artificial intelligence, but also algorithms, automated systems, and technical systems. In contrast, sophisticated statistical models and robots were perceived as less familiar. For **anthropomorphism**, our findings showed that the term artificial intelligence was more strongly anthropomorphized than the majority of the other terms, especially computers, computer programs, technical system and automated systems. Finally, participants associated relatively high **machine competence** with artificial intelligence, computers, and computer programs whereas they perceived especially less machine competence for decision support systems and sophisticated statistical models. Also machine learning, automated systems, and algorithms were perceived as having less machine competence than artificial intelligence.

Another aspect that revealed the influence of terminological differences is the R^2 statistic found for the properties associated with ADM systems (see Figure 3). This statistic reveals how much variance in participant responses can be explained by the included predictors. In the case of familiarity this means that 50% of variance is explained by affinity for technology and the terminological differences. When excluding affinity for technology, there was 44% explained variance due to different terminology. For the other variables, excluding affinity for technology from the model resulted in 36% explained variance for tangibility, 18% for machine competence, 14% for controllability and anthropomorphism, and 11% for complexity. These results revealed that the strength of the effect of terminological differences varies depending on the respective properties.

In summary, in response to Research Question 1, terminological differences do affect people's perceptions regarding the properties of ADM systems. This seems to be true for all the variables we captured in Study 1. We found the strongest effect of terminology on familiarity and tangibility, and less strong but still significant effects for machine competence, anthropomorphism, controllability and complexity.

4.1.2 Evaluations regarding the ability to conduct different tasks. Research Question 2 asked whether varying the term to refer to ADM systems affects people's evaluation regarding the performance of systems in various tasks. To investigate this research question, we used a linear mixed model, with a random-effects term for participants. We have included this random-effects term because every participant provided their evaluation of all thirteen tasks, thus the evaluation of tasks was nested within participants. Since all participants evaluated the performance of ADM systems in comparison to humans for all thirteen tasks, we added the tasks as dummy-coded independent variables into the model to investigate within-participant differences in reactions regarding whether humans or

[7] This was done because in a pilot study, the term artificial intelligence was associated with the highest machine competence and with the comparably highest potential to perform well on the thirteen tasks we included in Study 1.

[8] Our experimental design allows us to reason about the causal effects of terminology on the dependent variables. However, this is not the case for the control variable affinity for technology. Since this control variable does not reflect an experimental condition that participants were randomly assigned to but instead a characteristic of participants, we can only make claims about the correlation between affinity for technology and the dependent variables.

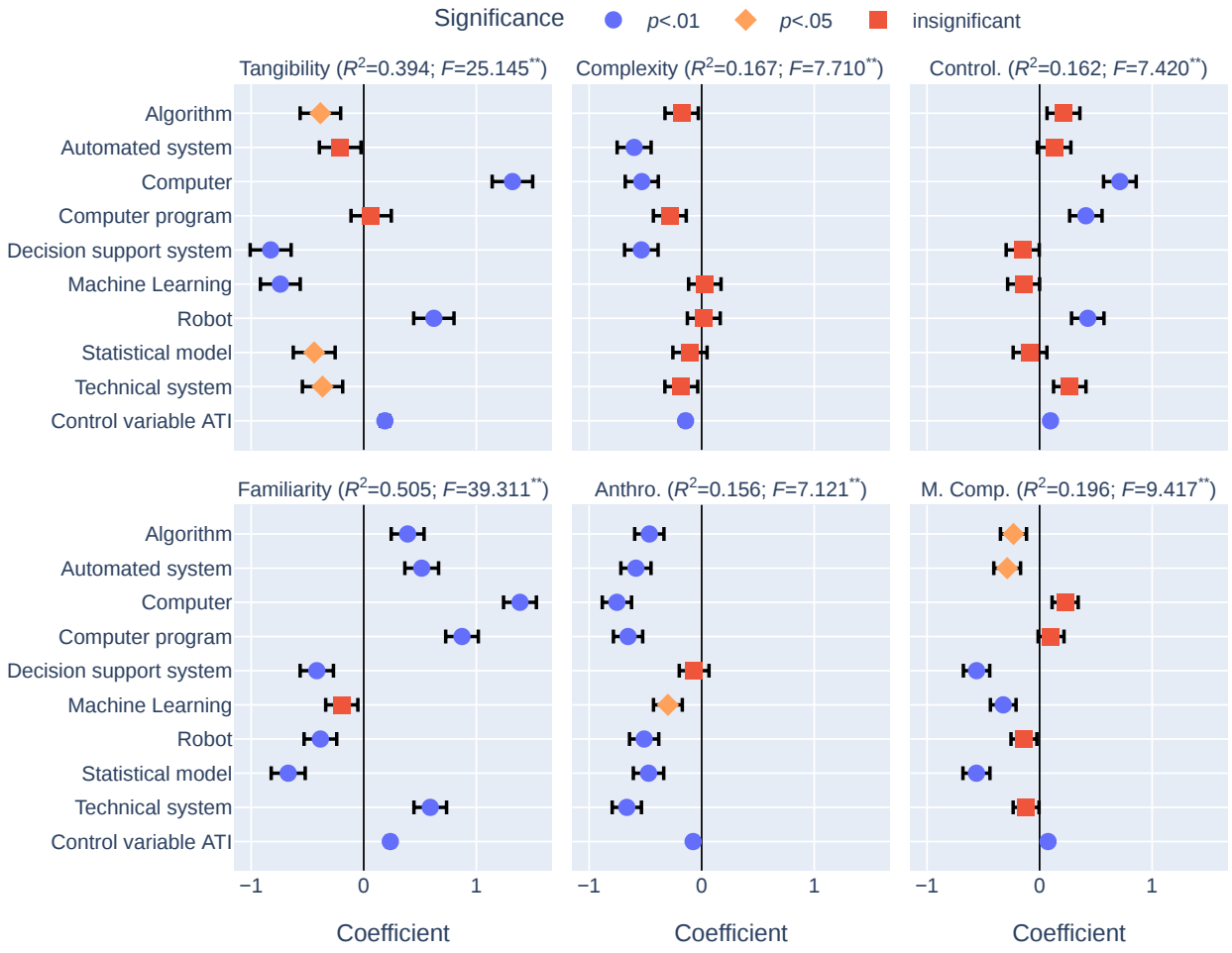

Figure 3: Linear regression coefficient plots for the perceptions of properties of ADM systems depending on the different terms. Dependent variables: Tangibility, Complexity, Controllability (Control.), Familiarity, Anthropomorphism (Anthro.), Machine Competence (M. Comp.). Independent variables: different terms used for ADM systems, and Affinity for technology (ATI) as control variable. The points show the estimated coefficients and respective standard errors. The effects for the terms can be interpreted in comparison to the reference group artificial intelligence (e.g., in the graph for Tangibility, all terms for which the coefficients are displayed on the right side of the black Zero-line received higher ratings for Tangibility than the term artificial intelligence, all left of the line received lower ratings). R^2 and F values were calculated for the respective full model. Mean values and standard deviations for the results can be found in Appendix Table 5. The intercept of the regression in the figure was omitted for readability purposes, and can be found in Appendix Table 6 that shows the results of the regressions in table format. $p < .01$. $N = 397$.**

systems are better able to conduct a respective task. We used "identify faces" as the reference task. [9] Furthermore, we included the ten terms dummy-coded with the term artificial intelligence as the reference group. We finally also added the mean-centered version of affinity for technology as control variable. The final set of variables in this model thus included the twelve dummy-coded variables for the different tasks, the nine dummy-coded variables for the different terms, and affinity for technology (for which we found that participants with a higher level of affinity for technology evaluated the performance of ADM systems more positively, see 2) all to predict the evaluation of the performance of ADM systems in comparison to humans (i.e. the dependent variable "Better than human").

Table 2 displays the results of the linear mixed model. There were no significant differences between the term artificial intelligence and the other terms. Consequently, in response to Research Question 2, varying the term did not significantly affect participant's evaluation of whether humans or the respective ADM system is

[9] We did this because a pilot study indicated that people associate the comparably highest performance with ADM systems that identify faces.

Table 2: Results of the linear mixed model with a participant random-effects term, for the comparison of whether humans or systems are better able to conduct a respective task depending on the different tasks and the different terminology.

	Better than human *Estimates* (*SE*)
Constant	3.671** (0.104)
Within-participant effects	
Predict weather	0.108 (0.063)
Make work schedules	−0.005 (0.063)
Predict the spread of infectious diseases	−0.060 (0.063)
Assess dangerous situations while driving	−0.428** (0.063)
Evaluate X-rays and MRIs	−0.542** (0.063)
Shopping recommendations	−0.657** (0.063)
Evaluate applicant documents	−1.010** (0.063)
Make medical diagnoses	−1.076** (0.063)
Make recidivism predictions	−1.121** (0.063)
Therapy recommendations in medicine	−1.214** (0.063)
Evaluate job interviews	−1.607** (0.063)
Diagnose mental illnesses	−1.728** (0.063)
Between-participants effects	
Algorithm	0.039 (0.135)
Automated system	0.002 (0.139)
Computer	0.044 (0.135)
Computer program	0.011 (0.134)
Decision support system	−0.140 (0.137)
Machine learning	0.023 (0.133)
Robot	−0.128 (0.135)
Statistical model	−0.007 (0.140)
Technical system	−0.039 (0.134)
Control Variable	
Affinity for technology	0.136** (0.030)

Note: Higher values for the dependent variable "Better than human" indicate that participants believed systems to perform better than humans. The results for the tasks can be interpreted in comparison to the task identify faces. The results for the terms can be interpreted in comparison to the term artificial intelligence. The column "Better than human" shows estimates and respective standard errors (SE) in brackets.
*$p < .05$, **$p < .01$. $N = 397$.

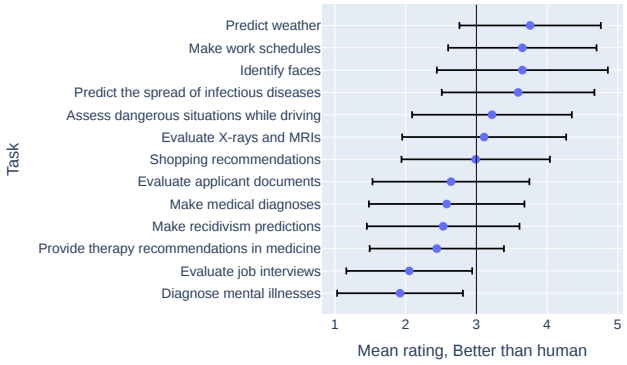

Figure 4: Rank order of participant evaluation of the comparison between humans and systems for the tasks presented in Study 1. A mean of 3 would mean that humans and systems can perform the task equally well (the black line reflects the mean of 3), above 3 means that systems are evaluated to perform better than humans, under 3 means that humans are evaluated as performing better than systems. *N* = 397.

better able to conduct various tasks. However, we need to highlight that in our case, participants were asked to explicitly compare the potential performance of an ADM system to a human being. This comparison might have reduced the possible effect of terminology because it may have affected how people think about the ADM system in question. With the explicit comparison to humans, people might reduce any term associated with ADM systems to a monolithic concept "not a human" instead of using the term to more elaborately think about the system in question. In contrast, it is conceivable that results would have differed if participants reported on how well they believe an, for instance, algorithm would be able to conduct a task on a scale from 1 (not at all) to 5 (very well).

Recent work shows emerging interest in understanding the tasks where people believe systems to perform better or at least equally well as human beings (see e.g., [8, 16, 36, 39]). Thus, the differences we found between the tasks might be of additional interest to readers. Figure 4 presents the mean values and standard deviations for the tasks ranked from most likely to be well-performed by systems to least likely. Our participants were most convinced that systems can perform better than humans for the tasks of predicting weather, identifying faces, scheduling, and predicting the spread of infectious diseases. However, people were convinced that humans perform especially better in the tasks of diagnosing mental illnesses, evaluating job interviews, and providing therapy recommendations in medicine.

4.2 Study 2 Results

Research Question 3 asked whether varying the term to refer to ADM systems affects people's evaluations of ADM systems. Furthermore, Research Question 4 asked whether the term used to refer to ADM systems and the task for which ADM systems are used interact to affect evaluations of ADM system. Figure 5 provides an overview on means and standard deviations for the dependent variables depending on the tasks and terms.

To investigate Research Questions 3 and 4, we used three separate linear regressions where we included fairness, trust, and procedural

Figure 5: Means and standard deviations for fairness, trust, and justice depending on the work evaluation and work assignment task and depending on the different terminology. The black line reflects the mean point of the respective scale and error bars reflect the standard deviation. The contents of this table can also be found in Appendix Table 7.

.

justice as dependent variable respectively. For the independent variable "task" (i.e., work evaluation vs. work assignment), we entered the tasks as a dummy-coded variable into the regression with work evaluation as the reference group. For the independent variable "terms", we included the six different terms using dummy-coded variables in the regression with the term artificial intelligence as reference group. To be able to respond to Research Question 4 that proposed a possible interaction effect between tasks and terms, we added the interaction between the tasks and the terms into the regression model. We finally added the mean-centered version of affinity for technology as control variable. The final set of variables in this model thus included a single dummy-coded variable for the tasks, five dummy-coded variables for the terms, five variables for the interaction between tasks and terms, and the control variable affinity for technology (which was positively associated with perceived justice).

Figure 6 displays the results for the regressions. Results showed that trust, fairness, and procedural justice were all stronger for ADM systems conducting the task work assignment compared to work evaluation. This replicates Lee' [39] results. Furthermore, the terms algorithm and sophisticated statistical model led to, overall, better fairness evaluations compared to the term artificial intelligence. Additionally, the terms algorithm, robot and sophisticated statistical model led to, overall, higher trust compared to the term artificial intelligence. However, terminological differences did not affect procedural justice evaluations. One interpretation for this finding could be that we used a multiple-item measure for the assessment of justice [11] whereas for fairness and trust evaluations we used the single-item assessments of these constructs also used in the original study by Lee [39]. For multiple-item measures, the influence of terminological differences might be weaker because when building a scale-mean over multiple items, terminological differences may average out. For instance, maybe the term algorithm in comparison to artificial intelligence leads to a higher evaluation for item one, but a lower evaluation for items two and three. Such a potential effect where the impact of terminology averages out over multiple items is of course not possible for single-item measures. In sum, in response to Research Question 3, varying the term to refer to an ADM system can affect evaluations of ADM systems. In our case, it affected fairness and trust but not procedural justice evaluations. Additionally, it might be that effects of terminological differences depend on the operationalization of the dependent variables.

In response to Research Question 4, Figure 6 additionally reveals that there were significant interactions for the term sophisticated statistical model and the tasks for fairness and trust. These interaction effects reflect the finding that the term sophisticated statistical model led to the most favorable fairness and trust evaluations for the task work evaluation but to the least favorable fairness and trust evaluations for the task work assignment. Apparently, terminological differences may not only affect fairness and trust evaluations but may also affect such evaluations differently for different tasks. Therefore, in response to Research Question 4, the task and the term may interact to differently affect people's evaluations of ADM systems – whereas in one task a term may be associated with comparably positive evaluations, this may not hold for another task.

Finally, in a direct replication of Lee [39], we investigated the difference between the human manager and ADM systems depending on the tasks as well as on the different terms to refer to ADM systems. Lee found that for work evaluation, her participants evaluated algorithms to be less fair and reported lower trust in algorithms than in human managers. In contrast, the differences she found between human managers and algorithms were much smaller, and non-significant, for work assignment tasks. In order to increase the interpretability of our results in comparison to Lee's results, we followed her example and provide analyses for the tasks separately in Figure 7. For these linear regressions, we entered six dummy-coded variables with the reference group human manager. We again added affinity for technology as control variable. The final set of variables in our models thus included six dummy-coded variables that reflect the comparison between human manager and the different terms, as well as the control variable affinity for technology (for affinity for technology see last row of each graph in Figure 7; affinity for technology was positively associated with fairness, trust and justice in the work evaluation task and negatively associated with fairness in the work assignment task). For transparency purposes, we also provide results for procedural justice but will not discuss them as Lee did not measure justice in her study.

Our results for the work evaluation task (Figure 7 top) showed that, although the terms differed in the extent to which they were evaluated as less fair and in the extent to which they evoked less trust (e.g., for the term sophisticated statistical model, there was a smaller difference to the human manager compared to the term artificial intelligence), the differences between human manager and all the terms were significant. This replicates the findings from Lee.

Yet, results for work assignment (Figure 7 bottom) only partly replicated Lee's findings. Specifically, our results replicated her findings with respect to fairness, where we also found no significant differences between the human manager and the ADM system for any of the terms. However, our results regarding trust replicated Lee's findings only for certain terminology. Specifically, there were

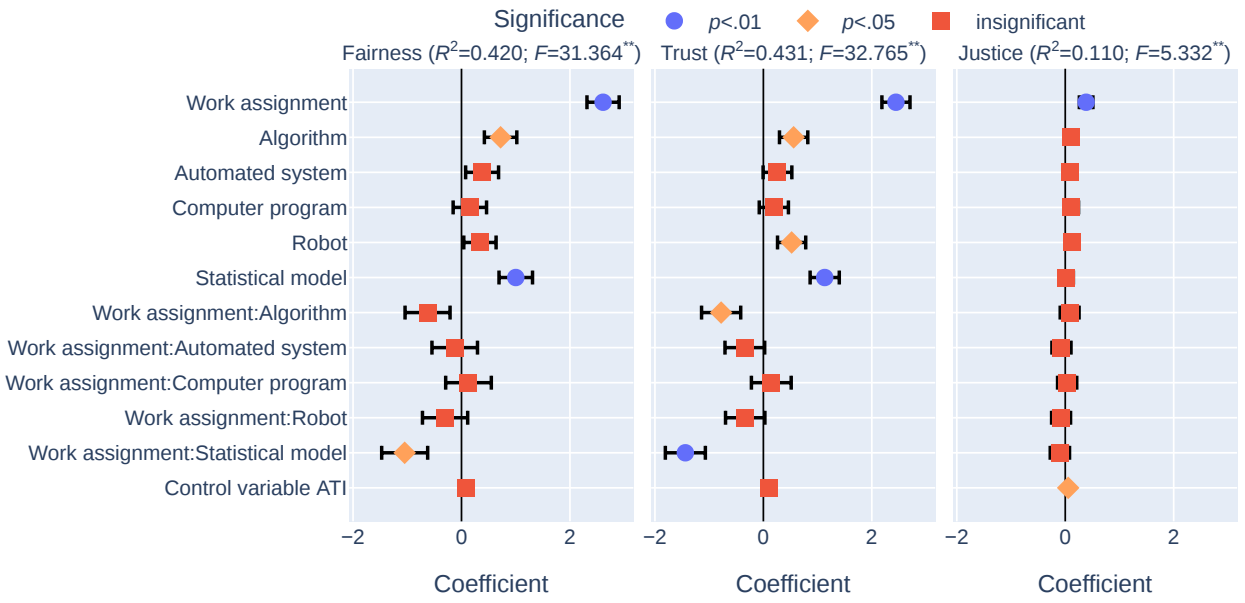

Figure 6: Linear regression coefficient plots for the evaluation of Fairness, Trust, and Procedural Justice depending on the different terms and tasks, including interaction effects. Dependent variables: Fairness, Trust, and Justice. Independent variables: term (six different terms for ADM systems), task (work evaluation or work assignment), and Affinity for technology (ATI) as control variable. The points show the estimated coefficients and respective standard errors. The effects for the terms can be interpreted in comparison to the reference group artificial intelligence, the effect for the tasks can be interpreted in comparison to the reference group work evaluation. R^2 and F values were calculated for the respective full model. The intercept is omitted for readability purposes, and can be found in Appendix Table 8. $^{**}p < .01$. $N = 533$.

no significant differences regarding trust for the terms computer program or robot which replicates Lee's findings. In contrast, for the terms artificial intelligence, algorithm, automated system, and sophisticated statistical model, we found significantly lower trust evaluations compared to the human manager. In other words, if we would have chosen one of the former terms, we would have found no statistically significant results, thus supporting Lee's findings. Instead, if we would have tried to replicate her study with one of the latter terms, we would have found significant differences between humans and ADM systems for trust regarding conducting work assignment tasks and would have concluded that we could not replicate Lee's findings. Consequently, the choice for or against one of the terms could have crucially affected whether our study would have supported or contradicted Lee's results. Similarly, if Lee would have chosen a different term for her study, she might have found different results and might have drawn different conclusions.

5 DISCUSSION

The goal of this paper was to investigate whether terminological differences affect human perceptions and evaluations of ADM systems. The main results are that, indeed, the terminology used to describe ADM systems affects people's perceptions of the properties of those systems. Furthermore, although terminological differences did not affect evaluations of the ability of ADM systems to conduct different tasks in comparison to humans, it did affect people's evaluations of system fairness and trust in systems. These effects might depend on the ways we measure perceptions and evaluations of ADM systems – in our case we hypothesize that it may depend on whether there was an explicit comparison to human task performance, and on whether we used single-item versus multiple-item measures. However, further research is necessary to evaluate these hypotheses. Overall, we conclude that terminology matters when describing ADM systems to participants in research studies because it can affect the robustness and replicability of research results, and terminology matters because it may shape perceptions and evaluations of ADM systems in communication about such systems (e.g., in public discourse and policy-making). Consequently, it is necessary to be aware that choosing the terms to describe ADM systems can have unintended consequences (e.g., varying research findings due to varying terminology) but that terminology can also be used strategically (e.g., referring to a system as artificial intelligence to make it sound complex and novel).

5.1 Terminology affects human perceptions and evaluations of ADM systems

One of the main implications of our study is that it is necessary to be mindful regarding what term to use when describing ADM systems to research participants because findings may vary due to using different terminology. Our Study 2 supports that this might have been

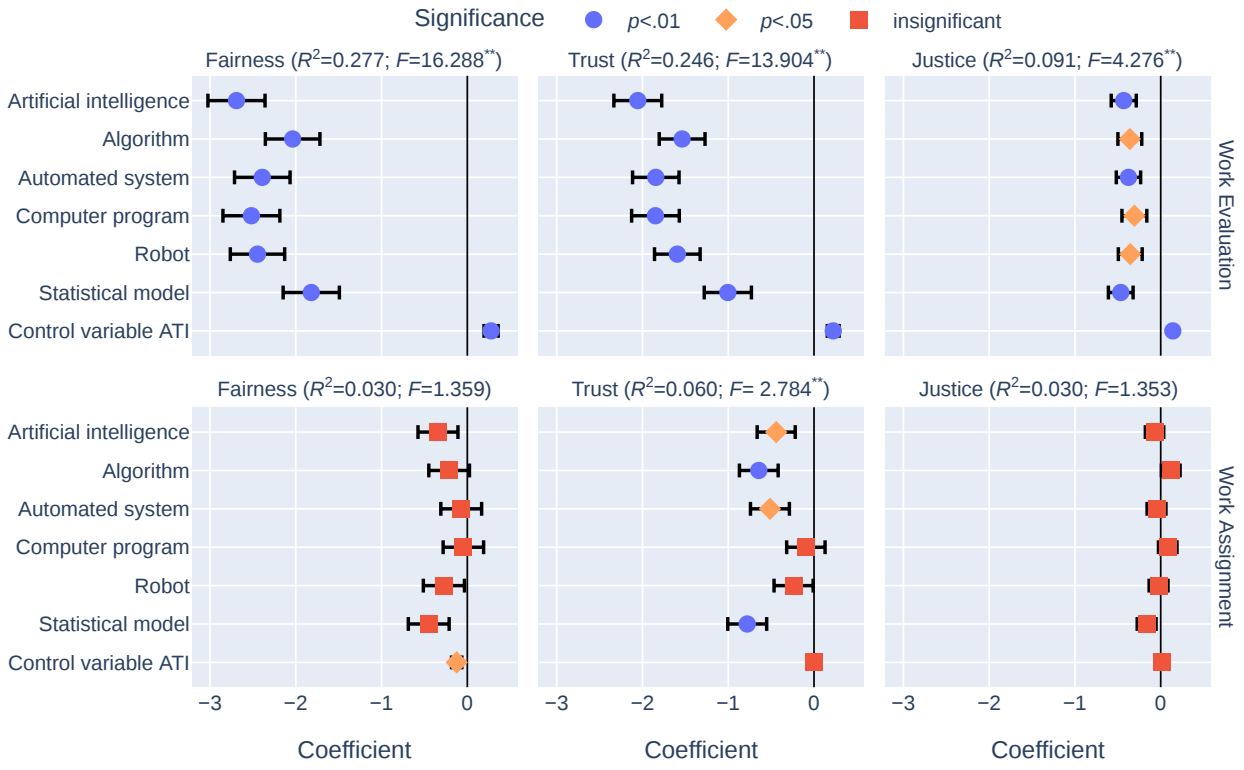

Figure 7: Linear regression coefficient plots for the evaluation of Fairness, Trust, and Procedural Justice for the tasks work evaluation (top) and assignment (bottom) depending on the different terms. Dependent variables: Fairness, Trust, and Justice. Independent variables: human manager vs. different terms for ADM systems, and Affinity for technology (ATI) as control variable. The points show the estimated coefficients and respective standard errors. The effects for the terms can be interpreted in comparison to the human manager as a reference group (e.g., in the graph for Fairness, all terms for which the coefficients are displayed on the left side of the black zero-line received lower ratings for Fairness than the human manager). R^2 and F values were calculated for the respective full model. The intercept is omitted for readability purposes, and can be found in Appendix Tables 9 and 10.

$^{**}p < .01$. Work evaluation $n = 306$, Work assignment $n = 316$.

Figure 1: Caption.

an issue in previous HCI research and thus is in line with Langer and Landers' [36] conclusion that terminological differences may have led to different conclusions for studies that examined similar research questions. For instance, whereas Lee [39] used the term algorithm to describe a system in a hiring scenario, Marcinkowski et al. [45] used the term AI technology for a similar task. Whereas Lee and Marcinkowski et al. may have had a similar idea as well as a similar technology in mind – a system that automatically evaluates applicant information and recommends rejection or approval of applicants – Lee found that her participants preferred the human manager over the algorithm in hiring, whereas Marcinkowski et al. reported that their participants preferred the AI technology over a human. Part of these differences in findings may be due to the varying terminology (see also [30]). Unfortunately, we cannot conclude that there is a simple main effect of different terminology where one term will always lead to more favorable evaluations than another. More precisely, our Study 2 showed that algorithms were perceived as more favorably than artificial intelligence to conduct work evaluations, whereas in the comparison of Lee and Marcinkowski et al.'s results, the term AI technology was associated with more favorable evaluations of ADM systems than the term algorithm. This suggests that terminological differences may differentially affect the evaluation of ADM systems for various tasks.

Our Study 2 further supports this interpretation because we found that the effect of the terminology depended on the task for which a system is used. Given that systems were perceived more negatively for work evaluation tasks, a preliminary interpretation of this finding might be that we can expect stronger effects of terminology in contexts where people are less positive about the use of systems. This could be the case because in tasks where people already have positive views about systems, they might already expect that ADM systems conduct respective tasks. However, in tasks where it is more controversial whether and to what extent we can and should use ADM systems (e.g., in work evaluation, diagnosis of mental illnesses; [36]), people's expectations will be

violated by the fact that "not a human" is conducting the respective task. In cases where expectations are violated, people might be more critical, may think more intensely about the use of systems in those situations, and may consequently scrutinize available information in order to determine how positive or negative they find the idea of an ADM system conducting a task [61]. In Study 2, this applies to the work evaluation task that was also found to be less positively evaluated by Lee's [39] participants and where our participants may have more intensely thought about what the respective term would tell them about the system conducting this task. In contrast, for work assignment our participants may have been less surprised by the fact that an ADM system conducts the task which led to less elaboration about the term used to describe the system. The mean values and standard deviations found for the single tasks and terms in Study 2 may support this interpretation (see Figure 5 and Appendix Table 7). Specifically, mean values for fairness in the work evaluation task showed stronger variance and ranged between 3.00 and 4.05, whereas those for the work assignment task ranged between 5.56 and 5.91. The same was found for trust, where the range for work evaluation was between 2.48 and 3.67, whereas for work assignment it was between 4.60 and 5.28. Also, the standard deviations for the evaluation of trust and fairness were almost consistently higher for the work evaluation task. This means that there was more variation between people in how (un)favorably they evaluated the terms for the task work evaluation compared to the task work assignment. Nevertheless, readers should be aware that this is a tentative interpretation of our findings. Shedding further light on the conditions that affect the strength of the influence of different terminology will be a task for future research.

Importantly, we do not claim that terminology is the only factor that contributes to variation in findings and conclusions between studies or that terminology is an especially strong determinant of research findings. In fact, there are many other important choices in studies that will have a larger effect on participants. For instance, both our studies showed that choices regarding the operationalization of constructs (e.g., single-item versus multiple-item measures; explicit comparison to human performance) can influence results. Moreover, both our studies support prior work suggesting that the task for which an ADM system is used more strongly affects participants' evaluations of ADM systems [8, 16, 36, 39, 43] than terminological differences. For instance, the rank order we found in Study 1 and also the large differences between the tasks in Study 2 support previous work where authors suspected that in high-stakes tasks (e.g., diagnosing mental illnesses), people will find humans to be better suited to perform these tasks than ADM systems [35, 36]. Yet, high-stakes versus low-stakes is clearly not the only dimension that explains differences in evaluations of ADM systems in these tasks. For example, predicting the spread of infectious diseases might also be considered a high-stakes task (especially given that our data collection was conducted during the COVID-19 pandemic) and our participants evaluated that ADM systems would be better able to perform this task than humans. Other dimensions might be whether the task requires human versus mechanical skills [39], the inherent uncertainty associated with the decision-making task at hand [16], and the complexity of the task [47] (as has been argued by Langer and Landers [36]).

Moreover, providing specific information on characteristics of ADM systems may reduce effects that terminology may have. Specifically, terminology effects might stem from what participants have in mind when thinking about an ADM system described with a specific terminology. As we described in Section 2.1, a large share of previous work only used the respective term to inform participants that an ADM system will make or support decisions without further information regarding different nuances of underlying technology or regarding how well a system works for a specific task. With this kind of ambiguity, terminology effects may be especially strong. However, providing information on, for example, training and validation of an ADM system or describing that during validation the respective system has been found to make accurate predictions in 95% of cases may attenuate terminology effects because participants are not left wondering how a system was developed or how well a system will work. In cases where there are more specific descriptions of system characteristics, it will be a task for future research to investigate whether it matters less that a system is described with different terminology.

5.2 Being mindful about terminology may enhance robustness and replicability of research

Even if the effects of terminology may depend on other methodological choices (i.e., the choice of operationalization of constructs), are comparably weaker than the effects of other considerations (e.g., the task performed by ADM systems), or are attenuated under certain circumstances (e.g., when adding specific information about characteristics of ADM systems) our results showed that different terminology is associated with variation in people's perceptions and evaluations of ADM systems. Given that the terminology to describe ADM systems to participants is easily controllable within studies, we suggest that researchers

- mindfully consider what term to use to describe ADM systems to their participants
- clearly report in the methodology of their papers what term they used

Following these suggestions may help increase the robustness and replicability of research findings. More precisely, when designing a study where participants are informed that an ADM system decides about the future of people or in studies where people interact with an ADM system to perform a task, it makes sense to screen previous literature to examine what terms other authors have used to describe the respective systems. If the goal of the research is to replicate or advance specific previous studies, it makes sense to use similar terms like the respective studies since it would at least control for unintended variation due to different terminology. Unfortunately, Langer and Landers' [36] literature review showed that there is a large variety of terms that have been used in previous studies. To date, there is limited information regarding how strongly varying terminology has affected findings and conclusions of previous work. We hope that our studies raise awareness of the effects different terminology can have on research findings and hope it will motivate future research to more actively consider what term to use when describing ADM systems to participants.

5.3 Terminology can be used strategically in communication about ADM systems

Our studies imply that when people communicate with someone about ADM systems describing this system with different terminology can impact listeners' perceptions and evaluations associated with the respective system (for similar results across disciplines see [19, 55, 58]). This supports that terminology may lead to different reactions in communication about ADM systems. An implication that needs further exploration is whether different perceptions and evaluations of ADM systems lead to different behavior in the interaction with ADM systems. For instance, if people are more likely to trust a statistical model compared to an artificial intelligence to conduct work assignment, they may also be more likely to actually use and rely on a system described as being a sophisticated statistical model. Similarly, if people associate higher machine competence with artificial intelligence compared to automated systems, they may more likely use outputs generated by a system that is described as an artificial intelligence. It is important to highlight that these are hypotheses we derived from our studies because we did not measure behavioral outcomes associated with using different terminology. Nevertheless, Study 2 showed that people's fairness and trust evaluations depended on the term used to describe the system and fairness as well as trust have been found to be antecedents of actual system use [29, 31, 37].

Overall, our studies also suggest that for communication about ADM systems in journalist reports, public discourse, and policy-making it is necessary to be aware that the choice of a certain term has effects. Terminology may affect how people receive the respective communication, how they evaluate the use of ADM systems for various tasks, and may influence what people do as consequence of respective communications. For example, if journalists write about artificial intelligence [68] versus algorithms in recruitment [13], this might lead to different evaluations of the general idea of ADM systems to support recruitment. If the term artificial intelligence would lead to a less favorable evaluation of using ADM systems to evaluate people and make decisions over people's careers (as found in previous work; [36, 39, 48]), this can lead to stronger public outcry and potentially even protests than the idea of algorithms doing the same thing. As another example, if public discourse on healthcare supported by ADM systems would use the term automated system instead of the term artificial intelligence to start respective discussions, this may lead to less engagement and controversy in the discussion because people may already know that automated systems are used in healthcare, thus less likely violating people's expectations.

In conclusion, our study supports that in communication practice, the choice for or against a term can be a strategic one. Take the example of policy-making documents where the authors may have the choice to use the term artificial intelligence compared to more familiar terms such as computer programs. Maybe if the European Commission's "Ethics Guidelines for Trustworthy AI" [50] had been called "Ethics Guidelines for Trustworthy Computer Programs" there would have been less public outreach. In other words, terminology could be used intentionally to engage people to contribute to discussions. Furthermore, terminology can also be used as a selling argument for companies who use ADM systems. There are reports that many European companies claim to use artificial intelligence in their products but actually never did so [65]. Our results showed that it might be the complexity as well as strong potential that people associate with artificial intelligence that may underlie this choice of terminology compared to equally plausible terminology. Consequently, in comparison to sometimes unintended consequences of using different terminology (e.g., variation in research findings), terminology can clearly also been used strategically in order to cause desired effects (e.g., engagement, interest; [19, 54]).

5.4 Limitations

There are four main limitations to our work. First, all captured data relied on self-reported information from participants. Although this led us to conclude that terminological differences affect perceptions and evaluations of ADM systems, we can only draw tentative conclusions with respect to behavioral consequences – consequences that have been found in other fields investigating terminological differences [19]. For instance, given that the term artificial intelligence was associated with comparably high machine competence, we imagine future studies where participants interact with a real system providing them with recommendations and where researchers capture to what extent participants rely on the recommendations by the system depending on whether the system is described as artificial intelligence or as a sophisticated statistical model [17]. Second, we gathered non-representative samples on Prolific so we cannot generalize our findings to broader populations. However, since many studies in HCI have been and still are conducted via crowdsourcing platforms such as Prolific or MTurk (e.g., [6, 67]), we are optimistic that our interpretation that terminology may have affected HCI research findings holds. Third, although we included a measure of participants' affinity for technology, participants' reactions to the terms may have been affected by more specific AI or computer science-related experience. However, by randomly assigning participants to the experimental groups, we would expect that there was no group where there was a significantly larger number of participants with a computer science background. Therefore, we believe that the interpretations drawn from our study remain valid. Nevertheless, future research could examine whether people with a strong computer scientific background would react less strongly to different terminology because they know that some of these terms can be used interchangeably, or whether they react more strongly because they are more aware of the nuances that distinguish the terms (e.g., with respect to underlying technology). Fourth, our studies were conducted in English so results may have been different if we had conducted our studies in other languages. This limitation may inspire future work that investigates whether terminology has different effects in languages other than English. For example, for scholars who conduct their research in the native languages of their participants it might be interesting to investigate whether terminology has effects similar to what we found in our studies. We can imagine that in some languages the respective terminologies (e.g., robots and artificial intelligence) might be more closely related or might be perceived as more different than in the English language. Furthermore, in some languages respective terms

may be more common compared to other languages where terms may reflect something more specific, potentially resulting in differences regarding perceptions of familiarity or complexity. Such nuances may lead to different effects of terminology in different languages.

5.5 Conclusion

When communicating, there are many terms that can be used to express similar ideas. This also applies to the terms to refer to ADM systems. Different terminology can strongly affect people's thoughts, feelings, and behavior [19, 55]. From a research point of view, our studies showed that it is necessary to be mindful when describing ADM system, especially when trying to better understand how people perceive and evaluate ADM systems, when investigating what people expect from such systems in application contexts, and when examining how people interact with systems in everyday life. From a practical point of view, our studies imply that in communicating about ADM systems in public discourse, the media, and policy-making, there might be strategic choices for different terminology because terminology may have the potential to engage people, make them more interested in a topic, or may lead to positive/negative evaluations of the use of ADM systems. In summary, our studies show that terminology needs to be chosen wisely as it can affect what kind of properties people ascribe to ADM systems, can influence people's perceptions of systems in application contexts, and can affect the robustness and replicability of research findings.

ACKNOWLEDGMENTS

Work on this paper and on the studies included in this paper was funded by the Volkswagen Foundation grant Az. 98513 "Explainable Intelligent Systems" (EIS), by the Deutsche Forschungsgemeinschaft (DFG) grant 389792660 as part of TRR248, and by the Bundesministerium für Bildung und Forschung (BMBF) Project "Ophthalmo-AI" grant 16SV8640.

REFERENCES

[1] Herman Aguinis and Kyle J Bradley. 2014. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods* 17, 4 (2014), 351–371. https://doi.org/10.1177/1094428114547952

[2] Kirk Allen, Teri Reed-Rhoads, Robert A Terry, Teri J Murphy, and Andrea D Stone. 2008. Coefficient alpha: An engineer's interpretation of test reliability. *Journal of Engineering Education* 97, 1 (2008), 87–94. https://doi.org/10.1002/j.2168-9830.2008.tb00956.x

[3] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189. https://doi.org/10.1016/j.chb.2018.03.051

[4] Christiane Atzmüller and Peter M Steiner. 2010. Experimental vignette studies in survey research. *Methodology* 6 (2010), 128–138. https://doi.org/10.1027/1614-2241/a000014

[5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[6] 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, 1–14. https://doi.org/10.31235/osf.io/9wqxr

[7] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 632–642. https://doi.org/10.18653/v1/d15-1075

[8] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825. https://doi.org/10.1177/0022243719851788

[9] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semantic textual similarity-multilingual and cross-lingual focused evaluation. *Proceedings of the 2017 SEMVAL International Workshop on Semantic Evaluation* (2017). https://doi.org/10.18653/v1/s17-2001

[10] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv:1803.11175* (2018).

[11] Jason A Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology* 86, 3 (2001), 386–400. https://doi.org/10.1037/0021-9010.86.3.386

[12] Kimberly E Culley and Poornima Madhavan. 2013. A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior* 29, 3 (2013), 577–579. https://doi.org/10.1016/j.chb.2012.11.023

[13] Oren Danieli, Andrew Hillis, and Michael Luca. 2016. *How to hire with algorithms*. Retrieved July, 17, 2021 from https://hbr.org/2016/10/how-to-hire-with-algorithms

[14] Maartje Ma De Graaf, Somaya Ben Allouch, and Tineke Klamer. 2015. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in Human Behavior* 43 (2015), 1–14. https://doi.org/10.1016/j.chb.2014.10.030

[15] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331–349. https://doi.org/10.1037/xap0000092

[16] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science* 31, 10 (2020), 1302–1314. https://doi.org/10.1177/0956797620948841

[17] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. https://doi.org/10.2139/ssrn.2466040

[18] Rob Eisinga, Manfred Te Grotenhuis, and Ben Pelzer. 2013. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health* 58, 4 (2013), 637–642. https://doi.org/10.1007/s00038-012-0416-3

[19] Melissa V Eitzel, Jessica L Cappadonna, Chris Santos-Lang, Ruth Ellen Duerr, Arika Virapongse, Sarah Elizabeth West, Christopher Kyba, Anne Bowser, Caren Beth Cooper, Andrea Sforzi, et al. 2017. Citizen science terminology matters: Exploring key terms. *Citizen Science: Theory and Practice* 2, 1 (2017), 1–20. https://doi.org/10.5334/cstp.96

[20] Kimberly D Elsbach and Ileana Stigliani. 2019. New information technology and implicit bias. *Academy of Management Perspectives* 33, 2 (2019), 185–206. https://doi.org/10.5465/amp.2017.0079

[21] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (2007), 864–886. https://doi.org/10.1037/0033-295x.114.4.864

[22] Doug Ertz. 2021. *Eight in ten leaders want intelligent systems success in five years but the time to start blueprinting is now.* Retrieved July, 17, 2021 from https://www.forbes.com/sites/windriver/2021/05/01/eight-in-ten-leaders-want-intelligent-systems-success-in-five-years-but-the-time-to-start-blueprinting-is-now/

[23] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150

[24] Murray Gell-Mann. 2002. What is complexity? In *Complexity and industrial clusters*, Alberto Quadrio Curzio and Marco Fortis (Eds.). Springer, Heidelberg, Germany, 13–24. https://doi.org/10.1007/978-3-642-50007-7_2

[25] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660. https://doi.org/10.5465/annals.2018.0057

[26] Manuel F Gonzalez, John F Capman, Frederick L Oswald, Evan R Theys, and David L Tomczak. 2019. "Where's the IO?" Artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions* 5, 3 (2019), 5. https://doi.org/10.25035/pad.2019.03.005

[27] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25. https://doi.org/10.2139/ssrn.3465622

[28] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 WWW World Wide*

Web Conference. ACM, 903–912. https://doi.org/10.1145/3178876.3186138

[29] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[30] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25. https://doi.org/10.1145/3479864

[31] Frederick M Howard, Catherine A Gao, and Christopher Sankey. 2020. Implementation of an automated scheduling tool improves schedule quality and resident satisfaction. *PloS One* 15, 8 (2020), e0236952. https://doi.org/10.1371/journal.pone.0236952

[32] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 FAccT Conference on Fairness, Accountability, and Transparency*. ACM, 624–635. https://doi.org/10.1145/3442188.3445923

[33] Stuart Keel, Pei Ying Lee, Jane Scheetz, Zhixi Li, Mark A Kotowicz, Richard J MacIsaac, and Mingguang He. 2018. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: A pilot study. *Scientific Reports* 8, 1 (2018). https://doi.org/10.1038/s41598-018-22612-2

[34] Markus Langer, Cornelius J König, and Victoria Hemsing. 2020. Is anybody listening? The impact of automatically evaluated job interviews on impression management and applicant reactions. *Journal of Managerial Psychology* 35, 4 (2020), 271–284. https://doi.org/10.1108/jmp-03-2019-0156

[35] Markus Langer, Cornelius J König, and Maria Papathanasiou. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment* 27, 3 (2019), 217–234. https://doi.org/10.1111/ijsa.12246

[36] Markus Langer and Richard N Landers. 2021. The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior* (2021), 106878. https://doi.org/10.1016/j.chb.2021.106878

[37] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

[38] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies* 64, 10 (2006), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002

[39] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. https://doi.org/10.1177/2053951718756684

[40] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. https://doi.org/10.1145/3411764.3445570

[41] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

[42] LiveCareer. 2018. *7 Things to Know about the role of robots in recruitment.* Retrieved July, 17, 2021 from https://careerenlightenment.com/7-things-to-know-about-the-role-of-robots-in-recruitment

[43] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650. https://doi.org/10.1093/jcr/ucz013

[44] Niklas Luhmann. 2000. Familiarity, confidence, trust: Problems and alternatives. In *Trust: Making and breaking cooperative relations*, Diego Gambetta (Ed.). Department of Sociology, University of Oxford, Oxford, UK, 94–107.

[45] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. Implications of AI (un-) fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 FAT* Conference on Fairness, Accountability, and Transparency*. ACM, 122–130. https://doi.org/10.1145/3351095.3372867

[46] Enid NH Montague, Brian M Kleiner, and Woodrow W Winchester III. 2009. Empirically understanding trust in medical technology. *International Journal of Industrial Ergonomics* 39, 4 (2009), 628–634. https://doi.org/10.1016/j.ergon.2009.01.004

[47] Rosanna Nagtegaal. 2021. The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly* 38, 1 (2021), 101536. https://doi.org/10.1016/j.giq.2020.101536

[48] David T Newman, Nathanael J Fast, and Derek J Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (2020), 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

[49] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Altered attitudes of people toward robots: Investigation through the Negative Attitudes toward Robots Scale. In *Proceedings of the 2006 AAAI workshop on Human Implications of Human-Robot Interaction*. 29–35.

[50] High-Level Expert Group on Artificial Intelligence. 2021. *Ethics guidelines for trustworthy AI.* Retrieved July, 17, 2021 from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[51] Sonja K Ötting and Günter W Maier. 2018. The importance of procedural justice in human–machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior* 89 (2018), 27–39. https://doi.org/10.1016/j.chb.2018.07.022

[52] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2020. Human–Autonomy teaming: A review and analysis of the empirical literature. *Human Factors* (2020), 0018720820960865. https://doi.org/10.1177/0018720820960865

[53] Robert A Peterson. 1994. A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research* 21, 2 (1994), 381–391.

[54] R Puhl, JL Peterson, and J Luedicke. 2013. Motivating or stigmatizing? Public perceptions of weight-related language used by health providers. *International Journal of Obesity* 37, 4 (2013), 612–619. https://doi.org/10.1038/ijo.2012.110

[55] Rebecca M Puhl. 2020. What words should we use to talk about weight? A systematic review of quantitative and qualitative studies examining preferences for weight-related terminology. *Obesity Reviews* 21, 6 (2020), e13008. https://doi.org/10.1111/obr.13008

[56] Victoria A Shaffer, C Adam Probst, Edgar C Merkle, Hal R Arkes, and Mitchell A Medow. 2013. Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making* 33, 1 (2013), 108–118. https://doi.org/10.1177/0272989x12453501

[57] Daniel B Shank, Madison Bowen, Alexander Burns, and Matthew Dew. 2021. Humans are perceived as better, but weaker, than artificial intelligence: A comparison of affective impressions of humans, AIs, and computer systems in roles on teams. *Computers in Human Behavior Reports* 3 (2021), 100092. https://doi.org/10.1016/j.chbr.2021.100092

[58] Daniel B Shank, Alexander Burns, Sophia Rodriguez, and Madison Bowen. 2020. Software program, bot, or artificial intelligence? Affective sentiments across general technology labels. *Current Research in Social Psychology* (2020). https://crisp.org.uiowa.edu/sites/crisp.org.uiowa.edu/files/2020-06/crisp_28_4_shank.pdf

[59] Daniel B Shank and Alyssa DeSanti. 2018. Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior* 86 (2018), 401–411. https://doi.org/10.1016/j.chb.2018.05.014

[60] Rania Shibl, Meredith Lawley, and Justin Debuse. 2013. Factors influencing decision support system acceptance. *Decision Support Systems* 54, 2 (2013), 953–961. https://doi.org/10.1016/j.dss.2012.09.018

[61] Stephen M Smith and Richard E Petty. 1996. Message framing and persuasion: A message processing analysis. *Personality and Social Psychology Bulletin* 22, 3 (1996), 257–268. https://doi.org/10.1177/0146167296223004

[62] Alarith Uhde, Nadine Schlicker, Dieter P Wallach, and Marc Hassenzahl. 2020. Fairness and decision-making in collaborative shift scheduling systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3313831.3376656

[63] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS Quarterly* 27, 3 (2003), 425–478. https://doi.org/10.2307/30036540

[64] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*. ACM, 351–362. https://doi.org/10.1145/2449396.2449442

[65] James Vincent. 2019. *Forty percent of 'AI startups' in Europe don't actually use AI, claims report.* Retrieved July, 23, 2021 from https://www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report

[66] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[67] 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*. ACM, 1–14. https://doi.org/10.1145/3313831.3376813

[68] David Windley. 2021. *Is AI the answer to recruiting effectiveness?* Retrieved July, 23, 2021 from https://www.forbes.com/sites/forbeshumanresourcescouncil/2021/06/16/is-ai-the-answer-to-recruiting-effectiveness/

A APPENDIX

Table 3: Items for Study 1

Scale	Item text	Response format
Tangibility	When I think of "the term", I have a clear picture in mind.	1 (strongly disagree) to
	When I think of "the term", it has a shape.	5 (strongly agree)
Familiarity	"The term" is something I encounter in everyday life.	1 (strongly disagree) to
	"The term" is familiar to me.	5 (strongly agree)
	"The term" is something novel. (r)	
Complexity	"The term" and how it works is easy to understand. (r)	1 (strongly disagree) to
	"The term" is understandable even for laypeople. (r)	5 (strongly agree)
	"The term" is complex.	
	I could predict the results generated by "the term". (r) (e)	
	"The term" works, even if I do not exactly understand how. (e)	
Controllability	"The term" is controllable by humans.	1 (strongly disagree) to
	"The term" and related processes can be controlled.	5 (strongly agree)
	"The term" acts independently. (r) (e)	
Anthropo-morphism	"The term" has a mind on its own.	1 (strongly disagree) to
	"The term" has intentions.	5 (strongly agree)
	"The term" has free will.	
	"The term" has beliefs.	
	"The term" has the ability to experience emotions.	
	"The term" has desires.	
	"The term" has conscientiousness.	
	"The term"'s decision making processes are similar to those of humans.	
Machine competence	"The term" has great potential in terms of what it can be used for.	1 (strongly disagree) to
	"The term" can be used flexibly for various tasks.	5 (strongly agree)
	I believe that "the term" has great capabilities.	
	"The term" can generate results just as good as human experts.	
	"The term" can adapt to changing situations.	
	"The term" and their decision outcomes are similar to that of humans.	
Tasks	"The term" can make shopping recommendations.	1 (worse than a human),
	"The term" can evaluate documents by applicants (e.g., applicant resumes).	2 (slightly worse than a human),
	"The term" can make recidivism predictions of convicted offenders.	3 (as good as a human),
	"The term" can make medical diagnoses.	4 (slightly better than a human),
	"The term" can evaluate X-ray and MRI images.	5 (better than a human)
	"The term" can predict the weather.	
	"The term" can evaluate job interviews.	
	"The term" can produce shift schedules at work.	
	"The term" can provide therapy recommendations in medicine.	
	"The term" can diagnose mental illness.	
	"The term" can identify faces.	
	"The term" can assess dangerous situations while driving.	
	"The term" can predict the spread of infectious diseases.	
Affinity for technology	I like to occupy myself in greater detail with technical systems.	1 (completely disagree) to
	I like testing the functions of new technical systems.	6 (completely agree)
	It is enough for me that a technical system works; I don't care how or why. (r)	
	It is enough for me to know the basic functions of a technical system. (r)	

Note: "The term" is used as a placeholder for the experimentally manipulated terms respectively. (r) = reverse-coded item, (e) = item was excluded from the final analysis because it led to a low Cronbach's α of the scale.

Table 4: Items for Study 2

Scale	Item text	Response format
Fairness	How fair or unfair is it for Chris that the "the term" assigns him to check a specific component of the machinery and he does the maintenance work on it? / How fair or unfair is it for Chris that the "the term" evaluates his performance?	1 (Very unfair) to 7 (Very fair)
Trust	How much do you trust that "the term" makes a good-quality work assignment? / How much do you trust that "the term" makes a good-quality work evaluation?	1 (No trust at all) to 7 (Extreme trust)
Procedural Justice	The following items refer to the procedures used to arrive at the decision. To what extent do you think:	
	Has Chris been able to express his views and feelings during those procedures?	1 (to a very small extent),
	Has Chris had influence over the decision arrived at by those procedures?	2 (to a small extent),
	Have those procedures been applied consistently?	3 (to some extent),
	Have those procedures been free of bias?	4 (to a large extent),
	Have those procedures been based on accurate information?	5 (to a very large extent)
	Has Chris been able to appeal the decision arrived at by those procedures?	
	Have those procedures upheld ethical and moral standards?	
Affinity for technology	I like to occupy myself in greater detail with technical systems.	1 (completely disagree) to
	I like testing the functions of new technical systems.	6 (completely agree)
	It is enough for me that a technical system works; I don't care how or why. (r)	
	It is enough for me to know the basic functions of a technical system. (r)	

Note: "The term" is used as a placeholder for the experimentally manipulated terms respectively. (r) = reverse-coded item.

Table 5: Means and standard deviations for perceptions regarding the properties of ADM systems for the terms in Study 1.

Condition	Tang.		Comp.		Cont.		Fam.		Anth.		M. Com.	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Artificial intelligence	2.89	0.83	3.76	0.61	3.52	0.69	2.98	0.63	2.39	2.39	4.01	0.49
Algorithm	2.41	0.63	3.66	0.65	3.69	0.78	3.26	0.68	1.97	1.97	3.74	0.54
Automated system	2.65	0.88	3.19	0.77	3.64	0.69	3.46	0.66	1.82	1.82	3.71	0.74
Computer	4.19	0.78	3.25	0.87	4.22	0.59	4.34	0.62	1.65	1.65	4.23	0.41
Computer program	2.91	0.79	3.51	0.77	3.91	0.60	3.80	0.77	1.76	1.76	4.10	0.53
DSS	2.05	0.85	3.24	0.59	3.37	0.63	2.55	0.69	2.33	2.33	3.45	0.43
Machine learning	2.15	0.83	3.79	0.59	3.38	0.73	2.79	0.69	2.09	2.09	3.69	0.54
Robot	3.46	0.82	3.82	0.67	3.93	0.69	2.54	0.74	1.90	1.90	3.85	0.46
Statistical model	2.43	0.91	3.68	0.66	3.43	0.44	2.29	0.59	1.93	1.93	3.44	0.37
Technical system	2.50	0.97	3.60	0.53	3.78	0.65	3.54	0.86	1.74	1.74	3.88	0.64

Note: The columns Tang., Comp., Cont., Fam., Anth., and M. Comp show the mean values for these variables. Tang. = Tangibility, Comp. = Complexity, Cont. = Controllability, Fam. = Familiarity, Anth. = Anthropomorphism, M. Com. = Machine Competence, DSS = Decision support system. $N = 397$.

Table 6: Results for the linear regressions analyzing the differences between the respective terms for the properties associated with ADM systems.

	Tangibility	Complexity	Controllability	Familiarity	Anthro.	M. Comp.
Constant	2.861**	3.786**	3.510**	2.947**	2.405**	4.001**
	(0.127)	(0.103)	(0.102)	(0.103)	(0.092)	(0.081)
Algorithm	−0.385*	−0.178	0.212	0.390**	−0.465**	−0.232*
	(0.180)	(0.148)	(0.146)	(0.146)	(0.130)	(0.116)
Automated system	−0.209	−0.600**	0.129	0.515**	−0.584**	−0.288*
	(0.185)	(0.151)	(0.149)	(0.150)	(0.134)	(0.119)
Computer	1.321**	−0.532**	0.713**	1.388**	−0.753**	0.227
	(0.180)	(0.147)	(0.145)	(0.146)	(0.130)	(0.116)
Computer program	0.066	−-0.283	0.411**	0.873**	−0.654**	0.102
	(0.179)	(0.146)	(0.144)	(0.145)	(0.130)	(0.115)
Decision support system	−0.827**	−0.536**	−0.150	−0.417**	−0.067	−0.560**
	(0.182)	(0.149)	(0.147)	(0.148)	(0.132)	(0.117)
Machine Learning	−0.741**	0.028	−0.142	−0.195	−0.300*	−0.323**
	(0.177)	(0.144)	(0.142)	(0.143)	(0.128)	(0.114)
Robot	0.623**	0.019	0.428**	−0.385**	−0.511**	−0.139
	(0.179)	(0.146)	(0.144)	(0.145)	(0.130)	(0.115)
Statistical model	−0.440*	−0.104	−0.085	−0.671**	−0.472**	−0.561**
	(0.186)	(0.152)	(0.150)	(0.151)	(0.135)	(0.120)
Technical system	−0.366*	−0.181	0.268	0.591**	−0.665**	−0.120
	(0.179)	(0.146)	(0.144)	(0.145)	(0.130)	(0.115)
Control Variable						
Affinity for technology	0.188**	−0.143**	0.097**	0.236**	−0.076**	0.075**
	(0.040)	(0.033)	(0.032)	(0.032)	(0.029)	(0.026)
R^2	0.394	0.147	0.162	0.505	0.156	0.196
F	25.145**	6.652**	7.420***	39.311**	7.121**	9.417**
	(df = 10; 386)	(df = 10; 385)	(df = 10; 385)	(df = 10; 386)	(df = 10; 386)	(df = 10; 386)

Note: The effects for the terms can be interpreted in comparison to the reference group artificial intelligence.
Anthro. = Anthropomorphism, M. Comp. = Machine Competence. The columns Tangibility, Complexity, Controllability, Familiarity, Anthro., and M. Comp. show estimates and respective standard errors in brackets.
*p < .05, **p < .01. N = 397.

Table 7: Means and standard deviations for fairness, trust, and justice depending on the work evaluation and work assignment task and depending on the different terminology in Study 2.

Condition	*n*	Fairness		Trust		Justice	
		M	*SD*	*M*	*SD*	*M*	*SD*
Manager, evaluation	44	5.80	1.07	4.61	1.15	2.70	0.76
Manager, assignment	45	5.96	1.11	5.38	0.83	2.69	0.63
AI, evaluation	40	3.00	1.45	2.48	1.09	2.22	0.66
AI, assigmnent	48	5.62	1.14	4.94	1.14	2.61	0.39
Algorithm, evaluation	47	3.74	1.65	3.06	1.39	2.33	0.66
Algorithm, assignment	45	5.73	1.18	4.73	1.44	2.80	0.54
Automated system, evaluation	44	3.41	1.62	2.77	1.27	2.33	0.79
Automated system, assignment	44	5.89	1.04	4.86	1.09	2.64	0.53
Computer program, evaluation	41	3.15	1.61	2.66	1.22	2.33	0.56
Computer program, assignment	46	5.91	1.15	5.28	0.91	2.77	0.47
Robot, evaluation	48	3.38	1.67	3.04	1.50	2.36	0.64
Robot, assignment	43	5.67	1.11	5.14	0.97	2.66	0.41
Statistical model, evaluation	42	4.05	1.64	3.67	1.34	2.27	0.66
Statistical model, assignment	45	5.56	1.18	4.60	0.99	2.52	0.59

Note: The columns Fairness, Trust, and Justice show the mean values for these variables.
$N = 622$.

Table 8: Results of the linear regressions for fairness, trust, and justice evaluations depending on the tasks and the terms in Study 2.

	Fairness	Trust	Justice
Constant	3.019**	2.500**	2.232**
	(0.220)	(0.192)	(0.092)
Work assignment	2.611**	2.445**	0.385**
	(0.297)	(0.259)	(0.125)
Algorithm	0.721*	0.558*	0.099
	(0.299)	(0.261)	(0.126)
Automated system	0.380	0.260	0.089
	(0.304)	(0.265)	(0.128)
Computer program	0.154	0.193	0.115
	(0.308)	(0.269)	(0.130)
Robot	0.340	0.521*	0.117
	(0.298)	(0.260)	(0.125)
Statistical model	1.001**	1.131**	0.020
	(0.309)	(0.269)	(0.130)
Work assignment:Algorithm	−0.626	−0.779*	0.081
	(0.415)	(0.361)	(0.174)
Work assignment:Automated system	−0.125	−0.342	−0.070
	(0.420)	(0.366)	(0.176)
Work assignment:Computer program	0.129	0.146	0.035
	(0.421)	(0.367)	(0.177)
Work assignment:Robot	−0.303	−0.335	−0.077
	(0.417)	(0.363)	(0.175)
Work assignment:Statistical model	−1.048*	−1.439**	−0.099
	(0.424)	(0.369)	(0.178)
Control Variable			
Affinity for technology	0.074	0.097	0.054*
	(0.058)	(0.050)	(0.024)
R^2	0.420	0.431	0.110
F (df = 12; 520)	31.364**	32.765**	5.332**

Note: The results for the tasks can be interpreted in comparison to the task work evaluation. The results for the terms can be interpreted in comparison to the term artificial intelligence. The columns Fairness, Trust, Justice show estimates for the coefficients and respective standard errors in brackets.
*p < .05, **p < .01. N = 533.

Table 9: Results of the linear regression for the comparison of human manager versus the different terms to refer to ADM systems for the task work evaluation in Study 2.

	Fairness	Trust	Justice
Constant	5.765**	4.589**	2.686**
	(0.229)	(0.192)	(0.100)
Artificial intelligence	−2.692**	−2.055**	−0.431**
	(0.333)	(0.279)	(0.146)
Algorithm	−2.037**	−1.538**	−0.360*
	(0.318)	(0.267)	(0.139)
Automated system	−2.391**	−1.845**	−0.376**
	(0.324)	(0.271)	(0.142)
Computer program	−2.518**	−1.849**	−0.307*
	(0.332)	(0.278)	(0.145)
Robot	−2.447**	−1.594**	−0.355*
	(0.317)	(0.266)	(0.139)
Statistical model	−1.820**	−1.005**	−0.466**
	(0.328)	(0.275)	(0.144)
Control Variable			
Affinity for technology	0.277**	0.224**	0.141**
	(0.082)	(0.069)	(0.036)
R^2	0.277	0.246	0.091
F (df = 7; 298)	16.288**	13.904**	4.276**

Note: The results for the terms can be interpreted in comparison to the human manager. The columns Fairness, Trust, Justice show estimates and respective standard errors in brackets.
*p < .05, **p < .01. N = 306.

Table 10: Results of the linear regression for the comparison of human manager versus the different terms to refer to ADM systems for the task work assignment in Study 2.

	Fairness	Trust	Justice
Constant	5.958**	5.378**	2.686**
	(0.168)	(0.160)	(0.077)
Artificial intelligence	−0.342	−0.441*	−0.072
	(0.233)	(0.222)	(0.107)
Algorithm	−0.212	−0.644**	0.117
	(0.237)	(0.226)	(0.109)
Automated system	−0.071	−0.514*	−0.049
	(0.238)	(0.227)	(0.110)
Computer program	−0.046	−0.095	0.082
	(0.236)	(0.224)	(0.108)
Robot	−0.273	−0.238	−0.025
	(0.240)	(0.228)	(0.110)
Statistical model	−0.450	−0.779**	−0.164
	(0.238)	(0.227)	(0.110)
Control Variable			
Affinity for technology	−0.124*	−0.003	0.010
	(0.060)	(0.057)	(0.028)
R^2	0.030	0.060	0.030
F (df = 7; 308)	1.359	2.784**	1.353

Note: The results for the terms can be interpreted in comparison to the human manager. The columns Fairness, Trust, Justice show estimates and respective standard errors in brackets.
*p < .05, **p < .01. N = 316.

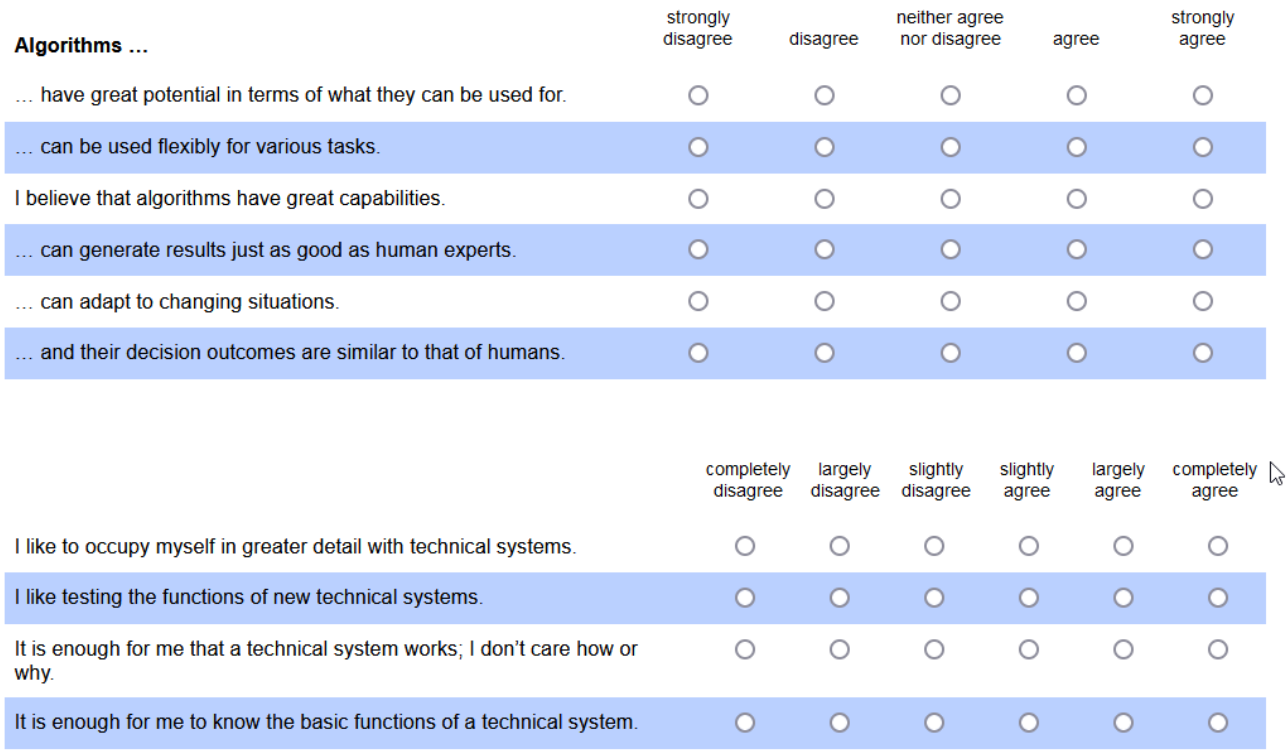

Figure 8: Screenshot from Study 1 where participants reported their perceptions regarding the different terms, here "the term" was algorithm.

Algorithms can ...	worse than a human	slightly worse than a human	as good as a human	slightly better than a human	better than a human
... make shopping recommendations.	○	○	○	○	○
... evaluate documents by applicants (e.g., applicant resumes).	○	○	○	○	○
... make recidivism predictions of convicted offenders.	○	○	○	○	○
... make medical diagnoses.	○	○	○	○	○
... evaluate X-ray and MRI images.	○	○	○	○	○
... predict the weather.	○	○	○	○	○
... evaluate job interviews.	○	○	○	○	○
... produce shift schedules at work.	○	○	○	○	○
... provide therapy recommendations in medicine.	○	○	○	○	○
... diagnose mental illness.	○	○	○	○	○
... identify faces.	○	○	○	○	○
... assess dangerous situations while driving.	○	○	○	○	○
... predict the spread of infectious diseases.	○	○	○	○	○

Figure 9: Screenshot from Study 1 where participants reported their evaluation regarding the performance of a respective term in comparison to a human, here "the term" was algorithm.

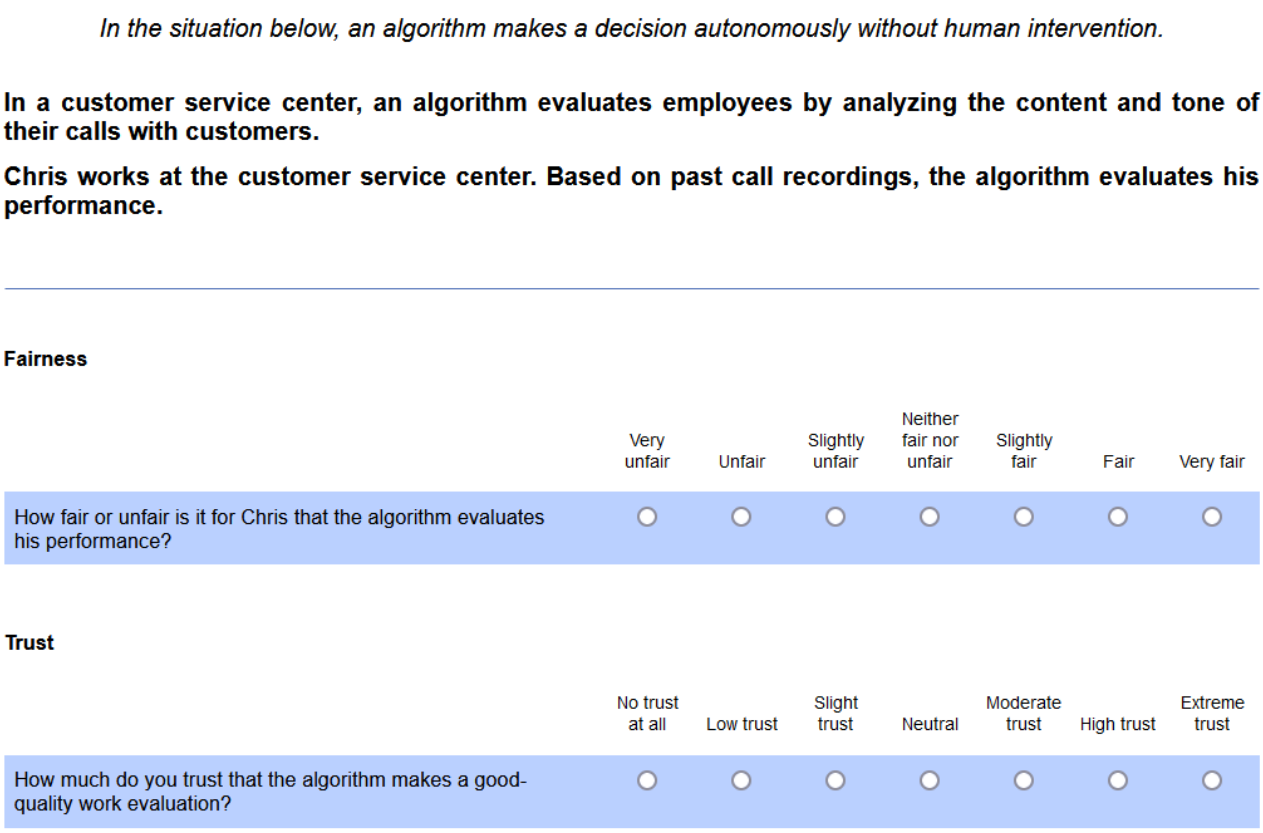

Figure 10: Screenshot from Study 2 where participants reported their evaluation of a respective term regarding the task work evaluation, here "the term" was algorithm.

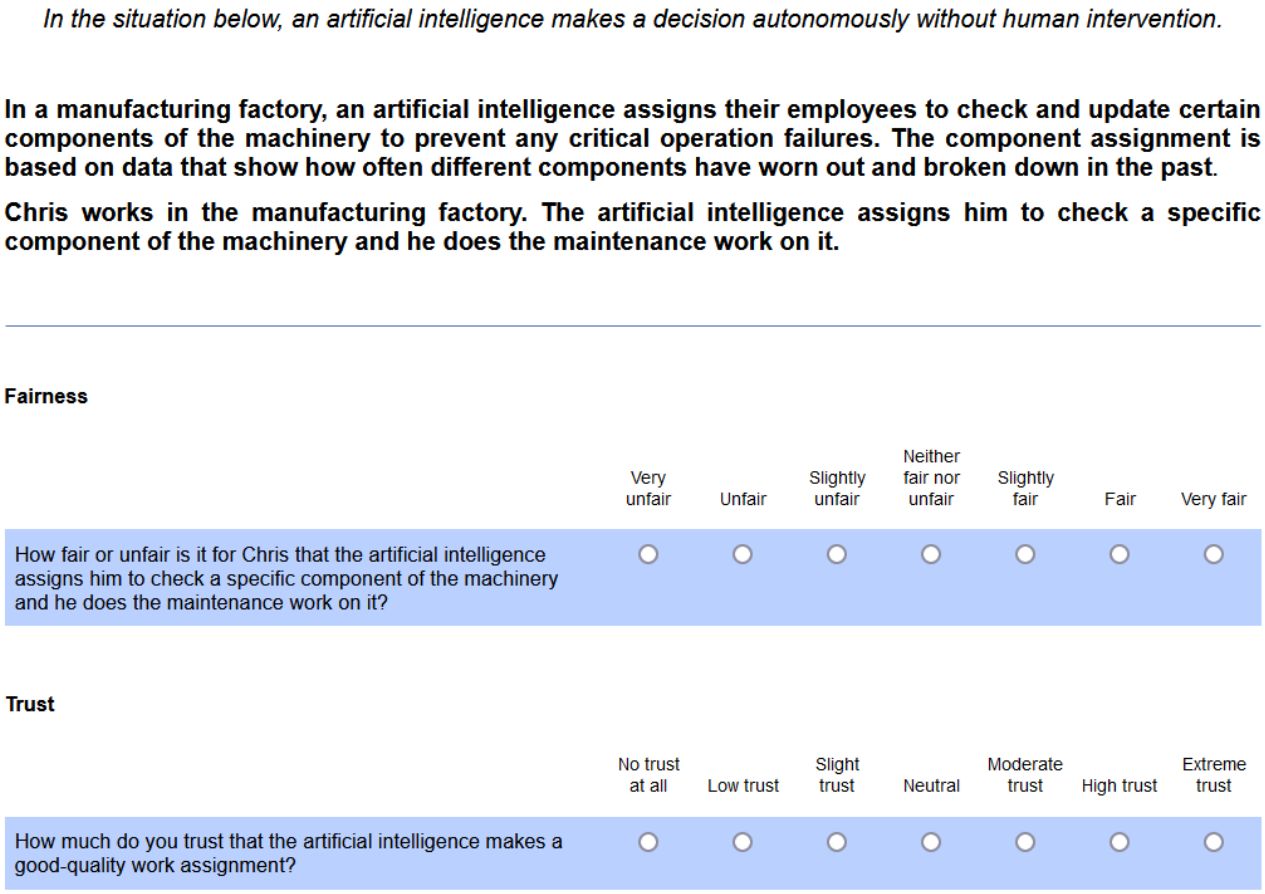

Figure 11: Screenshot from Study 2 where participants reported their evaluation of a respective term regarding the task work assignment, here "the term" was artificial intelligence.