

Master's thesis presented to the Department of Psychology of the University of Basel for the degree of Master of Science in Psychology

Transparency and Trust in AI:

Measuring the effect of human-friendly AI explanations on objective and subjective trust

Author: Nicolas Scharowski, Riehenstrasse 25, 4058 Basel

Immatriculation number: 15-055-759

Correspondence email: nicolas.scharowski@unibas.ch

Examiner: Prof. Dr. Klaus Opwis

Supervisor: Dr. Florian Brühlmann

Institute of Psychology, Center for Cognitive Psychology and Methodology

Submission date: 11. November 2020

Transparency and Trust in AI

Measuring the effect of human-friendly AI explanations on objective and subjective trust

Master's Thesis - Nicolas Scharowski

Center for Cognitive Psychology and Methodology, University of Basel

Submitted in November 2020

Thesis supervisors:

Dr. Florian Brühlmann

Prof. Dr. Klaus Opwis

Author Note

Nicolas Scharowski, 15-055-759

Riehenstrasse 25, 4058 Basel

nicolas.scharowski@unibas.ch

Acknowledgement

I want to thank my thesis supervisor Dr. Florian Brühlmann for his support in implementing the experiment, as well as for his advice concerning the data analysis. Further, I thank Dominik Schumacher for his proofreading and Julia Oplatka for her valuable feedback. Finally, I thank my great aunt Hildegard Mohr for talking and listening to me.

Declaration of scientific integrity

The author hereby declares that he has read and fully adhered the code for good practice in research of the University of Basel.

Abstract

In human-computer interaction research, transparency is widely regarded as crucial for user trust in artificial intelligence (AI). Transparency is expected to provide the end-user with understanding and knowledge about the functionality of an AI, which in turn creates trust. However, empirical investigations on this assumption have been largely omitted, and there are several proposals as to how transparency could be achieved. This thesis explores human-friendly AI explanations as a means for transparency and examines the effects of explanations on objective trust behavior and subjective measures of trust. An online experiment ($n = 387$) was conducted that compared two explanation techniques with a control. Study participants were asked to estimate subleasing prices of six apartments based on respective features and amenities. After this estimate, a pseudo AI provided participants with a price recommendation, which was accompanied by an explanation for all but the control group. Results showed that human-friendly explanations lead to higher trust behavior if participants were advised to *decrease* the initial price estimate. However, explanations had no effect if the AI recommended to *increase* the initial price estimate. Trust was further evaluated by validated trust questionnaires which revealed no effects of human-friendly explanations on subjective trust ratings. Possible reasons for this discrepancy between objectively observed trust behavior and subjectively rated trust are discussed and implications for the design of transparent AI through human-friendly explanations are suggested.

Contents

Abstract	3
Introduction	6
Theoretical Background	8
Algorithmic Transparency	8
Interpretability and explainability	9
Global vs. local	9
Intrinsic vs. post-hoc	10
Model-specific vs. model-agnostic	10
Human-friendly Explanations	12
Explanations are selective	13
Explanations are contrastive	13
Explanations are not about probabilities	13
Explanations are social	14
Explainability techniques	14
Feature importance	15
Counterfactuals	15
Trust	16
Trust in HCI and xAI	17
Measurements and models of trust in AI	18
Previous work and aim of this study	19
Research question and hypotheses	21
Method	21
Participants	22
Procedure and Task	23
Stimuli	24
Measures	26

TRANSPARENCY AND TRUST IN AI	5
Independent variables	26
Dependent variables	26
Data cleaning	27
Results	27
Descriptive statistics	27
Objective Trust - WOA	29
Subjective Trust - Trust in automation scale	32
Discussion	35
Limitations and future work	39
Conclusion	41
References	43

Introduction

It is generally recognized that there are specific tasks computers perform better than humans, such as numeracy, logical reasoning or storing information (Copeland, 2015). With the recent breakthroughs in artificial intelligence (AI), however, domains that used to be exclusively associated with human competence and considered computationally unattainable are likewise challenged by machines. Advances in AI in the last decade were achieved thanks to an increase in available data, major hardware improvements and new algorithms, especially with the emergence of machine learning (ML) algorithms (Došilović, Brčić, & Hlupić, 2018). While classical algorithms perform automated instructions that are rule-based, ML is a set of algorithms that can modify themselves and make varying decisions in response to different data inputs (Molnar, 2019). In a sense, ML is not programmed to perform a task, but programmed to learn to perform a task. This ability to change and learn is described as *intelligence* and makes ML a branch of modern AI. While recognizing these differences, the terms algorithms, ML and AI are used interchangeably in this thesis due to their conceptual proximity. The ML approach in AI led to improvements in speech recognition, image classification and object detection, and is now increasingly used in a variety of everyday applications such as video surveillance, email spam filtering, online customer support and product recommendations. Because of this general applicability and potential manifold consequences, voices are being raised that these algorithms should satisfy criteria like fairness, reliability, accountability and transparency. While all those criteria seem equally important, this work will focus on the notion of transparency. Since humans can be directly affected by the decision of an algorithm (e.g., an algorithm that *decides* if a requested loan will be granted to an applicant or not), or be involved in the decision-making process itself (e.g., an algorithm that *recommends* to a business corporation that their production should be decelerated), the *what*, *how* and *why* of those algorithmic decisions should be explained transparently to humans. This call for transparent algorithms has led to the research field of explainable artificial intelligence (xAI) that explores methods and models that make the behaviors, predictions and decisions of AI transparent and understandable to

humans. While xAI is typically referred to as a multidisciplinary field (Mohseni, Zarei, & Ragan, 2020), most of the work conducted today comes from the machine learning community. However, explanations from algorithms and AI to humans are not solely a computational problem. The interpretation of what is and what is not a useful explanation, is ultimately defined by people, not algorithms. Therefore, measured outcomes such as trust are consequently properties of human behavior (Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, & Wallach, 2018) and as a consequence, the endeavor of xAI should go beyond machine learning research, extending into human-computer interaction (HCI) and other disciplines. The HCI community has acknowledged the importance of their field in the academic discussion and has defined the need for transparent AI as one of the *grand challenges* for HCI researchers (Stephanidis et al., 2019). Transparency is not only legally required, but is also thought to contribute towards building a relationship of trust between humans and algorithms (Stephanidis et al., 2019). Others argue that transparency allows humans to question a system in order to develop appropriate trust and reliance, rather than blind faith (Rader, Cotter, & Cho, 2018). Nevertheless, there are still few empirical studies that evaluate the impact of transparency on trust (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018). Existing work in xAI often focuses on AI developers rather than end-users that encounter the decisions of such algorithms. We will see that for end-users, explaining the decision of an algorithm might be more feasible than merely making an algorithm transparent. This thesis aims to fill these gaps by investigating the effect of different AI explanations on objective and subjective user trust. A decision-making experiment on Amazon Mechanical Turk (MTurk) was conducted to compare two promising explanation techniques with a control condition. AI explanations are often axiomatically demanded to enable user trust, but so far, a comprehensive comparison of different explanation techniques has not been carried out and there is no evidence that explanations are indeed increasing trust. Given HCI's focus on technology that helps people better understand and collaborate with technology, more work from the HCI community is needed in this field, and the thesis presented here aims to make a contribution.

Theoretical Background

In this section, a detailed theoretical background of the relevant concepts, the terminology and currently discussed transparency approaches in HCI and xAI is provided. This serves as a starting point for the legal importance of transparent algorithms and AI, what current transparency methods are, what they try to accomplish, and to describe the research approach employed in this empirical work.

Algorithmic Transparency

As responsibilities and processes are increasingly delegated to automated decision-making systems, more attention is being paid to algorithmic and AI transparency (Rader et al., 2018). Generally speaking, algorithmic transparency describes the level to which a system provides information about its workings or structure (Ras, van Gerven, & Hase-lager, 2018). Promises about transparency are commonly driven by a certain chain of logic — transparency enables observations that produce insights, which in turn create knowledge about the goals, intent and behavior of a system (Ananny & Crawford, 2018). This knowledge, according to this line of reasoning, is required to govern algorithms, to hold them accountable and to ensure their fairness. Not only researchers, but also policy-makers advocate transparency as a remedy for identifying and preventing potentially negative effects of such systems. An example is the introduction of the EU general data protection regulation (GDPR) by the European parliament. Its frequently mentioned notion of “a right to an explanation” in automated decision-making processes is closely related to transparency. A *right to an explanation* constitutes an explanation for a specific output of an automated process, which is often carried out by an algorithm or AI. According to recital 71, article 12 and article 14 of the GDPR, an explanation should be given "after an assessment, in a concise, transparent, intelligible and easily accessible form, using clear and plain language, provided in writing" that should include "meaningful information to the subject about the logic involved" (European Parliament, 2018). Some authors do not agree that a legally binding *right to an explanation* exists in the final version of the GDPR (Wachter, Mittelstadt, & Floridi, 2017). Nevertheless,

the GDPR might provide a foundation for how algorithmic and AI explanations could be legally envisaged. While transparency is generally considered crucial for effective and responsible real-world deployment of such systems, significantly different transparency approaches exist, tailored to the algorithm’s goal, context and target group, such as developers, decision makers and end-users (Samek, Montavon, Vedaldi, Hansen, & Müller, 2019). Due to this, it is necessary to initially distinguish and examine two diverging approaches of transparency more thoroughly — interpretability and explainability.

Interpretability and explainability

In the context of xAI, explainability and interpretability are often used interchangeably as they are closely related terms. Both attempt to accomplish transparency, yet their rationale and level of implementation differ. In his influential work, Lipton (2018) specified the previously ill-defined concepts. He suggested that interpretability is the information that a system provides about its inner workings. In this sense, interpretability is associated with the notion of *transparent white box algorithms*, meaning algorithms whose internal mechanisms are accessible and not concealed. In contrast, *opaque black box algorithms* can only be viewed in terms of their inputs and outputs, without any direct observations of their inner workings. Explainability aims to give meaningful information by explaining how a specific output or decision of such black box algorithms was reached. In this thesis I define the two terms as follows: interpretability is achieved by using a white box algorithm that can be observed, whereas explainability implies using a black box algorithm and making it comprehensible by explaining its output after a computation has been carried out. Adadi and Berrada (2018) described the different transparency approaches on three complementary axes and presented a comprehensive overview of the xAI taxonomy:

Global vs. local. Describes the differentiation between explaining an individual output (local explanation) or interpreting the entire model (global interpretability) (Adadi & Berrada, 2018). It is therefore identical with the distinction between the white box and black box approach introduced above.

Intrinsic vs. post-hoc. Likewise, researchers distinguish whether transparency is achieved by restricting the complexity of the algorithm (therefore providing intrinsic interpretability), or by applying techniques that analyze the algorithm after a decision has been made (providing post-hoc explanation) (Molnar, 2019). Intrinsic interpretability is thus provided due to the simpler structure of an algorithm and theoretically, intrinsic interpretability can accomplish transparency with the white-box approach. Post-hoc techniques, on the other hand, do not describe the internal state of an algorithm, but nonetheless extract useful information from black boxes after a decision has been reached (Lipton, 2018). This is often enabled by applying a simpler model post-hoc that describes the more complex one.

Model-specific vs. model-agnostic. Model-specific methods are limited to specific model classes or a single type of algorithm (Molnar, 2019). Regression weights in a linear model, for example, are model-specific. Model-agnostic, on the other hand, signifies that the method can be applied to any type of algorithm. Model-agnostic techniques are applied after the model has been trained (post-hoc) and usually function by analyzing input and output. While intrinsic techniques are by definition model-specific, post-hoc techniques are usually model-agnostic (Adadi & Berrada, 2018).

This clarification follows the notion of Lipton (2018), which states that explainability is post-hoc interpretability. This definition is adopted for this thesis with the addition that explainability is achieved by local or global post-hoc explanations that are usually model-agnostic. On the other hand, interpretability is achieved by global intrinsically interpretable algorithms that are, by definition, model-specific. Both concepts attempt to increase transparency, yet there are certain limitations to interpretability. I argue that in the context of algorithmic transparency, whereas it is adequate to apply explainability, it might not be reasonable to apply interpretability. The reasons for this are outlined in the following.

- **Overwhelming Complexity:** Many algorithms, especially ML algorithms, can be overwhelmingly complex and may have an internal state composed of millions

of interdependent values and parameters (Mittelstadt, Russell, & Wachter, 2019). As such, trying to reveal the functions used to reach decisions may well be too complex, even for developers of such algorithms, let alone end-users. Wachter, Mittelstadt, and Russell (2018) argue that displaying the internal state of complex algorithms to end-users and achieving transparency in the form of interpretability is extremely challenging, and might be virtually impossible. Interpretability therefore does not necessarily contain meaningful information that allows end-users to reason about the decision-making process of an algorithm, as aimed to accomplish by transparency.

- **The interpretability trade-off:** In the pursuit of ever more accurate predictions, algorithms are getting increasingly complex and a certain trade-off between accuracy and interpretability becomes evident (Shmueli, 2010). ML algorithms, for example, have a high predictive power but are also multi-layered and more complex compared to classical algorithms. The trade-off leads to the difficult question what developers should optimize for — accuracy (how well the algorithm predicts an outcome) or interpretability (how well the algorithm can be observed and interpreted). Doshi-Velez et al. (2017) reason that it could be counterproductive to trade accuracy for interpretability, given that one of the biggest benefits of complex algorithms and AI is the ability to detect patterns that humans fail to identify.
- **Corporate compliance:** Companies like Spotify, YouTube and Amazon rely on business models for which algorithms and AI are paramount, since they create product recommendations that increase revenue. With the interpretability trade-off in mind, it seems unreasonable to assume that corporations would favor interpretable algorithms for better performing ones. Moreover, there exist concerns that interpretability would force trade secrets to be revealed, as it requires the innermost mechanism of an algorithm to be displayed (Doshi-Velez et al., 2017). From this perspective, corporate compliance for interpretability is expected to be low.

- **Legal requirements:** The *right to an explanation* that the GDPR demands on the one hand, and the ever more complex algorithms, on the other hand, seem to create a field of tension. From the legal viewpoint of the previously introduced GDPR, however, the differentiation between the concepts interpretability and explainability seems straightforward: the GDPR stipulates explanation, as in *right of an explanation*, not interpretation. Given Lipton’s emphasis of explainability as *post-hoc interpretability* and contrasting it with the GDPR’s request for *an explanation of the decision reached after such assessment*, the legal requirements appear to be better satisfied with explainability rather than interpretability.

On account of those arguments, applying interpretability to complex algorithms and AI might not be feasible and promising when aiming for transparency. Beyond that, making a system transparent does not necessarily mean that it leads to insight, understanding or *meaningful information* as the GDPR suggests. For this reason, this thesis exclusively focuses on explainability with regard to transparency and examines different explainability techniques. To identify, however, which explainability techniques are the most promising ones to empirically compare, it seems helpful to first consider what humans generally evaluate as good explanations.

Human-friendly Explanations

In the endeavor for meaningful explanations, researchers emphasize the importance of incorporating insights from philosophy, social sciences and psychology into the field of xAI because of their research on how people *define, generate, select, evaluate* and *present* explanations (Miller, 2019; Mittelstadt et al., 2019). De Graaf and Malle (2018) argue that because people assign human-like traits to artificial agents, they will expect explanations from them that are similar to the way in which humans explain their actions. Generally speaking, explanations are distinguished according to their completeness or the degree to which the entire causal chain of events can be explained (Mittelstadt et al., 2019; Ruben, 2015). This is often expressed as the difference between full *scientific explanations* and partial *everyday explanations* (Ruben, 2015). Miller (2019) argues that these *everyday*

explanations are, essentially, what people demand when they inquire about the decision-making process of another human being. To him, partially describing how certain factors were used to reach the decision in a specific situation is of greater importance than a description of the full decision-making process in general. For *everyday explanations*, any clarifying information can be considered an explanation (Doshi-Velez et al., 2017). Miller (2019), however, argues that not all explanations are equal, and that some information is more valuable for humans than others. This section further discusses the work of Miller (2019) on human-friendly explanations and adds implications for HCI and xAI that Molnar (2019) elaborated.

Explanations are selective. As aforementioned, humans rarely expect actual and complete *scientific explanations* of a decision P (or prediction, output, event etc.). People often select the most important or immediate causes from a sometimes infinite number of causes to be the *everyday explanation* (Miller, 2019). The findings of Lisman and Idiart (1995), that an average person is said to remember no more than 7 ± 2 pieces of information at a time, might be another reason for favoring selective explanations that can be successfully processed and remembered.

Explanations are contrastive. Miller (2019) argues that humans usually do not ask why a certain decision P happened, but why this decision P has happened instead of another decision Q . This approach has also been described as *counterfactual explanations*. Miller (2019) argues that people better process contrastive or counterfactual explanations. In the counterfactual case, only what is different between two events has to be explained. In this light, the best explanation is the one that highlights the greatest difference between the decision P and decision Q .

Explanations are not about probabilities. For Miller (2019), referring to probabilities in an explanation is not as effective as referring to causes. As a result, using statistics to explain why a decision P occurred can be unsatisfying for humans (Miller, 2019). Numerous psychological experiments demonstrated that humans struggle to comprehend probabilities, and it is heavily debated if people think in accordance with the laws of probability theory or merely use heuristics and biases (Gigerenzer & Selten, 2002).

Explanations are social. For humans, explanations are a form of social interaction or more specifically a transfer of knowledge, often presented as part of a conversation (Miller, 2019). As a consequence, practitioners should pay attention to the social environment when implementing explainability. Molnar (2019) points out that explaining something to an expert is different than explaining something to a layperson. Moreover, and as demonstrated by the confirmation bias (Nickerson, 1998), humans tend to ignore or devalue information that is inconsistent with their prior beliefs. This implies that for different target audiences, explanations are not equally valuable. However, this social aspect is not inherent in explanations coming from algorithms and AI.

While these four key points do not encompass all of Miller’s profound work on human-friendly explanations, they certainly contain its essence. I agree with the argumentation of Miller (2019) and Mittelstadt et al. (2019) that explanations should at least fulfill some of the above mentioned criteria to be truly meaningful to end-users. Taking this disparity into consideration, I limited the explainability techniques that were empirically compared in this thesis.

Explainability techniques

Explainability techniques broadly fall into four categories (Adadi & Berrada, 2018):

1. Visualization
2. Knowledge extraction
3. Influence methods
4. Example-based explanation

Taking into account the above-mentioned complementary axes (global vs. local, intrinsic vs. post-hoc, model-specific vs. model-agnostic), more than 17 different explanation techniques are being proposed and debated in the current HCI and xAI literature. The varying explanation approaches offer advantages for different target audiences

(experts, decision makers, AI practitioners etc.), but by applying the requirements of the GDPR and Miller’s insights of what constitutes a good explanation, the techniques that come into question for end-users can be narrowed down substantially. From the 17 explanation approaches, only a handful meet the GDPR’s requirement of *meaningful explanations after an assessment (post-hoc) using clear and plain language, provided in writing*. With the implications of Miller’s human-friendly explanations in mind that facilitate *meaningful information*, the two most promising explainability techniques seem to be **feature importance** and **counterfactuals**.

Feature importance. As the name suggests, feature importance explains which features are most important for an algorithmic outcome or the decision of an AI. As Miller (2019) suggested, the selection of the most important causes is also what humans are interested in when asking for an explanation. The feature importance method has the following notation: *Outcome P was returned because variable V had values (vi, vii, ...) associated with them*. To provide a simplified example: "Flu was returned because temperature had value 39°C". Feature importance allows end-users to determine which feature had the most impact on the outcome and hence fulfills Miller’s proposed *selective* criterion for explanations.

Counterfactuals. In addition to the main causes of an outcome, counterfactuals provide *if-then* statements that help a user identify what might be changed to achieve a desired outcome. Counterfactuals commonly have the following form: *Outcome P was returned because variables V had values (vi, vii, ...) associated with them. If V had values (vi', vii', ...) instead, and all other variables remained constant, outcome P' would have been returned*. Thinking in counterfactuals requires imagining a hypothetical reality that contradicts the observed facts, hence the name *counterfactual*. For the same simplified example as before, this could lead to the following statement: "Flu was returned because temperature had value 39°C. If temperature instead had value 37°C, cold would have been returned". Wachter et al. (2018) argue that counterfactuals help an end-user act rather than merely understand by altering future behavior for a desired outcome. Counterfactuals combine Miller’s *selective* and *contrastive* criteria.

Both techniques are structurally similar and can be classified as *local model-agnostic post-hoc explanations*, but while feature importance falls into the category of the *influence methods* approach, counterfactuals are categorized as *example-based explanations* (Adadi & Berrada, 2018). However, both techniques provide information to the end-user that is both human-friendly and meets the explicit requirements of the GDPR, while avoiding the major pitfalls of interpretability.

Despite this extensive introduction to algorithmic transparency and human-friendly explanations, the question remains what the goals and motivations of explainability are in the context of algorithmic-decision making and AI. While this thesis focuses on trust as a motivation for explainability, other purposes of explanations are being discussed in the scientific literature and Lipton (2018) points out that the HCI and xAI communities provide diverse and sometimes non-overlapping motivations. Some authors argue that explainability could help verify and improve the functionality of a system (i.e. for debugging), support developers to learn from a system (i.e. generating hypotheses), or to ensure fair and ethical decision-making (Mittelstadt et al., 2019). In this thesis, the focus lies on the rationale that explainability enhances trust in the AI system and its reached decisions. A brief and non-exhaustive introduction to trust is thus provided. Subsequently, trust will then be discussed in the context of HCI and xAI respectively, and established trust models and measurements are introduced.

Trust

The work of Andras et al. (2018) provides a comprehensive overview of the multi-layered facet of trust and sometimes diverging trust concepts from different disciplines. The authors argue that in the social world, trust is the expectation of non-hostile behavior; in the context of economics, trust is conceptualized through game theory; in psychological terms, trust represents cognitive learning from experiences, and philosophically speaking, trust is based on moral relationships between individuals (Andras et al., 2018). In his book *Vertrauen - die unsichtbare Macht*, Hartmann (2020) criticizes that the everyday use of the word trust is misleading when applied to technology and that

trust in this case must be differentiated from the concept of *reliance*. For Hartmann, trust is a property of the human condition, characterized as an acceptance of vulnerability and human expectation that this vulnerability will not be exploited. Reliance, on the other hand, is more related with the predictability of a behavior of someone or something. Since algorithms and AI have no knowledge about human vulnerability, Hartmann concludes that it is unreasonable to talk about trust when referring to interactions between humans and algorithms, or AI respectively. Instead, it is reliance that applies.

Trust in HCI and xAI. Despite this ambiguity, within the HCI and xAI communities, trust in algorithmic decisions and AI often seems to be demanded axiomatically without further clarification. Chopra and Wallace (2003) criticise that the HCI and xAI literature lacks a conclusive definition of trust, as well as a consensus about its desired effects, and a clear differentiation among the factors contributing to trust. Trust is believed to boost the performance of the human-system collaboration, is thought to be a key factor affecting the way people rely on automated systems, and is suggested to be closely connected to usability and user satisfaction (Stephanidis et al., 2019; Yu et al., 2017). Research on trust in human-system and human-machine has a long history in HCI, such as *online trust*, *trust in e-commerce*, *trust in technology*, *trust in recommender systems*, *content trust* and so on (Corritore, Kracher, & Wiedenbeck, 2003). Hoff and Bashir (2015) claim that some similarities are apparent across domains and contexts and that almost every definition of trust seems to include the following three characteristics:

1. First, there are two parties in a trusting relationship — a truster to give trust and a trustee to accept trust.
2. Second, the trustee must perform a task that the truster desires. The trustee on the other hand has an incentive to carry out the task. This incentive can consist of a monetary reward or the benevolent desire to help. In interactions with technology, the incentive for the trustee is usually based on the designer's intended use for a system.

3. Finally, there is the possibility that the trustee will not succeed to perform the desired task or perform it poorly. For the truster, trust therefore contains some sort of risk-taking behavior and that trust is needed when something is exchanged under uncertainty and vulnerability.

This applies to both interpersonal and human-automation trust, where trust in algorithms and AI is likewise positioned. As elaborated before, the GDPR considers algorithms and AI as cases of automated decision-making processes. In their systematic review of trust in automation, Hoff and Bashir (2015) proposed three layers of trust that conceptualize the variability of the concept: *Dispositional trust*, *situational trust* and *learned trust*. This distinction is in accordance with the proposal of Yu et al. (2017), defining trust as a multidimensional construct. *Dispositional trust* reflects the user's natural tendency to trust machines and technologies and involves cultural, demographic and personality factors. *Situational trust* refers to more specific factors, such as the task to be performed, the complexity and type of system, a user's workload, perceived risks, and even mood. And finally, *learned trust* encapsulates the experiential aspects of the construct which are directly related to the system itself.

Measurements and models of trust in AI. These different layers of trust and their variability imply that trust is a subjective construct experienced differently by different people. The differentiation between subjective and objective trust and the measurement of those constructs in xAI was addressed by Mohseni et al. (2020). They point out that subjective trust measures include *self-explanation during or after working with a system* and *Likert scale questionnaires*. To the best of my knowledge, no validated and established questionnaire exists that measures subjective trust in algorithms or AI. However, for human-automation trust, Jian, Bisantz, and Drury (2000) designed the *trust in automation scale* that has been evaluated several times (Gutzwiller et al., 2019; Spain, Bustamante, & Bliss, 2008). In their systematic approach on trust in machine learning and AI, Toreini et al. (2020) propose the trust model from Mayer, Davis, and Schoorman (1995) and the further developed ABI+ framework by Sanders, Schyns, Dietz, and Den Hartog (2006) as a possible candidate for trust evaluation. The ABI+ framework

consists of the the attributes *ability*, *benevolence*, *integrity* and *predictability*. *Ability* is the perception of the skill, competence, resources and capabilities needed that a trustee can have a successful influence within some specific domain. *Benevolence* is defined as the extent to which a trustee is believed to want to do good, is interested in the welfare of the truster and the absence of opportunistic behavior. To possess *integrity*, a trustee must fulfill their promises, be perceived to act in accordance to a set of principles and values that is shared with the truster. Finally, *predictability* reinforces the perception of the other attributes (ability, benevolence and integrity) over time by demonstrating consistent, regular and therefore predictable behavior. While Toreini et al. (2020) did not provide any questionnaires for their ABI+ framework, the *TrustDiff* developed by Brühlmann, Petralito, Rieser, Aeschbach, and Opwis (in press) captures the trust dimensions *benevolence*, *competence* and *integrity* that are closely related to the dimensions of the ABI+ framework.

For objective measures of trust, Mohseni et al. (2020) propose *perceived system competence*, *intention to return*, *user compliance* as well as *reliance with systems* and *user's perceived understanding*. Another objective measure of trust behavior was suggested by Dhurandhar, Iyengar, Luss, and Shanmugam (2018). They argue that a measured change in behavior or performance after being presented an explanation (e.g. a person reducing the speed of a semi-autonomous car after it explained to the driver that speeding on a wet road is dangerous) could be an objective and mathematically quantifiable measure of trust behavior. This interpretation of trust is measured by the parameter *weight of advice* (WOA) that stems from the advice-taking literature (Harvey & Fischer, 1997). WOA measures the degree to which people move their initial estimates towards an advice, thus changing their behavior as a consequence of trust.

Previous work and aim of this study

Past research on algorithmic transparency has focused primarily on interpretability rather than explainability (Krause, Perer, & Ng, 2016; Poursabzi-Sangdeh et al., 2018; Springer, Hollis, & Whittaker, in press). Narayanan et al. (2018) explored the impor-

tance of explanation properties and demonstrated that explanation size, the number of cognitive chunks and the number of repeated terms in an explanation had an impact on participants' response time and subjective satisfaction. In a similar experiment, Kizilcec (2016) showed that short textual explanations build trust with the algorithm's results. However, as is mostly the case when trust is assessed, the authors employed self-defined Likert scales with questionable validity instead of established trust questionnaires that are based on frameworks. An exception is the work of Cai, Jongejan, and Holbrook (2019): it was able to reveal that example-based explanations in a visual task lead to higher ratings in user's trust dimensions *benevolence* and *ability*. Contrary to the widespread believe that people are averse to algorithms (Dietvorst, Simmons, & Massey, 2014), results from experiments conducted by Logg, Minson, and Moore (2019) suggest that people are willing to rely on algorithmic advice in decision-making processes. Logg et al. (2019) showed that under certain circumstances, participants trusted advice that were supposedly coming from an AI more than advice that were labelled to come from a human. For this, Logg et al. (2019) quantified how much people relied on AI advice, using the previously introduced trust parameter WOA. So far, trust in AI has been investigated in arbitrary and rather hypothetical domains like estimate tasks of a persons level of attractiveness (Logg et al., 2019), forecasting romantic matches (Yin, Wortman Vaughan, & Wallach, 2019), and predicting which jokes people will find funny (Yeomans, Shah, Mullainathan, & Kleinberg, 2019). These highly subjective domains are selected to investigate whether AI recommendations are also perceived as trustworthy in the realm of typically human judgements. A more objective domain was chosen by Poursabzi-Sangdeh et al. (2018). They examined if interpretability had an effect on how people trust AI and change their house price estimates. However, contrary to what one might expect when applying interpretability, the authors found no significant differences in trust across conditions. To the best of my knowledge, no similar experiments have been conducted that compare and evaluate different explainability techniques with one another, let alone experiments conducted with end-users. This thesis aims to fill this gap and the work by Poursabzi-Sangdeh et al. (2018) served as an inspiration for the present study.

Research question and hypotheses

The following research question was investigated:

RQ: What effect do different explainability techniques have on objective and subjective end-user trust?

To answer this research question, the previously introduced explanatory techniques *feature importance* and *counterfactuals* were compared with a control condition in a scenario in which participants had to guess subleasing prices for different apartments. While feature importance provide selective explanations, counterfactuals are also contrastive by nature, making them a more promising candidate for human-friendly explanations. For this reason, the specific hypotheses are:

H₁ : The experimental conditions lead to higher objective trust scores compared to the control.

H₂ : Counterfactuals lead to higher objective trust scores compared to feature importance.

H₃ : The experimental conditions lead to higher subjective trust scores compared to the control.

H₄ : Counterfactuals lead to higher subjective trust scores compared to feature importance.

The presumption was that objective and subjective trust behave in an equal manner in so far as subjectively stated trust should also be reflected in objective trust behavior — the higher the subjectively stated trust, the higher the objectively observed trust behavior and vice versa.

Method

To answer the research question, an online experiment over the crowd-sourcing marketplace website Amazon Mechanical Turk (MTurk) was carried out (<http://mturk.com>). First, as a matter of principle, a power analysis using the software PANGAEA (Westfall, 2015) was performed. For a presumed small effect size ($d = .2$) a total of 480 participants was calculated to achieve a power of 80%. The experiment was implemented over the online survey tool Limesurvey (<http://limesurvey.org>).

Participants

A total of 913 participants were initially recruited over Amazon Mechanical Turk. Only workers from the USA with a human intelligence task (HIT) approval of 95% and at least 100 approved HITs were allowed to participate in the experiment. Workers who properly completed the task were reimbursed with 1.50 U.S. dollars and a bonus of 0.30 U.S. dollars for their participation. To ensure data quality, certain criteria were applied during data cleaning (see Figure 1). Participants that failed to provide correct answers in the bogus items ($n = 36$) and demonstrated careless responding in the control questions ($n = 310$) were removed. After excluding outliers ($n = 65$), 387 participants remained for the data analysis. The defining criteria for outliers are addressed in further detail in the section *data cleaning*.

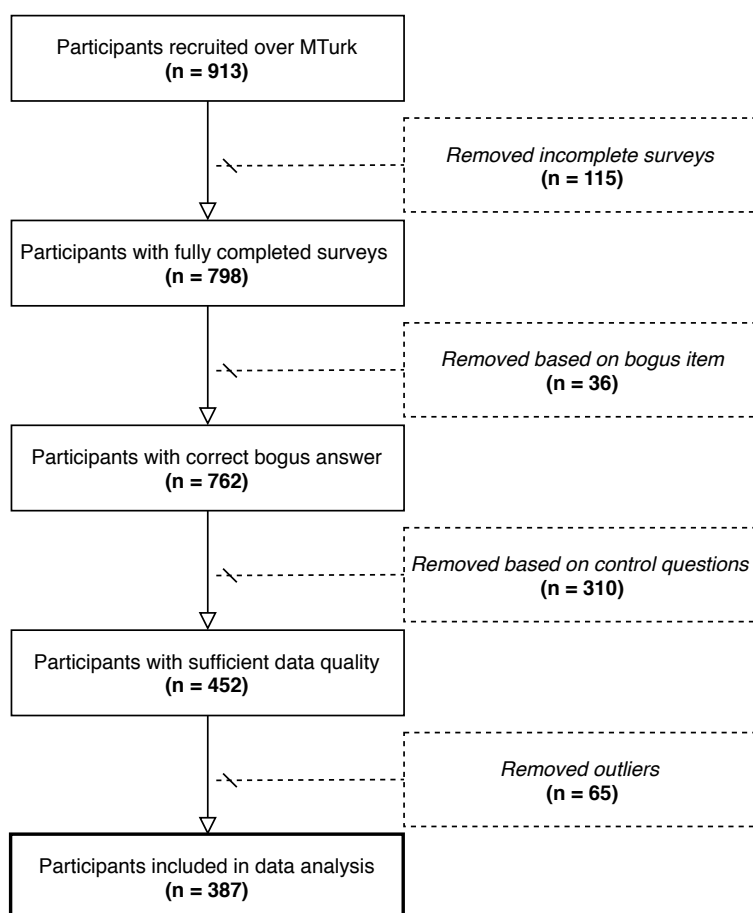


Figure 1. Flowchart of study sample inclusion and exclusion criteria.

Procedure and Task

After providing informed consent, participants were introduced to the study and their task. They were asked to imagine a scenario where their goal was to sublease six different apartments on a subleasing website. Based on the features and amenities of the apartment (e.g. number of bedrooms, distance to public transit, etc.), they had to guess an initial subleasing price (T1). Figure 2 illustrates how the apartments were presented to the participants.

The features and amenities of this apartment are the following:

Main features:

Size: 1'100 square feet / # Bedrooms: 2 / # Bathrooms: 2

Other features and amenities:

Year built:	1960
Elevator:	YES
Doorman:	NO
Balcony:	YES
Fitness center:	YES
Pets:	YES
Parking:	YES
AC:	YES
Dishwasher:	YES
Laundry:	On-Site
Distance to public transit:	0.2 miles
Distance to school:	0.5 miles




Figure 2. Example of a presented apartment as stimuli.

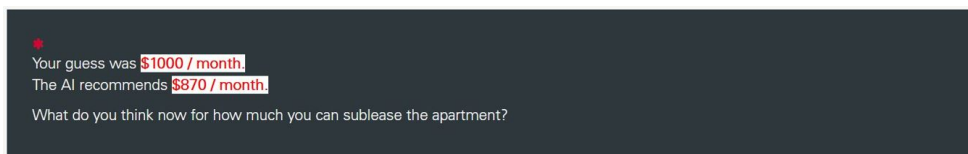
After guessing T1, an alleged AI from the website provided a computed price recommendation. In reality, however, a price recommendation based on basic arithmetic, rather than an actual AI, was given. How exactly this price recommendation was calculated is discussed in the section *stimuli*. After seeing the price recommendation, participants could decide if they wanted to approach the AI recommendation or not, settling for a final subleasing price (T2). Participants were informed that by choosing a lower price, they would be more likely to find a sublesser, but would receive less profit. When choosing a higher price on the other hand, they would be less likely to find a sublesser, but would potentially receive more profit. They were told that the AI's goal was to help them find

the optimal price so they would find a sublesser with a reasonable profit. To ensure performance and motivation, they were further informed that for good guessing they would be paid an additional bonus of 0.30 Dollars. However, every participant received the bonus, regardless of their performance. In order to better control price disparity between urban and rural regions, participants were asked to indicate the U.S. state they currently live in (e.g., Colorado) to ascertain their state capital (e.g., Denver). This not only made the guessing process easier for participants, but also made the AI more convincing since it was claimed that the AI would likewise base its price recommendations on data collected in that state capital. After an example that showed how the apartments and their amenities would be presented to them (see Figure 2), participants could start with the actual task. After completion, participants had to fill out questionnaires and give some demographic information. For ethical reasons, participants were debriefed about their deception at the end of the study and informed that the AI was a pseudo AI and did not actually use participants state capital for its recommendations.

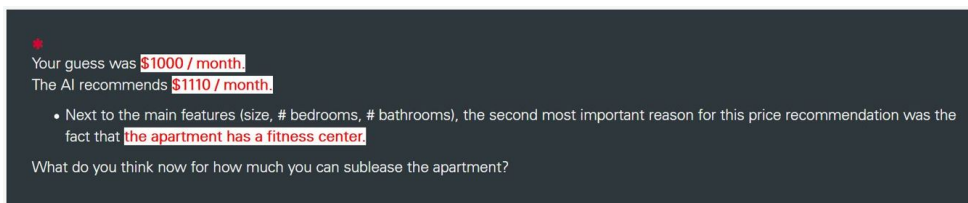
Stimuli

The apartments that participants had to evaluate were real apartments retrieved from the website Zillow (<http://zillow.com>) in May 2020. To create some variability, six different apartments of different sizes and consequently price ranges were selected: two small-sized apartments (500 - 750 square feet), two medium-sized apartments (751 - 1'000 square feet) and two large-sized apartments (1'001 - 1'250 square feet). Features and amenities were collected directly from the website Zillow, whenever available. If not available, a random value was chosen for continuous variables (e.g., distance to public transit from 0.1 - 2.0 miles) and a random choice for dichotomous variables was made (e.g., elevator YES / NO). All participants were presented with the same stimuli, that is, the same apartments and features. What differed was the price recommendation and the kind of explanation the recommendation was presented with (see Figure 3).

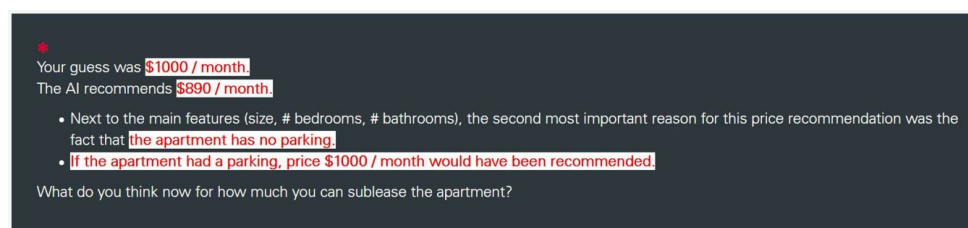
The price recommendation from the pseudo AI was designed so that a random number between 15 and 30 was picked. This random number was then transformed into



(a) Control.



(b) Feature Importance.



(c) Counterfactual.

Figure 3. Examples how the three different conditions (a) Control, (b) Feature Importance and (c) Counterfactuals were presented to the participants for three different apartments.

percentages and either added or subtracted to the initial subleasing price (T1), which always led to a random deviation between 15 and 30 percent. By applying this procedure, it was ensured that no participant could guess the price correctly, inasmuch as there was no 'correct price'. This has been a major limitation of past studies (Poursabzi-Sangdeh et al., 2018), since the interpretation of WOA becomes meaningless if T1 and T2 are equal to one another. It was randomly assigned that for three of the six apartments the recommendation was negative, meaning lower than the initially guessed subleasing price (e.g., if T1 was 1'000 and the random number 17, the AI recommendation was 830) and for the other three apartments positive, meaning higher than the initially guessed subleasing price. By doing this, the AI informed participants that their initial guesses were either too low or too high which made it possible to compare positive and negative recommendations.

Measures

Independent variables. The previously introduced explainability techniques *feature importance* and *counterfactuals*, as well as the *control condition*, served as independent variables. Figure 3 shows how the price recommendation was explained to the participants. The output was designed in such a way that the subsequent explanations can be seen as an extension of the preceding ones. To make the stimuli more convincing, the most relevant outputs were presented as if it was a console output.

Dependent variables. For objective trust behavior the introduced weight of advice from the advice-taking literature was used. It has the following notation:

$$\mathbf{WOA} = \frac{|T2 - T1|}{|P - T1|}$$

Here, P is defined as the model’s prediction, T1 is the participants’ initial prediction of the apartment’s price before seeing P, and T2 is the participants’ final prediction of the apartment’s price after seeing P. WOA measures the degree to which people update their beliefs and quantifies how much people weigh the received advice (i.e. the AI recommendation). To better understand this, consider the scenario in which T2 is closer to P compared to T1 — meaning that the participant made a significant update when estimating T2 based on the received explanations, towards P and away from T1. In this case, WOA is positive. WOA is equal to 1 if the participants’ final prediction matches the AI recommendation and equal to 0.5 if the participant averages their initial prediction with the AI recommendation P. WOA of 0 occurs when a participant ignores the AI recommendation (T1 = T2) and a negative WOA signifies that a participant discounts the recommendation completely and moves further away from the recommendation. WOA can be seen as percentages and this straightforward interpretation is a key advantage of this objective trust measurement.

As previously mentioned, no validated and established questionnaires exist that measure subjective trust in AI and algorithms. For this reason, the introduced *trust in automation scale* by (Jian et al., 2000) was used. Thus far, the questionnaire was cited in

309 studies and of those, at least 100 used the scale in its original form (Gutzwiller et al., 2019). Additionally, the *TrustDiff* (Brühlmann et al., in press) was utilized. Their proposed semantic differential for user trust on the web shares the trust dimensions *benevolence*, *competence* and *integrity* of the ABI+ framework by Toreini et al. (2020) for AI. Both scales range from 1 to 7, which simplifies comparison.

Data cleaning

After removing participants that did not fulfill the inclusion criteria (see Figure 1), participants identified as outliers were excluded. For WOA, outliers were defined in the following way:

1. Following prior research (Gino & Moore, 2007; Logg et al., 2019), all participants were excluded that showed unrealistic WOA's. In this case, an unrealistic WOA was defined as being < -1 (more than 100 percent discount of the recommendation) and > 2 (more than 100 percent overshoot of the recommendation).
2. Additionally, the interquartile rule was used to define outliers. By applying this rule, outliers are defined as observations that fall below $Q1 - 1.5$ interquartile range or above $Q3 + 1.5$ interquartile range.

By applying this approach, 65 participants were removed and 387 participants remained for final data analysis.

Results

Descriptive statistics

Table 1 shows participants' characteristics, as well as the descriptive statistics, split by condition. The sample was predominantly male (61%) and had an average age of 37 years ($M = 36.98$, $SD = 10.16$). A majority of the participants (68%) possessed a higher educational qualification (i.e. bachelor's degree, master's degree, Ph.D. or higher). See Table 1 for more detailed demographic information. On average, participants across all conditions approached the AI recommendation, resulting in a positive WOA ($M = 0.69$,

$SD = 0.36$). Overall, TrustDiff showed low average ratings ($M = 2.93$, $SD = 1.47$), while *benevolence* showed the highest ratings ($M = 3.63$, $SD = 1.64$) compared to the other two subscales *competence* ($M = 2.61$, $SD = 1.56$) and *integrity* ($M = 2.72$, $SD = 1.56$). The *trust in automation scale* showed higher overall ratings than TrustDiff ($M = 5.00$, $SD = 0.86$).

Table 1

Participants characteristics and descriptive statistics of mean (M), standard deviation (SD) and median (Mdn) for WOA, trust in automation and TrusDiff, split by condition.

Overall (n = 387)										
Gender	n	%	Age	M	Mdn	Range	Education	n	%	
Male	235	60.7		36.98	35.00	18 - 69	High School	36	9.3	
Female	148	38.2					College, no degree	83	21.4	
Non-binary	3	0.8					Bachelor's degree	194	50.1	
No Answer	1	0.3					Master's degree	66	17.1	
Total	387	100					Ph.D. or higher	3	0.8	
							Other	5	1.3	
							Total	387	100	

	Control (n = 133)			Feature Importance (n = 146)			Counterfactual (n = 108)			
	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	
Weight of advice										
Positive Recommendation	0.66	0.37	0.68	0.66	0.34	0.67	0.67	0.33	0.66	
Negative Recommendation	0.67	0.37	0.69	0.73	0.37	0.79	0.76	0.36	0.78	
TrustDiff (1 - 7)										
Overall	2.92	1.44	2.70	3.04	1.45	2.70	2.79	1.51	2.60	
Benevolence	3.47	1.64	3.67	3.78	1.53	4.00	3.61	1.78	3.67	
Competence	2.57	1.48	2.25	2.74	1.61	2.25	2.49	1.57	2.13	
Integrity	2.72	1.54	2.50	2.78	1.55	2.33	2.64	1.61	2.00	
Trust in automation (1 - 7)										
Overall	4.98	0.92	4.83	4.93	0.83	4.83	5.09	0.84	5.00	

Objective Trust - WOA

For WOA, residuals were checked for normal distribution via quantile-quantile plots (Q-Q plots), as well as if the residuals' variance was equal across groups (homoscedasticity). The normality assumption seemed satisfied and Levene's test indicated equal variances ($F = 0.61$, $p = .54$) that did not differ between groups. To address H_1 and H_2 , corresponding contrasts were created. The first contrast allowed the comparison if at least one of the two experimental conditions was significantly different from the control condition (planned contrast 1: control condition vs. experimental conditions for answering H_1). By defining another contrast, it was possible to test if the two experimental conditions were significantly different from one another (planned contrast 2: feature importance vs. counterfactual for answering H_2). The effect of the three conditions on WOA was analysed by employing linear mixed effect models (LMEMs), using the *lme4* package for R (Bates, Mächler, Bolker, & Walker, 2015). β -estimates, t -values, as well as their corresponding p -values and the .95 confidence interval (CI) are reported.

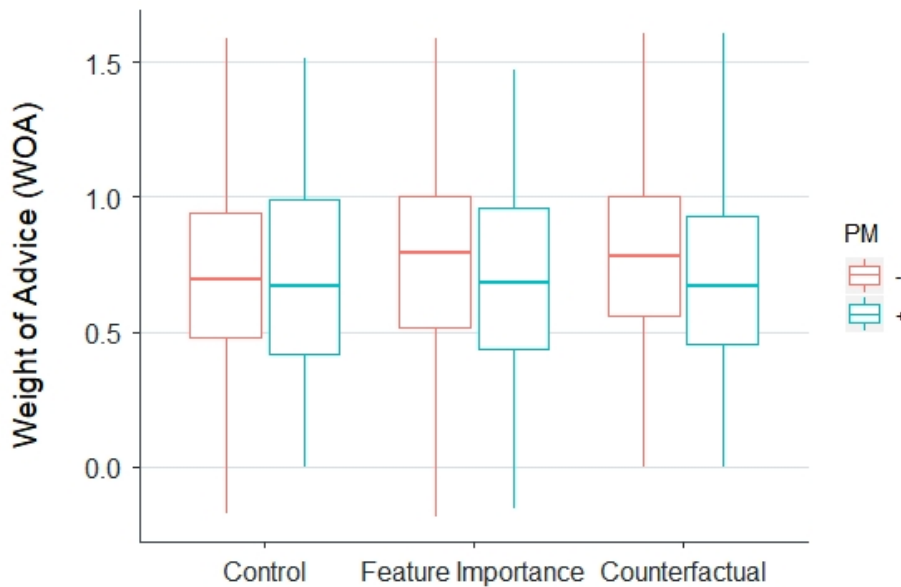


Figure 4. Boxplots for WOA's of the different conditions split according to positive and negative recommendations.

The first model contained two fixed main effects: The first contrast as well as the difference in the positive and negative recommendations (PM). Under the assumption that the stimuli and conditions have varying random effects for the different participants, a random intercept was additionally introduced in the model. Expressed in R-syntax, the model was defined as:

$$WOA \sim 1 + Contrast1 + PM + (1 | participants)$$

For this first model, the experimental conditions were not significantly different from the control ($\beta = 0.04$, $t = 1.39$, $CI = [-0.01, 0.08]$, $p = .17$), while the distinction between positive and negative recommendation was highly significant ($\beta = -0.05$, $t = -4.26$, $CI = [-0.08, -0.03]$, $p < .001$). See Figure 4 for the boxplots of the different conditions, split by positive and negative recommendations. Since the first contrast was not significant, further examination of the second contrast was redundant. To check whether there was an interaction effect between conditions and the negative and positive recommendations, a second model was defined:

$$WOA \sim 1 + Contrast1 * PM + (1 | participants)$$

This second model revealed that there was indeed a significant interaction effect ($\beta = -0.07$, $t = -2.64$, $CI = [-0.11, -0.03]$, $p = .01$). Comparing the two models confirmed that the inclusion of the interaction term in the model was justified since it significantly improved the model fit ($\chi^2(1) = 6.95$, $p = .01$). To better understand this relationship, an interaction plot was created (see Figure 5). The non-parallel lines indicate that the condition effect on WOA was different for positive and negative recommendations. Since the interaction effect was significant, the main effects of the second model can not be interpreted in a meaningful way. Depending on whether the recommendation was positive (the AI recommended a higher subleasing price) or negative (the AI recommended a lower subleasing price), the two explanation techniques had different effects on WOA. For positive recommendations, explanations had a negligible effect on WOA, yet for negative recommendations, the effect was substantial. Therefore, the effect of explanations can not be clearly understood without taking into account the differences in positive and negative recommendations.

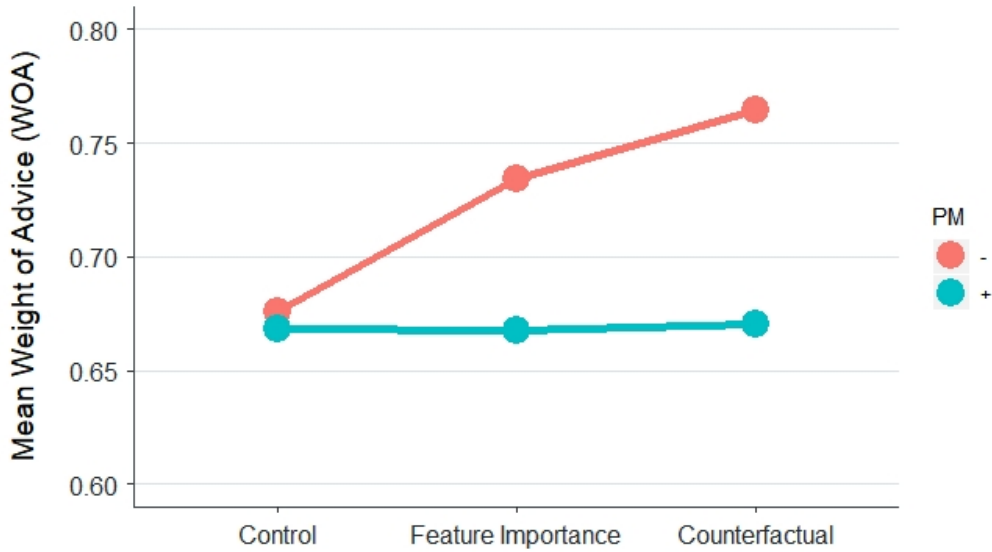


Figure 5. Interaction plot, capturing the interaction between the effects of conditions and positive and negative recommendations. Note that the y-axis is scaled to better visualize the effect.

To better understand the obtained results, positive and negative recommendations were controlled for by dividing the data into two subsets. One subset contained the three apartments with the positive recommendations, the other subset contained the three apartments with the negative recommendations. Following this, the first model was executed again for the two discrete subsets, thus omitting PM as a main effect. For negative recommendations, the first contrast was significant ($\beta = 0.07$, $t = 2.17$, $CI = [0.02, 0.13]$, $p = .03$), while the same was not true for the positive recommendations ($\beta < -0.001$, $t = -0.03$, $CI = [-0.06, 0.06]$, $p = .98$). This result implied that for negative recommendations, at least one of the two experimental conditions leads to significantly higher WOA compared to the control. To identify which one, a model without contrasts was defined that allowed the distinction between the experimental conditions.

$$WOA \sim 1 + Conditions + (1 | participants)$$

Running this model revealed that the condition counterfactual led to significantly higher WOA's compared to the control ($\beta = 0.09$, $t = 2.19$, $CI = [0.02, 0.15]$, $p =$

.03), while feature importance did not ($\beta = 0.06$, $t = 1.61$, $CI = [-0.01, 0.12]$, $p = .11$). The β -estimates indicated that on average, counterfactual explanations increased WOA by an approximated 9%. Note that feature importance explanations increased WOA by roughly 6%, but this difference was not significant for the .05 significance level. A further analysis with the second contrast (planned contrast 2: feature importance vs. counterfactual) revealed that the difference between the two experimental conditions was not significant ($\beta = -0.03$, $t = -0.71$, $CI = [-0.09, 0.04]$, $p = .48$).

Subjective Trust - Trust in automation scale

To test H_3 (the experimental conditions lead to higher subjective trust ratings compared to the control) and H_4 (counterfactuals lead to higher subjective trust ratings compared to feature importance), multiple one-way analysis of variance (ANOVAs) were intended. However, a visual inspection of the data using Q-Q plots indicated a non-trivial violation of assumption of normality. A subsequent Shapiro–Wilk test confirmed this presumption. Additionally, Levene’s test for homogeneity of variance between groups suggested additional violations of ANOVA’s assumptions ($F = 0.80$, $p = .45$). The ANOVA results might thus not be interpretable and meaningful. Under these circumstances, a Kruskal-Wallis test (Kruskal & Wallis, 1952) was carried out. Kruskal-Wallis test is a non-parametric alternative to ANOVA that does not assume a normal distribution of the residuals (Kruskal & Wallis, 1952).

The results showed that neither the ratings for the *trust in automation scale* ($H(2) = 1.76$, $p = .42$) nor the overall TrustDiff ($H(2) = 2.26$, $p = .32$), as well as its subscales *benevolence* ($H(2) = 2.62$, $p = .27$), *competence* ($H(2) = 2.10$, $p = .35$) and *integrity* ($H(2) = 1.28$, $p = .53$) were significantly different between conditions. Figure 6 captures the similar subjective trust ratings for the two experimental conditions and the control. It is interesting, however, that *benevolence* ranks higher than the other two subscales, *competence* and *integrity*. As mentioned in the descriptive statistics, the rating differences between the *trust in automation scale* ($M = 5.00$, $SD = 0.86$) and the TrustDiff ($M = 2.93$, $SD = 1.47$) are substantial.

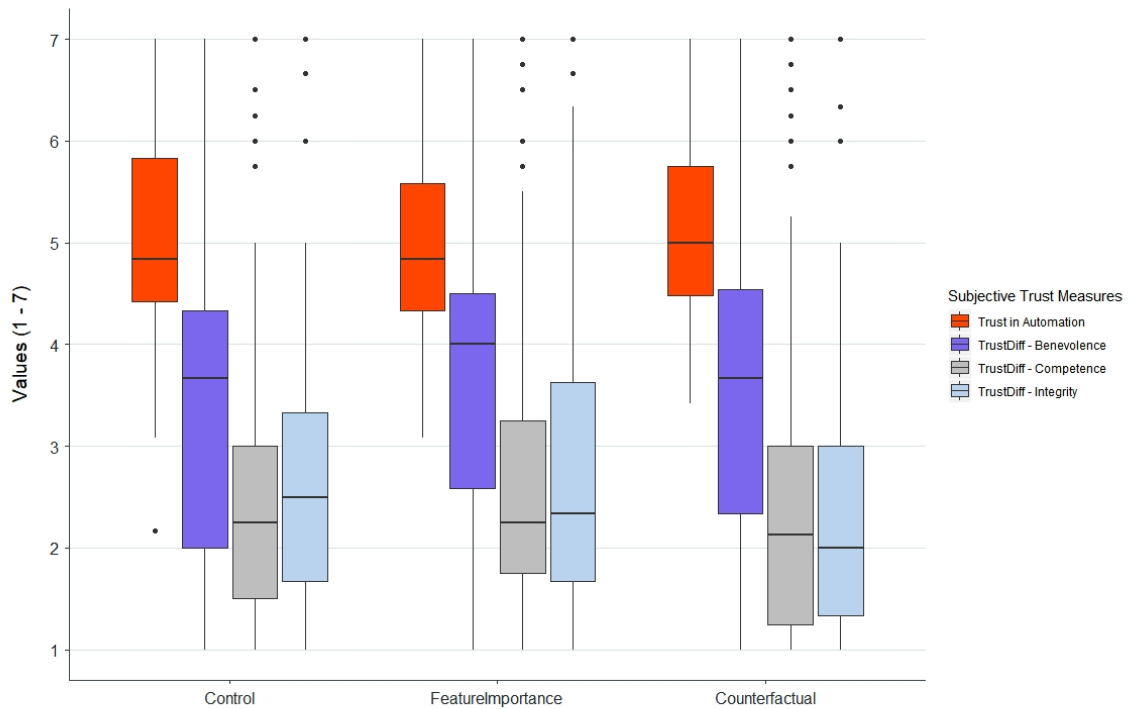


Figure 6. Boxplots of the ratings for the subjective trust measurements *trust in automation scale*, *TrustDiff* as well as its subscales *benevolence*, *competence* and *integrity* for each condition.

To evaluate the association between objective trust behavior and subjective trust ratings, a correlation analysis was performed. The assumption was that participants with high WOA scores also reported higher subjective trust ratings and vice versa. Because of the violation of normality assumption, Kendall's non-parametric rank-based correlation coefficient was applied. Kendall's tau statistics indicated a significant positive relationship between *trust in automation* and WOA scores ($r_{\tau}(385) = .07$, $p = .047$). However, according to Cohen (1988), the effect size of the relationship is low if the value of r_{τ} varies around 0.1 and squaring the correlation coefficients indicated that a mere 0.49% of the variance in WOA is explained by subjective trust. This positive relationship was strongest for feature importance ($r_{\tau}(144) = .10$, $p = .08$), followed by the control condition ($r_{\tau}(131) = .05$, $p = .37$) and counterfactuals ($r_{\tau}(106) = .05$, $p = .48$), while not being significant for all conditions. Surprisingly, for *TrustDiff*, the relationship changed

direction ($r_{\tau}(377) = -.10, p < .01$). This circumstance complicates the interpretation of the results. The negative relationship was again strongest for feature importance ($r_{\tau}(141) = -.15, p < .01$), this time, however, followed by counterfactuals ($r_{\tau}(104) = -.09, p = .16$) and the control condition ($r_{\tau}(128) = -.05, p = .37$). The varying degrees of freedom are caused by missing values in the TrustDiff, since participants could choose not to answer. To better understand the obtained results, these missing values were further examined. Overall, there were 319 missing values. Comparing this number with the total number of responses, that is the number of participants ($n = 387$), multiplied by the number of items ($n = 10$), roughly 8% of all TrustDiff responses were missing values. What was striking, however, is their uneven distribution. The subscale *benevolence* accounted for 53% of all missing values. See Table 2 for a comprehensive overview of the distribution of missing values for each subscale of the TrustDiff.

Table 2

Distribution of missing values ($n = 319$) for TrustDiff in the original enumeration by (Brühlmann et al., in press), expressed in whole numbers, as percentages of the total number of missing values, and corresponding Kendall's correlation coefficients for trust in automation and WOA.

TrustDiff item	Number of missing values	In percentage (%)	Kendall's tau TIA	Kendall's tau WOA
1 (<i>Benevolence</i>)	57	18	.03	-.08*
2 (<i>Benevolence</i>)	55	17	.07	-.11**
3 (<i>Benevolence</i>)	57	18	-.02	-.05
4 (<i>Integrity</i>)	33	10	-.21***	-.09*
5 (<i>Integrity</i>)	16	5	-.29***	-.11**
6 (<i>Integrity</i>)	30	9	-.26***	-.11**
7 (<i>Competence</i>)	13	4	-.27***	-.15***
8 (<i>Competence</i>)	18	6	-.34***	-.10*
9 (<i>Competence</i>)	15	5	-.33***	-.11**
10 (<i>Competence</i>)	25	8	-.29***	-.11**
Total	319	100	–	–

Note. * $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$. TIA = trust in automation.

Finally, the correlations on the TrustDiff’s item level with *trust in automation* and WOA was examined. Interestingly, for the relationship between *trust in automation* and two of the three items of the subscale *benevolence*, Kendall’s tau was positive. All other items showed a significant negative correlation. The overall negative relationship between TrustDiff and WOA was reconfirmed by the exclusively negative correlation on the item level (see Table 2).

Discussion

This study was able to show that participants generally accept AI recommendations in low-stake decision making processes — in this case, receiving AI advice to find an optimal apartment subleasing price. Regardless of the different conditions, participants on average showed high overall WOA ($M = 0.69$, $SD = 0.36$). A WOA of 0.69 implies that participants adopted nearly 70% of the AI recommendations when updating their prior beliefs. This finding supports the idea of people displaying *algorithm appreciation* in decision-making processes (Logg et al., 2019). Finding an optimal apartment subleasing price is a more objective domain than AI recommendations was studied in the past. XAI and HCI researchers often focus on subjective domains like joke recommendations (Yeomans et al., 2019) and forecasting romantic partners (Yin et al., 2019).

It was further demonstrated that under certain conditions, human-friendly explanations have a significant effect on objective trust behavior, captured by weight of advice. Our findings therefore support H_1 , that the experimental conditions lead to higher objective trust scores compared to the control. However, the effect of explanations depended on the nature of the decision that participants had to make. Participants were presented two types of recommendations in a randomized order — for the first type, the pseudo AI recommended participants to *increase* their initial guess for the apartment prices (referred to as positive recommendations in this thesis) as to potentially receive more money. For the second type, participants were advised to *decrease* their initial price (referred to as negative recommendations in this thesis) and hence potentially earn less money. The results of the experiment indicate that for positive recommendations, different explana-

tions techniques had no effect on WOA. Contrarily, for negative recommendations, there was a significant effect of explanations on WOA, meaning that participants in the experimental conditions updated their initial guess and approached the AI recommendation up to 9% more than participants in the control. This seems counterintuitive at first glance, since one might expect that participants would always choose to embrace the prospect of obtaining a higher subleasing price. It is conceivable, however, that the two types of recommendations can be thought of as two distinct decision-making processes. The well-studied concept of *loss aversion* by Tversky and Kahneman (1991) could account for this discrepancy and serve as an explanation attempt for these findings. Loss aversion suggests that, psychologically speaking, losses are twice as powerful than gains because people assign more utility to losses than to gains (Tversky & Kahneman, 1991). In practical terms, this means that a person who *loses* \$100 will subjectively experience more dissatisfaction than another person would experience satisfaction from *gaining* \$100. The study design seems to satisfy the preconditions for a possible loss aversion effect, since it accounts for positive and negative recommendations. When participants were advised by the AI to increase their initial price guess, it is likely that they were concerned that this potential price raise would cause an unsuccessful sublease. The prospect of getting more money (gain) mattered less in this decision-making process than the possibility of not being able to sublease at all (loss). When faced with loss aversion, the explanations from the pseudo AI might not be convincing enough to overcome the participants' higher assigned utility to losses. Within the framework of loss aversion, a recommendation from the AI to decrease the initial price was a different kind of decision-making process. In this case, no loss aversion was induced and participants were encouraged by the AI explanations, that demanding less money was the right decision to successfully sublease the apartment. Naturally, a price reduction also leads to small losses, but compared to the possibility of not being able to sublease at all, these smaller losses might be negligible. When not being faced with loss aversion, AI explanations seem to convince people to adjust their initial sublease price, compared to the control where no additional explanation was present. What is different from the classical loss aversion scenario, is that no

concrete possibilities were assigned to the two potential outcomes. Participants had no indication of how their decision might affect the likelihood of a successful subleasing.

The interpretation under consideration of loss aversion is speculative, but the findings indicate that trust behavior induced by explanations occurs only in certain cases and that people might demonstrate certain cognitive biases, here loss aversion. With the notion of human-friendly explanations by Miller (2019) in mind, the findings suggest that in addition to explanations being *selective*, *contrastive*, *non-probabilistic* and *social*, explanations ought to be *case-specific*. Case-specific signifies that depending on the individual case, a specific explanation that accounts for potential cognitive biases in this circumstance may be more human-friendly than another explanation. An example with regard to this study could be that in the case of negative recommendations that do not induce loss aversion, feature importance and counterfactuals seem sufficiently comprehensible explanations for increased trust behavior. However, in the case of positive recommendations that induce loss aversion, those explanations do not seem to convince people and in this case, other explanations might be more persuasive.

Other possible explanations for the findings of our study are conceivable. An alternative reason for the obtained results could be that participants already maximized the expected price when making an initial guess (T1) and that explanations accompanying the AI recommendation were not convincing enough for participants to further *increase* the subleasing price for their final guess (T2). *Decreasing* the subleasing price, however, was not in conflict with the maximized expected profit. The second hypothesis H_2 , that counterfactuals lead to higher objective trust scores compared to feature importance, can not be supported by these results. Although counterfactuals increase WOA nearly 3% more than feature importance for negative recommendations, this difference was not significant under the .05 significance level.

For subjectively stated trust, the results showed no significant differences between conditions for the *trust in automation scale*, the overall TrustDiff, as well as its subscales *benevolence*, *competence* and *integrity*. The results therefore do not support H_3 , which states that the experimental conditions lead to higher subjective trust ratings compared

to the control. Neither H_4 can be supported, i.e., that counterfactuals lead to higher subjective trust ratings compared to feature importance. The findings, however, are not conclusive. While the *trust in automation scale* indicated a relatively high overall trust score ($M = 5.00$, $SD = 0.86$), the overall score for TrustDiff was low ($M = 2.93$, $SD = 1.47$). Examining the correlations between WOA and the *trust in automation scale* ratings, as well as the TrustDiff ratings, revealed further contradictions. The expected result was that objectively observed trust would be reflected in subjectively expressed trust – i.e., the higher the observed trust captured by WOA, the higher the subjectively expressed trust which was assessed by the trust questionnaires. However, while Kendall’s correlation coefficient was positive for *trust in automation* and WOA, the direction changed for the relationship between TrustDiff and WOA. According to Cohen (1988), all correlations were associated with low effect sizes that explained only a fraction of the variance of objective trust, captured by WOA. Analyzing the TrustDiff revealed that there was a relatively high number of missing answers of 8%, and a more thorough investigation confirmed that the subscale *benevolence* was conspicuous in several ways. To begin with, *benevolence* had the highest number of missing values, accounting for more than half of all missing values in the TrustDiff. Since the TrustDiff was designed for evaluating user trust in online services and websites, it is not surprising that the transfer to trust in AI might not be straightforward. The *benevolence* differentials *insensitive* vs. *sensitive*, *ignoring* vs. *caring* and *inconsiderate* vs. *empathic* may not be applicable to AI in particular, since these differentials in their original form apply to the human developers or operators of online content. When applied to AI, those differentials possibly lead to anthropomorphism, the attribution of human traits to non-human entities, which has been demonstrated to make people feel uncomfortable (MacDorman, Green, Ho, & Koch, 2009). This could be an explanation for the high number of missing values for *benevolence* and resembles the argumentation of Hartmann (2020), that it might be unreasonable to apply certain aspects of trust when referring to interactions between humans and AI. Additionally, *benevolence* was the subscale with the highest rating of the TrustDiff ($M = 3.63$, $SD = 1.64$), being more similar to the rating of the *trust in automation scale*. This

similarity was also evident when looking at Kendall's correlation coefficients between the TrustDiff items and the overall score for *trust in automation*. The *benevolence* items were the only items that did not show a significant negative correlation with the overall score of the *trust in automation scale*. This might imply that the *trust in automation scale* and the subscale *benevolence* cover similar aspects of subjective trust, as opposed to the subscales *integrity* and *competence*. The obtained results give rise to doubts if the ABI+ framework by Sanders et al. (2006) is a promising candidate for trust evaluation in AI, since the TrustDiff should capture its proposed trust dimensions *benevolence* and *integrity*. Alternatively, the *trust in automation scale* might not capture subjectively stated trust accurately. Whatever the reasons for this discrepancy may be, it seems that objectively measured trust behavior and subjectively stated trust are not fully transferable to each other. While explanations can lead to higher objective trust scores, the same is not true for subjective trust ratings. Unfortunately, for subjective trust measures, it was not possible to differentiate between positive and negative recommendations as it was for WOA. It would have been interesting to see if the effects of explanations for negative recommendations are likewise observable in the subjective trust measures. The pressing question how trust behavior and the perception of trust relate to one another and if human-friendly explanations effectively promote genuine trust thus remains.

Limitations and future work

This study has certain limitations that future work should consider. Firstly, this study used a pseudo AI and not a real AI to recommend participants a specific price. While a screening of the open-ended questions did not show any critical comments addressing this, it was not properly controlled if the deception worked. However, by removing WOA outliers, participants were excluded that indicated unreasonable values. Next, these findings only apply to low-stake decisions, including small and fictive amounts of money. As mentioned earlier, a core assumption within the trust literature is that trust contains some sort of risk-taking behavior and vulnerability (Hoff & Bashir, 2015). While our study design encouraged participants to perform well to receive an additional mon-

etary bonus, future studies should conduct more realistic experiments, where a more tangible loss depends on the participants' decision to trust AI. Those studies should also focus on domains other than apartment prices to investigate if the findings of this study are transferable to different scenarios. Another methodological limitation was the small number of evaluated apartments and that the positive and negative AI recommendations were apartment-dependent, meaning that it was arbitrarily defined for which apartments the AI would always advise increasing or decreasing the initial price. Creating a pool of apartments, randomly picking some apartments out of it and assigning either positive or negative recommendations would have been a more reasonable methodological approach. By doing so, it would have been possible to differentiate between positive and negative recommendations for subjective trust measures, as was the case with WOA. Also, a high number of MTurk participants did not satisfy the quality checks and had to be removed ($n = 310$). Due to those exclusions, the originally envisaged sample size ($n = 480$) was not achieved. It's conceivable that with an adequate sample size, the effect of feature importance would have turned out to be significant. On the positive side, the fact that there were nevertheless significant differences indicates that the effect size is larger than expected. Lastly, the study design does not distinguish between *dispositional trust*, *situational trust* and *learned trust*, as suggested by Hoff and Bashir (2015). It is recommended that future studies investigate those varying manifestations of trust. Researchers could measure user trust before participants are exposed to an AI system and compare it with the stated trust that participants report afterwards (learned trust). Or they could expose participants to AI recommendations while inducing different emotional valence (situational trust). Further factors such as participants' housing situation and task familiarity that could also have an effect on subjective trust behavior were not considered in this study.

Conclusion

The findings provide evidence that in some cases, transparency in the form of human-friendly explanations can enhance objectively observed trust behavior in end-users. Explainability as opposed to interpretability appears to be a real option for transparent AI that is both human-friendly and meets the legal GDPR requirements, while avoiding the major disadvantages of interpretability. It was demonstrated that depending on the type of decision-making process, it matters what kind of explanation AI provide. More specifically, only in cases where an AI recommended to decrease the price (negative recommendations), explanations led to increased trust behavior, captured by weight of advice. This issue of case-specificity has been given little attention in the current HCI and xAI literature. People seem to retain certain cognitive biases in their decision-making processes involving AI. In this study, loss aversion was introduced to explain the differences in objective trust behavior, but other biases and heuristics might also play a role. Future experiments should include means of validating that these cognitive biases are indeed taking place and AI explanations must account for those when expected to be human-friendly. Ideally, future AI might provide explanations, perfectly tailored to a specific decision-making process. Researchers as well as practitioners must be aware of this when designing AI, since they are not inherent in the systems themselves. Referring to Miller's human-friendly explanations, the findings suggest that in addition to explanations being *selective*, *contrastive*, *non-probabilistic* and *social*, explanations should also be *case-specific* as design principle for human-friendly AI explanations. This is important, since more and more AI systems are deployed in real-world applications and for a successful adaption, human biases in decision making processes must be accounted for. A behavior change of 8%, induced by explanations, may not appear to amount to much, but in large-scale scenarios it could add up over time and be of real significance in an increasingly automated world.

The relationship between objective and subjective trust, however, remains ambiguous. While participants in this study generally showed high objective trust behavior with AI advice adoption of nearly 70%, the results for subjective trust are inconclusive. It is

also unresolved whether people subjectively experience trust when they show trust behavior. The behavior change might be induced by blind-trust or other subjective experiences like uncertainty, ignorance or mental overload. These findings can not answer conclusively if there exists a measurement problem when it comes to assessing subjective trust via questionnaires. I suggest that more work needs to be invested in validated AI trust questionnaires, especially designed for end-users. Such questionnaires could ensure that only genuine trustworthy AI are being employed in decision-making processes that tangent humans. Transparency and trust in AI just began to appear on the HCI landscape and more research should be conducted in the area to aim for human-friendly AI.

References

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–18). New York, USA: ACM Press. doi: 10.1145/3173574.3174156
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. *IEEE Access*, *6*, 52138–52160. doi: 10.1109/ACCESS.2018.2870052
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. doi: 10.1177/1461444816676645
- Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., . . . Wells, S. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine*, *37*(4), 76–83. doi: 10.1109/MTS.2018.2876107
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01
- Brühlmann, F., Petralito, S., Rieser, D. C., Aeschbach, L. F., & Opwis, K. (in press). Trustdiff: Development and validation of a semantic differential for user trust on the web. *Journal of Usability Studies*.
- Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In W. Fu, S. Pan, O. Brdiczka, P. Chau, & G. Calvary (Eds.), *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 258–262). New York, USA: ACM Press. doi: 10.1145/3301275.3302289
- Chopra, K., & Wallace, W. A. (2003). Trust in electronic environments. In *Proceedings of the 36th annual hawaii international conference on system sciences* (pp. 10–15). Los Alamitos, USA: IEEE Computer Society Press. doi: 10.1109/HICSS.2003.1174902

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, USA: Lawrence Erlbaum Associates.
- Copeland, J. (2015). *Artificial intelligence: A philosophical introduction*. Hoboken, USA: John Wiley & Sons.
- Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *International journal of human-computer studies*, 58(6), 737–758. doi: 10.1016/S1071-5819(03)00041-7
- De Graaf, M., & Malle, B. (2018). People’s judgments of human and robot behaviors: A robust set of behaviors and some discrepancies. In *Companion of the international conference on human-robot interaction* (pp. 97–98). New York, USA: ACM Press. doi: 10.1145/3173386.3177051
- Dhurandhar, A., Iyengar, V., Luss, R., & Shanmugam, K. (2018). *Tip: Typifying the interpretability of procedures*. ArXiv. Retrieved from <https://arxiv.org/abs/1706.02952>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. doi: 10.1037/xge0000033
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O’Brien, D., . . . Wood, A. (2017). *Accountability of ai under the law: The role of explanation*. SSRN Electronic Journal. doi: 10.2139/ssrn.3064761
- Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *41st international convention on information and communication technology, electronics and microelectronics* (pp. 210–215). Los Alamitos, USA: IEEE Computer Society Press. doi: 10.23919/MIPRO.2018.8400040
- European Parliament, C. o. E. U. (2018). *Eu general data protection regulation (gdpr). regulation 2016/679 §§71-12-14*. Retrieved from <https://gdpr-info.eu/>
- Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. Cambridge, USA: MIT Press.
- Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of*

- Behavioral Decision Making*, 20(1), 21–35. doi: 10.1002/bdm.539
- Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., & Hsiung, C.-P. (2019). Positive bias in the ‘trust in automated systems survey’? an examination of the jian et al.(2000) scale. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 217–221). Los Angeles, USA: SAGE Publications. doi: 10.1177/1071181319631201
- Hartmann, M. (2020). *Vertrauen – die unsichtbare macht*. Berlin, Germany: Fischer Verlag.
- Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, 70(2), 117–133. doi: 10.1006/obhd.1997.2697
- Hoff, K., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors The Journal of the Human Factors and Ergonomics Society*, 57(3), 407-434. doi: 10.1177/0018720814547570
- Jian, J.-Y., Bisantz, A., & Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71. doi: 10.1207/S15327566IJCE0401_04
- Kizilcec, R. F. (2016). How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2390–2395). New York, USA: ACM Press. doi: 10.1145/2858036.2858402
- Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions. In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 5686–5697). New York, USA: ACM Press. doi: 10.1145/2858036.2858529
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260), 583–621. doi: 10.1080/01621459.1952.10483441
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the

- concept of interpretability is both important and slippery. *Communications of the ACM*, *16*(3), 31–57. doi: 10.1145/3236386.3241340
- Lisman, J. E., & Idiart, M. A. (1995). Storage of 7+/-2 short-term memories in oscillatory subcycles. *Science*, *267*(5203), 1512–1515. doi: 10.1126/science.7878473
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. doi: 10.1016/j.obhdp.2018.12.005
- MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? uncanny responses to computer generated faces. *Computers in human behavior*, *25*(3), 695–710. doi: 10.1016/j.chb.2008.12.026
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734. doi: 10.2307/258792
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. doi: 10.1016/j.artint.2018.07.007
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279–288). New York, USA: ACM Press. doi: 10.1145/3287560.3287574
- Mohseni, S., Zarei, N., & Ragan, E. D. (2020). *A survey of evaluation methods and measures for interpretable machine learning*. ArXiv. Retrieved from <https://arxiv.org/abs/1811.11839>
- Molnar, C. (2019). *Interpretable machine learning*. Retrieved from <https://christophm.github.io/interpretable-ml-book/>
- Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). *How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation*. ArXiv. Retrieved from <https://arxiv.org/abs/1802.00682>
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220. doi: 10.1037/1089-2680.2.2.175
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H.

- (2018). *Manipulating and measuring model interpretability*. ArXiv. Retrieved from <https://arxiv.org/abs/1802.07810>
- Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13). New York, USA: ACM Press. doi: 10.1145/3173574.3173677
- Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In E. H. et al. (Ed.), *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Berlin, Germany: Springer. doi: 10.1007/978-3-319-98131-4_2
- Ruben, D. H. (2015). *Explaining explanation*. London, England: Routledge.
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable ai: interpreting, explaining and visualizing deep learning*. Berlin, Germany: Springer.
- Sanders, K., Schyns, B., Dietz, G., & Den Hartog, D. N. (2006). Measuring trust inside organisations. *Personnel Review*, *35*(3), 557–588. doi: 10.1108/00483480610682299
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. doi: 10.1214/10-STS330
- Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 1335–1339). Los Angeles, USA: SAGE Publications. doi: 10.1177/154193120805201907
- Springer, A., Hollis, V., & Whittaker, S. (in press). Dice in the black box: User experiences with an inscrutable algorithm. *AAAI*. Retrieved from <https://arxiv.org/abs/1812.03219>
- Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., ... Zhou, J. (2019). Seven hci grand challenges. *International Journal of Human-Computer Interaction*, *35*(14), 1229–1269. doi: 10.1080/10447318.2019.1619259
- Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel,

- A. (2020). The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 272–283). New York, USA: ACM Press. doi: 10.1145/3351095.3372834
- Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, 106(4), 1039–1061. doi: 10.2307/2937956
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76–99. doi: 10.2139/ssrn.2903469
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard journal of law technology*, 31, 841–887. doi: 10.2139/ssrn.3063289
- Westfall, J. (2015). *Pangea: Power analysis for general anova designs*. Retrieved from <http://jakewestfall.org/publications/pangea.pdf>
- Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4), 403–414. doi: 10.1002/bdm.2118
- Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12). New York, USA: ACM Press. doi: 10.1145/3290605.3300509
- Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics. In G. A. Papadopoulos, T. Kuflik, F. Chen, C. Duarte, & W.-T. Fu (Eds.), *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 307–317). New York, USA: ACM Press. doi: 10.1145/3025171.3025219