I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Malin Eiband
LMU Munich
Munich, Germany
malin.eiband@ifi.lmu.de

Felicitas Buchner
LMU Munich
Munich, Germany
felicitas.buchner@campus.lmu.de

Adrian Krüger
LMU Munich
Munich, Germany
adrian.krueger@campus.lmu.de

Andreas Butz
LMU Munich
Munich, Germany
butz@ifi.lmu.de

ABSTRACT

Unintended consequences of deployed AI systems fueled the call for more interpretability in AI systems. Often explainable AI (XAI) systems provide users with simplifying local explanations for individual predictions but leave it up to them to construct a global understanding of the model behavior. In this work, we examine if non-technical users of XAI fall for an illusion of explanatory depth when interpreting additive local explanations. We applied a mixed methods approach consisting of a moderated study with 40 participants and an unmoderated study with 107 crowd workers using a spreadsheet-like explanation interface based on the SHAP framework. We observed what non-technical users do to form their mental models of global AI model behavior from local explanations and how their perception of understanding decreases when it is examined.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; *User studies.*

KEYWORDS

explainable AI; Shapley explanation; cognitive bias; understanding

ACM Reference Format:
Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450644

1 INTRODUCTION

There is a growing awareness that machine learning-based *intelligent systems* (IS) need to be capable of explaining their behavior in human-understandable terms to prevent unintended consequences in sensitive contexts of society (e.g., credit scoring, recruiting, predictive policing, or criminal justice) [18]. Driven by this concern, the field of *explainable artificial intelligence* (XAI) develops models, methods, and explainable interfaces that are interpretable to human users by providing some notion of explanation [16]. Organizations aspire to deploy explainability techniques to wider non-technical audiences to comply with demands and regulations [5]. Such users of XAI, also referred to as *operators* or *executers* [56], consume machine learning (ML) predictions to inform their decisions. They are centered between the developers and the individuals affected by the predictions [56]. Because they may be accountable for their decisions, they utilize explanations to assure the underlying models is *trustworthy* (i.e., *"they can reasonably trust a model's outputs"* [5]) (*operator-interpretability* [56]).

Many empirical XAI studies limit their explanation approaches to *outcome explanations* [21] for individual ML predictions (*local* explainability) without examining if users build an accurate mental model of the overall ML model behavior (*global explainability*). Local explanations based on *Shapley values* [51] are widely used in practice [5]. For a single observation, they perfectly distribute the difference between the average prediction and the actual prediction between its features [30]. Thus, much of the inherent ML model complexity (e.g., feature interactions) is simplified into accessible Shapley values [20]. Relying on them alone might leave users with a false sense of understanding that is merely illusive. Further, the explainability of explanations is often assessed through subjective user ratings [41]. In this type of evaluation, users are asked to report their perceived understanding, trust, or other relevant mental factors through one-shot ratings with little to no incentives for self-reflection or self-calibration [34, 35]. It has been shown, however, that people are *"often miscalibrated about their own judgments"* [35]. Psychological research has demonstrated in many contexts that humans have a robust bias of overconfidence regarding their understanding of how complex concepts work [46]. After being asked to explicate and actively reflect on their understanding, people significantly reduce their estimation of their own knowledge.

In this paper, we argue that because of this *illusion of explanatory depth* (IOED) [46], XAI explanations (especially in the form of additive local explanations for individual predictions) may be misleading for non-technical XAI users. Rather than stipulating effective gains in human understanding, they might cause them

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450644

to form false or incomplete beliefs about the explained ML model. Some researchers already speculated that an IOED could be at play when users are confronted with XAI explanations [14, 35, 54]. To our knowledge, however, this has not yet been systematically investigated. We examine if end users fall for an IOED when consuming XAI explanations in a decision-support scenario. In particular, we focus on the effect of *local post-hoc* explanations using *Shapley* values. We conducted human-grounded evaluations [16] in a *crowd lending* scenario using a tabular real-world data set. The scenario leverages a functional black-box ML model (a random forest classifier) and functional Shapley explanations generated by the widely-used explainability framework *SHAP* [30][1]. We followed a mixed methods approach. First, we moderated 40 participants through the study and observed their interactions. Second, we verified our hypotheses in an unmoderated study with 107 crowd workers. The studies has been approved by our internal IRB.

With our work, we follow the call to improve the user experience of XAI for a wider range of stakeholders [6]. The majority of current XAI research targets ML experts (e.g., data scientists) [25] or specific domain experts (e.g., physicians) [2, 15]. In contrast, we focus on the understanding of users with low expertise in AI. Our work contributes to the HCI community in three ways: First, we present *SHAPTable* an explanation interfaces targeting end non-technical users of XAI systems that embeds Shapley explanations in an accessible spreadsheet-like user interface (section 4). Second, based on an empirical examination we show that non-technical users fall for an IOED when relying on Shapley explanations (section 6).

2 BACKGROUND AND RELATED WORK

2.1 Explanations from Intelligent Systems

The research field of XAI aims to make black-box ML models interpretable by generating some notion of explanation that can be used by humans to interpret the behavior of an ML model [58]. An ML model is considered as a black-box if humans can observe the inputs and outputs of the model but have difficulties understanding the mapping between them. This may result from the model either being too complex, such as many deep neural networks, or being proprietary, such as with the COMPAS system [47]. Black-box models are often reported to yield a high *predictive accuracy* with less effort [47]. There are two broad categories of explainability approaches: *transparency-based* and *post-hoc* explainability [29]. *Transparency-based* approaches focus on how the model works and leverage model characteristics to explain it. This may involve using simpler models with intrinsic explainability that may yield a lower predictive accuracy. In contrast, *post-hoc* approaches ignore model characteristics. Instead, they observe the inputs and outputs of the ML model and try to detect regularities in its behavior in an inductive manner. Thus, post-hoc approaches have no impact on the predictive accuracy of a model but may oversimplify the true model behavior. The ability of an explanation method to accurately describe the behavior of an ML model is referred to as *descriptive accuracy* [36] or *fidelity* [47]. Human understanding in XAI can be fostered either by offering means of introspection or through explanations [7]. A large variety of methods exist for both approaches [21]. XAI research distinguishes two types of explanations - local and global [2, 21]. *Local* explanations of an ML model explain why an individual model prediction was made. In contrast, *global* explanations aim to convey the overall structure of the model by looking at model predictions on an aggregated level. Some definitions of explainability are rather *system-centric*. Doshi-Velez and Kim [16] describe it as a model's *"ability to explain or to present in understandable terms to a human."* Miller [32] takes a more *human-centered* perspective calling it *"the degree to which an observer can understand the cause of a decision"*. For an explanation to be effective, it does not only need to have a sufficient level of fidelity but must *"provide insight for a particular audience into a chosen domain problem"* [36].

2.2 Illusion of Explanatory Depth

Insights emerge when humans gain *"a clear, deep [...] understanding of a complicated problem or situation"*[2]. Human understanding, however, is often impacted by various cognitive biases. Research in cognitive sciences showed that people often form an inaccurate understanding of complex systems and often overrate the depth of their knowledge [35]. Rozenblit and Keil coined this type of overconfidence bias as the *illusion of explanatory depth (IOED)* [46]. They observed that laypeople consistently reduced the estimation of their own knowledge of different phenomena or devices after they were inquired to provide explanations about them or apply their understanding. Furthermore, people are often surprised by their limited explanations [4]. The IOED is more pronounced for *explanatory knowledge*, i.e., knowledge that involves complex causal patterns, than it is for *descriptive knowledge*, i.e., knowledge about facts (names of capitals), procedures (baking), or narratives (movie plots) [28, 46]. The IOED has first been demonstrated for people's understanding of causally complex systems in mechanical (bicycles, crossbows) [28, 33, 46] and natural (tides, rainbows) [46] domains. Subsequent work reproduced the IOED for social and policy domains (voting, mental disorder) [4, 60].

The illusion is believed to be caused by the way humans build their *conceptual knowledge*. Conceptual knowledge refers to the entirety of a person's *concepts* that are causally related to each other. According to the *theory-based* approach, people form *theories* about all their concepts, not just for those that they use regularly [46]. For instance, people form their own theories of what causes volcanic eruptions or how AI-based systems derive their predictions even though they were never confronted with one. These theories often consist of vague explanations that are not necessarily accurate nor coherent with each other [37]. When inquired to explicate parts of our conceptual knowledge to ourselves or others, we fall for the illusion to think we know more about a system than we actually do. Four factors are believed to influence the emergence of an IOED [46]: (i) *Representation/recovery confusion*: We overestimate our abilities to remember what we have observed. People tend to store observations as mental images. If the stored mental images do not correspond to the original facts, the IOED occurs. (ii) *Label/mechanism confusion:* Most complex systems are hierarchical with various levels of sub components. If we can name and describe individual parts on the first level of the hierarchy, we often assume

[1] https://github.com/slundberg/shap

[2] https://dictionary.cambridge.org/dictionary/english/insight

to understand how the overall system works, even though we have little insight into the levels further down the hierarchy. (iii) *Undefined end states*: Because of the hierarchical and related structure of complex topics, we have difficulties to imagine what constitutes a good and complete understanding or explanation. The end states for descriptive knowledge about facts or procedures are much clearer (e.g., naming the capital of a country or reverse engineering how to book a flight). (iv) *Lack of practice*: in everyday life most people regularly retrieve facts or reconstruct procedures. However, many people lack the practice of giving an explanation of complex topics. Just because we consume or make up explanations does not mean that we can produce effective explanations when needed.

2.3 IOED and Cognitive Biases in XAI and IS

Building on the IOED theories, it can be assumed that users of XAI systems form their own theories about the global behavior of the underlying ML model during interaction with the explanation facility. These also overlap with the widespread HCI concept of mental models. According to Norman, people form theories about how objects and systems work to explain what they observe [39]. A *mental model* refers to a person's understanding of how a system works and how the person's behavior affects it. People form mental models for all kinds of systems including objects, people, and services. The respective mental model is adjusted with every interaction (e.g., exposure to an XAI explanation) and helps the person to reflect on their belief about the system (e.g., the ML model behavior) [39].

Little research has been published on a potential IOED in the context of XAI or IS. Some researchers speculated that an IOED may be at play when users deal with explanations from XAI systems [14, 35, 54]. Collaris et al. observed during their XAI evaluation that their users did not question the validity of local explanations, even when provoked to do so [14]. Sokol and Flach call for an XAI validation protocol that addresses the IOED [54]. Kaur et al. observed that even data scientists and ML engineers took visual explanations of interpretability tools at face value and missed to effectively use them to uncover data or model issues. The provided XAI explanations encouraged the users to apply their heuristic thinking instead of activating their analytical thinking [25]. Even though the IOED itself received little attention in the context of intelligent systems and XAI, there is prior research on cognitive biases of explanations from intelligent systems investigating automation [9, 31, 40, 49, 52], anchoring [19, 27, 57], framing [26], and confirmation biases [23, 55]. A cognitive bias related to the IOED is the Dunning-Kruger effect of illusory overconfidence, which states that people with low competence at a given task tend to overestimate their task performance [49]. It occurs only with individuals with low competence while the IOED affects almost everyone.

3 RESEARCH QUESTIONS AND HYPOTHESES

Our work investigates the formation and the accuracy of operators' understanding of the ML model behavior from Shapley based local explanations. **The overarching research question (ORQ) of our work is to examine whether non-technical users of such XAI systems are prone to an IOED**. It is driven by the following research questions:

- **RQ1**: How robust is a self-reported global understanding gained from local explanations when examined?
  - H1: When participants are exposed to local explanations, this leads to an increased perception of understanding how the XAI system works (compared to no explanations)
  - H2: The participants' perception of understanding decreases after they have been examined for their understanding (IOED applies)
- **RQ2**: What do non-technical XAI users do to construct a global understanding from local explanations?

We focus on Shapley based explanations because, despite their vulnerability to adversarial attacks [53] and potential infidelity [20], we consider them as relevant for end users for two reasons: (i) enabled by the mathematical properties of accuracy and consistency, multiple local explanations can be combined to be contrastive and counterfactual [43] as well as interactive [13], (ii) the *SHAP* framework is widely used by XAI practitioners[3] and thus end users will likely come across Shapley based explanations, (iii) model agnostic approaches allow system designers to offer uniform explanation interfaces even when the underlying ML models differ.

However, human cognition is biased towards simple explanation [11]. Thus, if users' expectations are not properly calibrated, we hypothesize they may be prone to an IOED for two reasons: (i) *Representation/recovery confusion through abstraction of local insights:* User that are provided with local justifications of an XAI system may perceive to understand why those explanations were chosen by the system. However, under the influence of prior beliefs and misconceptions about AI, they may abstract their local insights into higher-level anecdotal evidence that may not be consistent with the predictions of other observations. End users may only become aware of these inconsistencies when they recall their abstractions to self-explain their understanding of the global ML model behavior [22]. (ii) *Label/mechanism confusion through subtle interactions:* Shapley explanations hide much of the model's complex behavior behind accessible feature value attributions [20]. Knowing what features a model has access to and the effect of feature values for some observations might results in the impression that the user understands how the model comes to its predictions for all observations. However, especially in state-of-the-art black box ML models, feature values may interact with one another in non-linear ways and significantly influence the predictions for some observations while having little effect on others.

4 SHAPTABLE

We outline the exemplary XAI system *SHAPTable* that serves as the apparatus for our user studies. First, we describe the setting and implementation details. Second, we provide details on the used explanation-generation method and the rationale for our explanation interface.

4.1 Scenario, Data Set, and ML Model

Scenario. Our scenario resembles a decision-support situation in which the human decision-maker is accompanied by an intelligent

[3] compared with other open-source XAI frameworks (such as LIME, AIX360, or DALEX), SHAP has the most engagements on GitHub

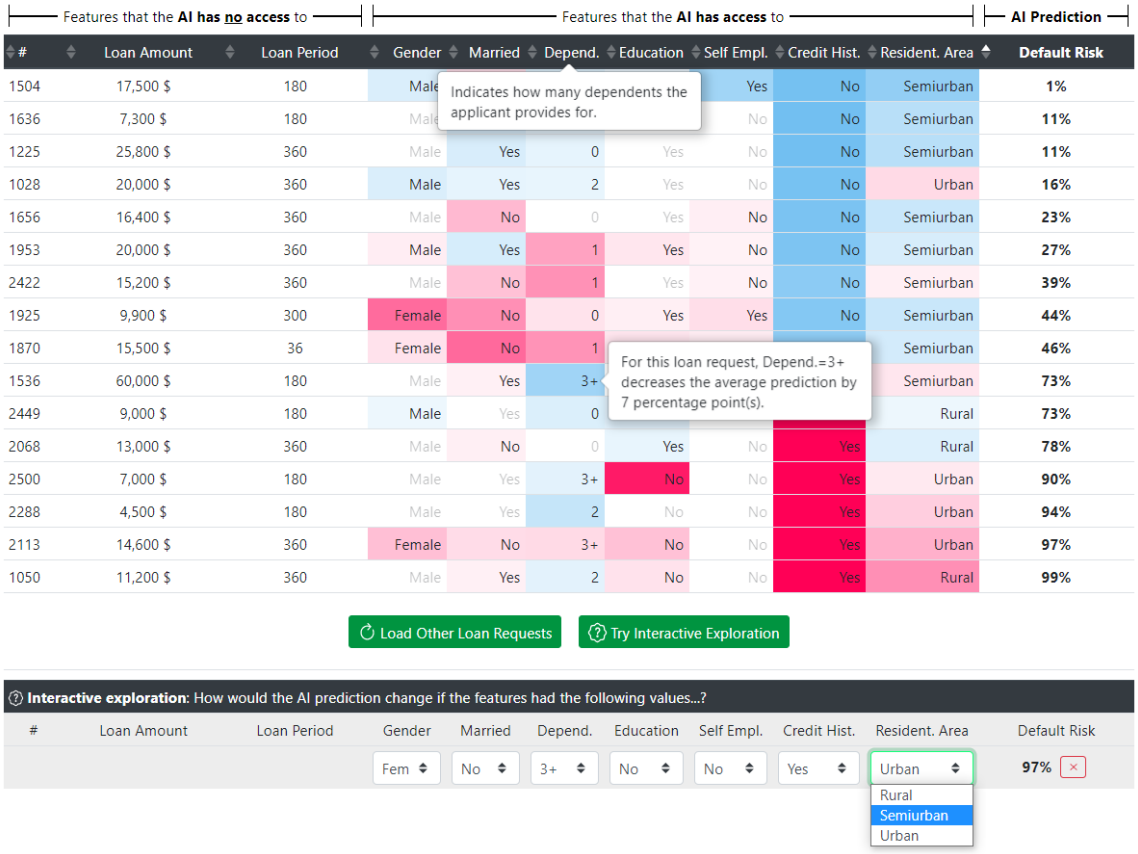

Figure 1: Overview of the explanation interface. Participants were presented a representative sample of 16 loan requests and their respective default risk prediction. Feature values of our ML model were shaded depending on their their Shapley values.

and interpretable system. Following [56], we take an XAI *operator* perspective in a loan application scenario. In such a scenario, the operating user of the XAI system is centered between the developer of the system and a decision-subject individual affected by the decision. We put our study participants in the shoes of a private lender on a fictional crowd lending platform[4]. Participants can see demographic information, loan details, and credit history of individuals that request a loan on the platform. Each request is accompanied by an "AI-based intelligent prediction" of the *default risk*, i.e., the probability that the borrower fails to service a loan installment some time during the loan period. The prediction is introduced as an "AI-based" feature that is based on machine learning from historic cases. As part of the scenario, participant evaluated a novel feature that explains the default risk prediction for each lending request through Shapley explanations. People utilize explanations for learning [32]. Thus, participants were instructed to give feedback to the platform if the provided explanation facility supports them in learning about the behavior of the default risk prediction feature (*operator interpretability* [56]).

Dataset. We chose a tabular data set for our user studies as many ML models deployed in practice build on this type of data. This applies especially to regulated domains such as healthcare, finance, and public services [5, 30]. Tabular data is often characterized by individually meaningful features and, unlike images or time series, lacks strong temporal or spatial structures [30]. Thus, each feature represents a distinct concept of a person's conceptual knowledge (e.g., gender, education, credit history). We built on the *Loan Prediction*[5] data set that is widely used for educational purposes. It consists of 614 loan requests with 13 columns. We relabeled two columns of the data set to be consistent with our scenario[6].

ML Model. We calculated the default risk prediction via a *random forest classifier (RFC)*. RFCs are widely used in many real-world contexts because of their practicability. They often yield competitive performances even without extensive ML engineering efforts. Especially for tabular data, tree-based models often outperform other black-box models [30]. However, random forests are considered black-box ML models. They consist of many decision trees. Each tree is trained on a random selection of features. The classifications

[4]a platform that facilitates the matchmaking between private lenders and borrowers over the internet

[5]https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/ or https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

[6]we re-framed the *Loan_Status* column to represent the default risk and the *Credit_History* column to represent a negative item on a credit report.

of individual trees are then combined into a final classification by a majority vote. Although individual decision trees are interpretable, it is unfeasible to understand the prediction behavior of their ensemble. To limit the cognitive load for participants we chose to train our model on a subset of columns. We used only the seven categorical columns (5 binary, 1 ternary, and 1 with four possible values). We trained a binary RFC with 100 decision trees using a 80:20 split for the training and validation sets. The split was stratified to have the same distribution of binary predictions between training and test sets. Other than that, we used the default hyperparameters of the *scikit-learn* package. The accuracy of the predicted default risk on the validation set was 0.83.

4.2 Explanation Facility

Explanation-generating Method. To algorithmically generate explanations for the default risk predictions, we build on the widely used post-hoc explanation framework *SHAP (SHapley Additive exPlanations)* [30]. SHAP belongs to the class of *additive feature attribution methods* where the explanation is represented as a linear function of feature contributions towards an ML prediction. It trains a surrogate model by slightly changing the inputs and testing the impact on the model outputs. The SHAP framework unifies the ideas of other feature attribution methods (such as LIME [44]) with *Shapley values*, which originate from game theory [51]. Applying Shapley values to XAI, an ML prediction can be modelled as a cooperative game between the features to produce a prediction. As the features may influence one another through interactions, the game is a cooperative one. With Shapley values we can assign a unique and fair contribution to each feature over all possible coalitions of features despite the presence of interactions. SHAP assigns a number for each input feature (the Shapley value) that is guaranteed to be consistent under mathematical guarantees: (i) *local accuracy* ensures that the sum of the feature contributions matches the ML prediction of an instance, (ii) *missingness* ensures that feature values that have no effect on the model prediction (e.g., because they are constant) have a Shapley value of zero, (iii) *consistency* ensures that changes in the contribution of an individual feature value in the black-box model result in a consistent change of the respective Shapley value. Consistency is interesting because it allows users to compare contributions between multiple observations, groups of observations, or even models. All contributions are relative to the *expected value*. The expected value equals the percentage of defaulted loan requests in the data set (32% for our data set). As such, it serves as a base value for all requests. The Shapley value for a feature value describes the direction and strength of the contribution relative to the expected value.

Explanation Interface. The SHAP framework provides information-dense visualizations of local and global feature attributions out-of-the-box. However, prior research showed that even ML experts face challenges to interpret them correctly without assistance [25]. Thus, for our explanation facility, we borrowed ideas from these visualizations but worked with the raw Shapley values. We assumed that most explanation-seeking end users in the decision-support context are familiar with spreadsheets. Thus, our explanation interface resembles a spreadsheet-like user interface that is overlaid with a heat map of Shapley values. We show 16 loan requests from the data set with their respective default risk prediction in percentage (i.e., 0%=no risk and 100%=highest risk of defaulting). The initial loan requests were sampled according to the confusion matrix to represent a representative range of default risk probabilities.[7] Each loan request is depicted as a table row. For each request, we show its column values in a separate cell. For columns that were used for the default risk prediction, the corresponding cell is shaded depending on their effect on the prediction. We chose a heatmap-like representation as it supports counterfactual reasoning through comparison of loan requests [58]. The direction and strength of the effect is given by the Shapley value. A red shading indicates a positive effect (increases the expected value) while a blue one a negative effect (decreases the expected value) on the ML prediction. The opacity of the shading indicates the strength of the effect. Details about the strength are provided in a tooltip when the user hovers the cell. For example in Figure 1, the fact that request #1536 has 3+ dependents decreases the expected value of 32% by 7 percentage points. We reviewed research on explanation design approaches that foster user understanding. In general, the design of explanation facilities should follow the guidelines of *contrastive*, *selective*, and *interactive* explanations [32]. Our explanation style is similar to the *input influence* explanations in [6] where each feature value is accompanied by the direction and strength of its effect on the prediction. Prior work reported that providing users with interactive explanation facilities improved their subjective and objective model understanding [12]. These mechanisms informed the designs of our explanation facilities as follows: (i) *contrastive*: we show multiple instances and their respective explanations at once so that users can contrast a local explanation with local explanations of other instances. Further, users can sort the data by columns to contrast instances with equal values to spot regularities; (ii) *selective*: we excluded neglectable feature values with absolute effects of less than one percentage point from the explanation; (iii) *interactivity*: following the call for more interactive explanation interfaces that *"allow users to explore the system's behavior freely"* [1], we provided participants with two basic interactive functionalities: (a) to *resample* a different set of 16 loan requests to get a more holistic understanding of the ML model behavior[8] and (b) to *simulate* a prediction for a hypothetical loan request with user-defined features values [12]. Figure 1 shows the final explanation interface from a participant's perspective.

5 METHODS

We pre-tested and iterated our scenario, apparatus, and procedure with 10 people to ensure they are comprehensible from a participant perspective. We applied a mixed methods approach. First, we moderated 40 participants through the study (6 of them followed a think aloud protocol to not bind cognitive capacities). Second, we conducted an unmoderated study with 107 crowd workers. Following [45], we describe our participants as educated lay users of XAI. We used a combination of moderated and unmoderated studies to account for dual process model of human reasoning [24, 58]. For the moderated study, the presence of a moderator motivated

[7] 4 requests for the 4 different combinations of predicted and actual values, i.e., *true positives*, *false positives*, *true negatives*, and *false negatives*
[8] again sampled according to the confusion matrix

participants to invest more resources and apply high-effort rational thinking throughout the procedure (*system 2 thinking*). There, we used a slightly shorter procedure to qualitatively investigate what users do to form their mental model of the global ML model behavior. In contrast in the unmoderated study, we assumed participants to be guided more by low-effort heuristic thinking (*system 1 thinking*).

5.1 Participants

5.1.1 Moderated Study (N=40). We recruited 40 participants via our internal university mailing list. All participants were supervised by a moderator during the study to ensure participants understand and follow the instructions. We randomly selected a subset of 6 participants to additionally follow a think aloud protocol. We used a subset as the think aloud puts additional cognitive load on the participants and might *"impact how people perform on cognitively-demanding tasks"* [8]. 17 participants self-identified themselves as female and 23 as male. Of these, at the time of the study 65% aged 18-24 years, 32.5% aged 25-34 years and 2.5% in the age of 35-44 years. Among the participants, 20 (50%) hold a high school degree, 10 (25%) an undergraduate degree, 8 (20%) a graduate degree, while 2 had other educational backgrounds. On average, participants took 37.8 minutes (SD=10.1 minutes) to complete the study and were compensated 10 EUR per completion. 29 (72.5%) participants disagreed and rather disagreed to have practical knowledge of AI (e.g. application of statistical learning methods or training of machine learning models), 8 agreed or rather agreed, while 3 were undecided. 29 (72.5%) agreed to or rather agreed to frequently explain complex things to other people (e.g. seminar contents to fellow students or smartphone features to friends), 11 were undecided. 19 (47.5%) participants stated they use spreadsheet applications at least weekly, while 21 used them once a month or less.

5.1.2 Unmoderated Study (N=107). We recruited participants via the crowd sourcing platform *Prolific*. The posting included a short description about the study, the expected duration, and the compensation. We only recruited workers with a 100% approval rate and at least 10 previous submissions. Further, we required all participants to hold at least an undergraduate degree. 116 participants started the study of which 8 only partly finished it. We screened the answers of all completed sessions and excluded 1 participant due to low quality verbalization that was most likely generated by a bot. Participants' demographics were quite diverse. 48 participants self-identified themselves as female and 59 as male. Participants were located in the United Kingdom (42), Portugal (13), the United States (10), and other countries (42). At the time of the study, 22.5% of participants were aged 18-24 years, 49.5% aged 25-34 years, 18.4% aged 35-44 years, and 9.6% 45+ years. Among the participants, 57.2% stated they hold an undergraduate degree, 35.9% a graduate degree, 2.9% a PhD, and 3.8% stated other as highest educational level. On average, participants took 28.5 minutes (SD=15.8 minutes) to complete the study and were compensated £3.75 per completion (=£7.09/hour). 68 (63.5%) participants disagreed and rather disagreed to have practical knowledge of AI, 25 agreed or rather agreed, while 14 were undecided. 81 (75.6%) agreed to or rather agreed to frequently explain complex things to other people, 12 (11.2%) were undecided, and 14 disagreed or rather disagreed (13.2%). 65 (60.7%) participants stated they use spreadsheet applications at least weekly, while 42 used them once a month or less.

5.2 Procedure

The goal of our user studies is to investigate if an IOED can be observed when end users are exposed to Shapley explanations of an ML model. For this purpose, we query the participants' model comprehension through different tasks and repeatedly measure their self-assessment of perceived understanding using a uniform scale. Our procedure was inspired by the study designs of the initial IOED studies [46] but adjusted to the XAI context. Participants used the apparatus described in section 4 to complete the five tasks illustrated in Figure 2. The moderated user studies were conducted via video conferencing to observe how users interact with the apparatus. We describe the stages below:

Introduction. After consenting with the participation and data processing information, participants reported their demographics (i.e., age, gender, and educational background), their frequency of use of spreadsheet applications, their frequency of giving explanations about complex topics to others, and their level of practical experience in the field of AI, (the last three questions were illustrated with example statements and rated on 5-point Likert scales). Next, we explained in multiple steps the crowdlending scenario, the "AI-enabled prediction" of the default risk, the explanation facility, and the scale self-rating scales. In the moderated study, participants were encouraged to ask clarifying questions to the moderator.

Task 1: Exploration of Black Box (only used in unmoderated study). Participants were presented with a table of 16 observations. For each observation the ML prediction was presented without any explanation. Participants were asked to spend 5 minutes and *"try to understand how the AI forms its default risk predictions"*. Afterwards, they were asked to rate their perceived understanding. To give them an indication, a timer showed how much time they already spent on this task. We used this task in the unmoderated study to ensure that our explanation interface was perceived to improve understanding of participants.

Task 2: Exploration of Explanation Facility. Next, we provided participants with the explanation facility presented in section 4.2. We asked them to freely explore the decision-making behavior of the prediction model for no longer than 10 minutes and re-rate their gained understanding. To give them an indication, a timer showed how much time they already spend on this task.

Task 3: Verbalization of Understanding. According to psychological research deliberate self-explanation results in a more realistic assessment of a user's own understanding and may potentially refine it [22, 35]. It does not matter whether the self-explanation is self-motivated or prompted by an instructor [22]. Further, retrospection techniques such as (self)-explanation, can provide rich information about a user's mental model [22]. Thus, as a next step, participants had to write a detailed explanation of their global understanding of the ML model's prediction behavior. Their explanation was to be between 50 and 100 words long and address three guiding questions. After the participants verbalized their understanding, they re-rated their perceived understanding.

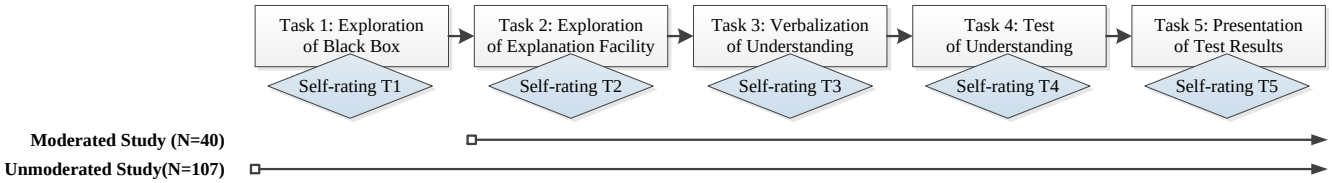

Figure 2: Stages of the procedure in the moderated and unmoderated study. First, we observe in task 1 (only in unmoderated study) and task 2 what end users do to form their mental models of the global ML model behavior. Second, we assess in tasks 3 to 5 what end users think they know about the behavior of an ML model in relation to what they actually know. Through multiple tests of comprehension, we assess how stable their self-reported understanding is if users need to put it into action.

Task 4: Test of Understanding. For the diagnostic questions, we based our questions on prediction tasks, where the participants had to simulate the prediction of the ML model for given sets of features. Afterwards, participants re-rated their perceived understanding.

Task 5: Presentation of Test Results. Rozenblit and Keil confronted their participants after the diagnostic questions with an expert statement [46]. In our case, we showed the participant's answers and contrasted them with the default risks predicted by the ML model. Further, we showed the Shapley explanations for each observation. We summarized their results as *"You predicted <n> out of 8 loan requests like the AI"*. This allowed the participants with incorrect predictions to re-examine the ML model behavior. Afterwards, participants re-rated their perceived understanding. Each session ended with a short questionnaire.

5.3 Dependent Variables

Self-Rating of Perceived Understanding. We used a uniform 7-point Likert scale that measures each participants' perceived understanding at multiple points throughout the study. We adopted the scale from the original IOED experiments and fitted it to the XAI context. To calibrate participants' usage of the scale, we demonstrated the scale during the introduction and provided explanations for levels 1, 4, and 7. On level 1, respondents think they can name features that the ML model has access to and what it predicts. On level 4, they think they understand the relative importance of individual features. At the highest level, level 7, they think they understand the absolute importance of individual feature values as well as possible interactions between them.

Objective Understanding. Following [12] and [59], a user "understands" an ML model *"if the human can see what attributes cause the algorithm's actions and can predict how changes in the situation can lead to alternative algorithm decisions"*. We built upon two question types from the explanation evaluation framework proposed by [12] to measure participants' objective model understanding. In total, we asked 8 questions (6x forward simulation, 2x relative simulation). For the first question type, we presented them with an observation and asked *"What do you think will the ML predict?"* (*forward simulation task*). We selected the observations according to the default risk predicted by the ML model: two at the extremes (0%, 100%), two with low risks (11%, 29%) and two with high risks (68%, 69%). We provided participant with five answer options of prediction ranges (from *0-20%* to *81-100%*). Following [22], participants had to rate their confidence for each prediction on a 5-point Likert scale (1=*very unconfident* to 5=*very confident*). As a second question type, we asked them to select the loan request with the highest (lowest in the second question) predicted default risk from a set of three given requests (*relative simulation task*). We offered three loan applications that differed in three (five in the second question) of the seven features that had on average a medium to low effect. The ML prediction of the correct option differed by at least 30 (66 in the second question) percentage points from the other options. Again, they had to rate their confidence in their simulation. We counted the number of correct answers and the mean deviation from the correct answer.

Demographics, Literacy, and Interaction. We asked participants on their age, gender, and level of education. Subject to participants' approval, we screen recorded their interactions in the moderated study. Further, we measured how much time participants spent at each step and logged their interactions with the explanation facilities (e.g., number of resamples and simulations). We used those measures as additional levels of control for analysis.

5.4 Design and Analysis

Both studies used a within-subjects design. Following the analysis in the original IOED experiments, we analyzed the differences in self-ratings through a *repeated measures ANOVA* [46]. None of the self-ratings of understanding were normally distributed. As a paired Student's t-test is not valid in such a case, we used a Wilcoxon signed-rank (WSR) test to analyze the planned linear contrasts for T1<T2, T3<T2, T4<T3, T5<T4 and T5<T2. If not stated otherwise, we based our significance at α=.05.

6 RESULTS

6.1 Robustness of Perceived Understanding

To answer RQ1, we present the distribution of participants' self-ratings throughout the moderated and unmoderated studies (see Figure 3). For comparison with the original IOED studies, the differences in the reported understandings were significant across the stages (repeated measures ANOVA: $F(4,424)=28.260$, $p<.001$, $\eta_p^2=.21$)

Shapley Explanations Increased Self-Ratings (T1<T2). In the unmoderated online study, participants reported on average rather high understanding levels even without explanations of the ML model (median=4, mean=4.33). 53 participants increased their understanding by at least one level after being exposed to Shapley

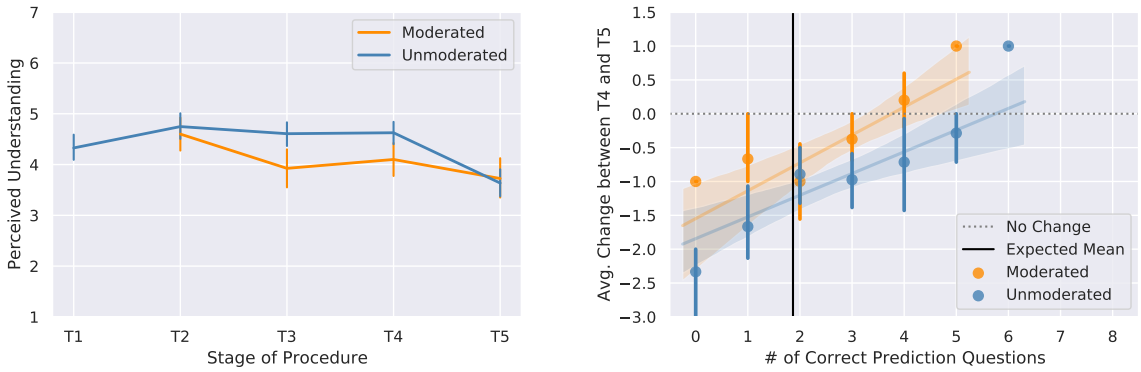

Figure 3: ***(left)*** **The means of self-ratings throughout the procedure for the moderated and unmoderated studies. In the moderated setting, we observed two large drops, one after the verbalization in T3 and another after the presentation of test results in T5. In the unmoderated setting, the drops remained insignificant until the last stage.** ***(right)*** **The average change in self-ratings between T4 and T5. After the participants saw their test results, most of them downgraded their perceived understandings.**

Table 1: The left side shows the mean and standard deviation of participants' self-ratings of understanding in the moderated and unmoderated studies. The right side presents the number of participants that *decreased* (*increased* for T1<T2) their self-rating by at least one level (#) and the results of our hypotheses tests using non-parametric Wilcoxon signed-rank test (w). The significance levels are reported as following: * p<.05; ** p<.01; *** p<.001

		T1	T2	T3	T4	T5		T1<T2	T3<T2	T4<T3	T5<T4	T5<T2
Moderated Study	Mean		4.60	3.95	4.10	3.73	#		19	7	18	25
(*N = 40*)	SD		1.06	1.21	1.03	1.28	w		6.5***	150.5	66.0**	76.0***
Unmoderated Study	Mean	4.33	4.75	4.61	4.63	3.64	#	53	27	20	72	77
(*N = 107*)	SD	1.30	1.33	1.25	1.12	1.38	w	2055.5**	408.5	384.0	358.0***	564.0***

explanations. Across all participants, the average reported understanding increased significantly (median=5, mean=4.75, w=2055.5, p<.01). Thus, **H1** was supported and our explanation interfaces was at first perceived as valuable to participants.

Examination Decreased Participants' Self-Ratings (T5<T2). Most participants in both studies significantly (p<.001) decreased their perception of understanding over the course of the procedure: 63% of participants in the moderated study and 72% in the unmoderated. Thus, **H2** stating that participants fell for an IOED was supported. Below, we report the changes in the self-ratings at individual stages of the procedure. *Verbalization (T3<T2):* In the original IOED studies, deliberate self-explanations decreased the perceived understanding. In our moderated studies, 48% of participants decreased their rating at this stage. The drop was significant. In the unmoderated online study, we observed a drop for only 25%. The drop was not significant. *Test of Understanding (T4<T3):* Participants remained confident in their understanding during the prediction tasks. Only, 19% decreased their rating in the unmoderated setting, compared to 18% in the moderated study. The drops were not significant. Contrary to our expectations, the prediction tasks increased the perceived understanding in the moderated study. *Test Results (T5<T4):* Confronting participants with their results of the prediction tasks caused a significant drop in understanding in both studies. In the unmoderated study, 67% decrease their understanding compared to the previous stage. In the moderated study, 45% did so. The drops in both studies were significant.

Moderated Participants Devoted More Resources. Participants in the moderated setting spent significantly more time on the study tasks than in the unmoderated setting. In the moderated study, participants spent on average 9.8 minutes (SD=4.9) exploring SHAP-Table, 10.9 minutes verbalizing their understanding (SD=4.2), and 7.1 minutes solving the prediction tasks (SD=3.0). In contrast in the unmoderated study, they spent only 3.8 minutes (SD=2.6), 6.7 minutes (SD=4.9), and 3.3 minutes (SD=1.9).

Moderated Participants Performed Better in Test of Understanding. We analyzed the number of correct predictions and the mean error of participants' predictions. The mean error describes the average number of bins between the participant prediction and the AI prediction over all questions (e.g., error between "0-20%" and "41-60%" is 2). On average, participants answered 2.85 (SD=1.05) questions correctly in the moderated and 2.66 (SD=1.20) questions in the unmoderated study. Both are significantly better than a random guess (expected mean) that would result in 1.86 correct questions. Further, on average, the mean error of participants in the moderated study (1.07, SD=0.29) was significantly lower compared to participants in the unmoderated study (1.22, SD=0.37). Both are significantly better than a random guess (expected mean) that would result in a mean error of 1.7 (see Figure 4).

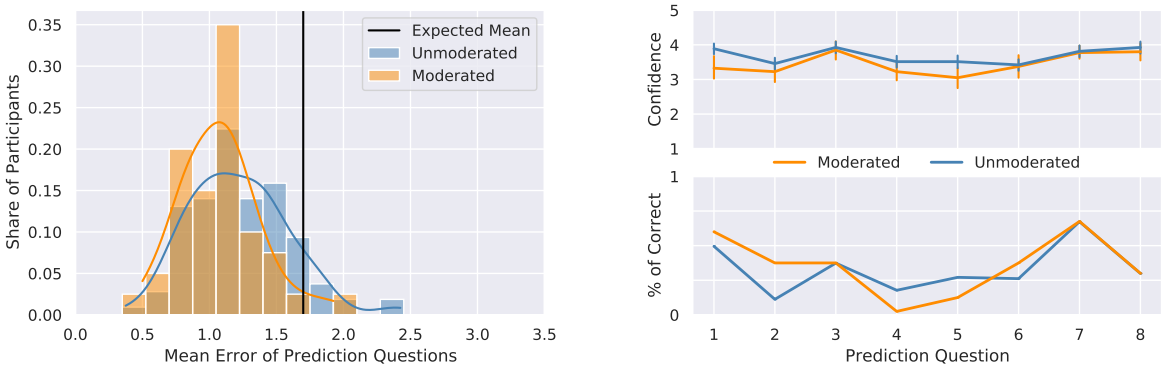

Figure 4: *(left)* The proportion of participants by their average distance to the correct answer (*mean error*). Participants in the moderated study were significantly closer to the correct answers than their unmoderated counterparts. The average confidence *(top right)* and share of correct answers *(bottom right)* for each prediction question in T4. On average, participants in the unmoderated study were more confident throughout the procedure.

6.2 How Users Formed Their Understanding

To answer RQ2, we report observations gained from the 6 think aloud protocols. We revisited the screen and audio recordings and openly coded recurrent themes during participants' interaction with SHAPTable. ***Orientation***: Participants used the coloring in SHAPTable to gain a first overview. They visually looked for inconsistencies in the heatmap. To calibrate their understanding of the coloring and the associated feature contributions, participant TA1 studied multiple tooltips. TA2 looked for *"global heuristics that always apply"* by shifting the attention from one feature to another (*column-wise comparison*). TA4 used a combination of sorting and rapidly resampling *"to look for [visual] patterns"*. Soon he stated that *"credit history correlates with the prediction without dependencies"*. After some resamples, the participant spotted an outlier that violated this hypothesis. TA1 identified an outlier in the heatmap where an effect was unusually strong. By this the participant realized that there are interactions in place. This discovery served as a starting point for deeper analysis. ***Analysis***: Single outliers guided the reasoning process of most participants. After they visually spotted one in the heatmap, they often replicated an observation in the simulation feature to *"live edit"* single feature values to understand their contributions. Further, participants often performed *pairwise comparisons* between two observations to understand differences. ***Abstraction***: All participants realized that interactions are present, but often over- or underestimated their impact. If they stumbled upon effects that violated their prior beliefs (e.g., that fewer dependents decrease the default risk), they searched for anecdotal memory aids for what they saw. Sometimes these were built from fragmented insights consisting of few features (e.g. *"self-employed in rural areas are high risk. That does not make sense."*) and missed that another feature (e.g. gender) had an impact too. Some participants stated it was difficult for them to assess when they should generalize from outliers and when not. Also, some participants assumed monotonic features effects (e.g., the effects of 0 vs. 2 dependents). If they found cases that violated this assumption, they judged the AI behavior *"as illogical"*. During verbalization, some participants recovered the effects of feature values from their memory aids and from the colors they remembered. ***Additional functionalities***: Some participants wished for aggregated *"scenarios"* consisting of similar observations (e.g. combinations of feature values that have consistent effects) and examples that illustrate interactions between features for easier orientation. TA1 and TA2 wished for an improved sorting feature that allows sorting by SHAP values to group observations with similar effects close to each other to identify regularities. TA1 wished for multiple rows in the simulation feature to simultaneously explore multiple combinations at once. Further, he wished to duplicate one observation into the simulation feature for improved usability. ***Reflection***: Participants perceived the study procedure as valuable. For example, participant TA1 considered the study procedure as a feedback loop that helped *"to learn from mistakes and expose my misconceptions [about the ML model behavior]"*. TA4 would have liked to complete the cycle multiple times to refine their insights: *"If I were to do this task again, I would gain a much better understanding."*

7 DISCUSSION

With a moderated and unmoderated study, we examined if and why an illusion of explanatory depth (IOED) emerges when non-technical users of XAI are exposed to local Shapley explanations. Our results indicate that participants overrated their understanding of the ML model behavior after freely exploring it with SHAPTable. On average, participants in both studies significantly decreased their perceived understanding throughout the procedure. What differed were the stages at which the drops occurred. In the moderated setting, we observed two large drops. One after the self-explanation stage (48% decreased their self-rating by at least one level) and another after the presentation of test results (45%). In the unmoderated setting, the self-ratings of participants remained mostly unchanged until the last stage. After they had seen and analyzed their test results, 67% decreased their self-rating. The IOED was more pronounced for participants in the unmoderated study. They spend significantly less time at each stage and had a significantly narrower objective understanding according to our prediction tasks. Still, on average they were more confident about the correctness

of their prediction questions. The magnitude in the decrease in self-ratings in the last stage depended on the number of correct predictions and was stronger in the unmoderated setting. It seems that participants in the unmoderated study expected more correct answers of themselves. While moderated participants with 4 out of 8 correct answers refrained from downgrades, unmoderated participants downgraded it even with 5 out of 8 correct questions. We interpret that participants in the unmoderated setting were guided by heuristic thinking and did not realize the incompleteness of their understandings until they saw their test results. We believe, they were less aware of irregularities of feature values effects and feature interactions than participants in the moderated setting. Overall, 85% of participants in the moderated and 69% in the unmoderated study agreed or agreed completely that the study procedure *"helped me to better assess my own understanding of the AI prediction behavior"*.

Humans will most likely never be able to correctly predict the behavior of complex non-linear ML models. Our results highlight the importance of XAI systems to not only provide non-technical users with static justifications, but also guiding user interactions that support them in building an accurate mental model – even if this means exposing complexities and irregularities of the ML model behavior. Otherwise, providing them with seemingly simple local justifications of complex ML behavior (as with Shapley values) may leave them with an *"easiness effect"* [50]. Below we discuss implications for the design of XAI systems derived from our findings and outline its limitations.

Calibrating Understanding as Part of XAI Interaction: An effective XAI system need to capture a wrong or incomplete mental model of the user and adjust its explanations accordingly [48]. An implication for XAI designers is that calibrating user perception of understanding through a structured procedure, as outlined in our studies, might expose that the system is more complex than it seems at first. For example, Cai et al. [10] described the onboarding phase to an XAI system as a key phase that forms users' initial impressions of an XAI system. It is during the onboarding that users form their mental model of the capabilities and limitations of the XAI system. Deliberate self-explanation has been proposed as being an effective way to calibrate XAI understanding [22, 35]. However, our results indicate this is only the case if users are willing to devote the required cognitive capacities. Buccinca et al. [8] describe cognitive forcing strategies, such as forcing users to form an own prediction before being confronted with the AI prediction. Our multi-stage procedure extends this idea in a playful way. Future work could explore how to leverage the individual results of such procedures to automatically learn about the mental model of the user and personalize explanations during the interaction with the XAI.

Forming (Global) Rationales from Local Explanations: Like [3], our results indicate that participants had difficulties in abstracting their local insights to a global understanding. They understood the justifications provided for individual observations but struggled to assess how representative they were for the average model behavior. The properties of SHAP enable novel ways for interactivity [13, 43] to provide selective, contrastive, and interactive explanations [32] that might bridge the gap between local and global understanding [42]. Future research could explore how to condense multiple local explanations into accessible higher order explanations to contextualize them. Such novel ways of interactivity could support the interpretation strategies applied by participants in our studies. This resonates with the concept of *rationales* [15, 17]. These aim to provide end users with contextually appropriate reasons for an ML prediction in natural language.

Limitations. There are several limitations to our studies. First, we examined a simplified extrinsic [38] scenario around a tabular data set. Thus, the external validity beyond this scenario (i.e., different decision-making situation) and type of data (i.e., visual data or natural language data) is uncertain. Second, the emergence and strength of an IOED might highly depend on the target audience. Physicians and risk managers may have very different reasoning strategies than the educated lay users in our studies. Future work could investigate different extrinsic as well as intrinsic [38] scenarios with varying ML model complexities or XAI methods. Still, we are confident that our insights highlight the importance of keeping cognitive biases in mind when designing and deploying XAI.

8 CONCLUSION

With XAI systems expected to be deployed deeper into organizations and society, it is important to understand how non-technical users of XAI consume explanations. In this work, we examined how non-technical XAI users form their mental model of the global ML behavior. Our results indicate that users overestimate the understanding they gain because of the illusion of explanatory depth. Further, we describe reasoning and interaction strategies that users applied. Future work could investigate how these strategies can be included into interactive explanation facilities to make them aware of potential fallacies and to support their reasoning. We offer starting points for XAI designers on how to support non-technical users to form a more appropriate mental model of ML model behaviors.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 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* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[2] A. Adadi and M. Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Ahmed Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Bianchi-Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. *Proceedings of the 25th International Conference on Intelligent User Interfaces* (2020).

[4] Adam L. Alter, Daniel M. Oppenheimer, and Jeffrey C. Zemla. 2010. Missing the trees for the forest: a construal level account of the illusion of explanatory depth. *Journal of personality and social psychology* 99 3 (2010), 436–51.

[5] Umang Bhatt, Alice Xiang, S. Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and P. Eckersley. 2020. Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[6] Reuben Binns, M. V. Kleek, M. Veale, Ulrik Lyngs, Jun Zhao, and N. Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *ArXiv* abs/1801.10408 (2018).

[7] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[9] Adrian Bussone, S. Stumpf, and D. O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics* (2015), 160–169.
[10] C. J. Cai, S. Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3 (2019), 1 – 24.
[11] N. Chater. 1999. The Search for Simplicity: A Fundamental Cognitive Principle? *Quarterly Journal of Experimental Psychology* 52 (1999), 273 – 302.
[12] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).
[13] Michael Chromik. 2020. reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP. In *Proceedings of 18th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).
[14] Dennis Collaris, Leo M. Vink, and Jarke J. van Wijk. 2018. Instance-Level Explanations for Fraud Detection: A Case Study. *ArXiv* abs/1806.07129 (2018).
[15] Devleena Das and S. Chernova. 2020. Leveraging rationales to improve human task performance. *Proceedings of the 25th International Conference on Intelligent User Interfaces* (2020).
[16] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretability. *CoRR* abs/1702.08608 (2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608
[17] Upol Ehsan, Pradyumna Tambwekar, L. Chan, B. Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
[18] Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *ArXiv* abs/2004.07780 (2020).
[19] Joey F. George, Kevin Duffy, and Manju K. Ahuja. 2000. Countering the anchoring and adjustment bias with decision support systems. *Decis. Support Syst.* 29 (2000), 195–206.
[20] Alicja Gosiewska and Przemyslaw Biecek. 2020. Do Not Trust Additive Explanations. arXiv:1903.11420 [cs.LG]
[21] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (aug 2018). https://doi.org/10.1145/3236009
[22] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR* abs/1812.04608 (2018). arXiv:1812.04608 http://arxiv.org/abs/1812.04608
[23] Hsieh-Hong Huang, J. Hsu, and Cheng-Yuan Ku. 2012. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decis. Support Syst.* 53 (2012), 438–447.
[24] D. Kahneman. 2011. Thinking, Fast and Slow.
[25] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219
[26] Tae-Nyun Kim and Hayeon Song. 2020. The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
[27] A. Lau and E. Coiera. 2009. Research Paper: Can Cognitive Biases during Consumer Health Information Searches Be Reduced to Improve Decision Making? *Journal of the American Medical Informatics Association : JAMIA* 16 1 (2009), 54–65.
[28] Rebecca Lawson. 2006. The science of cycology: Failures to understand how everyday objects work. *Memory and Cognition* 34 (2006), 1667–1675.
[29] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3, Article 30 (June 2018), 27 pages. https://doi.org/10.1145/3236386.3241340
[30] Scott M. Lundberg, G. Erion, Hugh Chen, Alex J. DeGrave, J. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2 (2020), 56–67.
[31] J. McGuirl and N. Sarter. 2006. Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors: The Journal of Human Factors and Ergonomics Society* 48 (2006), 656 – 665.
[32] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1 – 38. https://doi.org/10.1016/j.artint.2018.07.007
[33] Candice M. Mills and Frank C. Keil. 2004. Knowing the limits of one's understanding: the development of an awareness of an illusion of explanatory depth. *Journal of experimental child psychology* 87 1 (2004), 1–32.
[34] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *CoRR* abs/1811.11839 (2018). arXiv:1811.11839 http://arxiv.org/abs/1811.11839
[35] Shane T. Mueller, Robert R. Hoffman, William J. Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *CoRR* abs/1902.01876 (2019). arXiv:1902.01876 http://arxiv.org/abs/1902.01876
[36] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and B. Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116 (2019), 22071 – 22080.
[37] Gregory L. Murphy and Douglas L. Medin. 1985. The role of theories in conceptual coherence. *Psychological review* 92 3 (1985), 289–316.
[38] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR* abs/1802.00682 (2018). arXiv:1802.00682 http://arxiv.org/abs/1802.00682
[39] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
[40] M. Nourani, Donald R. Honeycutt, Jeremy E. Block, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, and V. Gogate. 2020. Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
[41] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0
[42] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* (2019), 1–19.
[43] Shubham Rathi. 2019. Generating Counterfactual and Contrastive Explanations using SHAP. arXiv:1906.09293 [cs.LG]
[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).
[45] Mireia Ribera and Àgata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *IUI Workshops*.
[46] Leonid Rozenblit and Frank C. Keil. 2002. The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive science* 26 5 (2002), 521–562.
[47] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[48] Heleen Rutjes, M. C. Willemsen, and W. IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019*.
[49] James Schaffer, J. O'Donovan, James Michaelis, A. Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
[50] Lisa Scharrer, Yvonne Rupieper, M. Stadtler, and R. Bromme. 2017. When science becomes too easy: Science popularization inclines laypeople to underrate their dependence on experts. *Public Understanding of Science* 26 (2017), 1003 – 1018.
[51] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.
[52] L. Skitka, K. Mosier, M. Burdick, and B. Rosenblatt. 2000. Automation Bias and Errors: Are Crews Better Than Individuals? *The International Journal of Aviation Psychology* 10 (2000), 85 – 97.
[53] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and H. Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020).
[54] Kacper Sokol and Peter A. Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).
[55] Jacob Solomon. 2014. Customization bias in decision support systems. In *CHI '14*.
[56] Richard Tomsett, Dave Braines, Dan Harborne, A. Preece, and S. Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *ArXiv* abs/1806.07552 (2018).
[57] Michelle Vaccaro and Jim Waldo. 2019. The effects of mixing machine learning and human judgment. *Commun. ACM* 62 (2019), 104 – 110.
[58] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.
[59] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62 (2019), 70 – 79.
[60] Andrew Zeveney and Jessecae Marsh. 2016. The Illusion of Explanatory Depth in a Misunderstood Field: The IOED in Mental Disorders. In *CogSci*.