On Interpretability and Overreliance

Julian Skirzyński
UC San Diego, CSE
jskirzynski@ucsd.edu

Elena Glassman
Harvard University, SEAS
eglassman@g.harvard.edu

Berk Ustun
UC San Diego, HDSI
berk@ucsd.edu

Abstract

One of the underlying drivers to create interpretable models is that they may help humans make better decisions. Given an interpretable model, a human decision-maker may be able to better understand the model's reasoning and incorporate its insights into their own decision-making process. Whether this effect occurs in practice is difficult to validate. It requires accounting for individuals' prior beliefs and objectively measuring when reliance on the model goes beyond what is reasonable given the available information. In this work, we address these challenges and validate if interpretability improves decision-making. Concretely, we compare how humans make decisions given a black-box model and an interpretable model, while controlling for their prior beliefs and rigorously quantifying rational behavior. Our results show that interpretable models can lead to overreliance and that the level of overreliance varies across models that we would consider to be equally interpretable. These findings raise fundamental concerns about current approaches to AI-assisted decision-making. They suggest that making models transparent is insufficient—and currently counterproductive—for promoting appropriate reliance.

1 Introduction

Machine learning models are already applied in high-stakes domains. For instance, hundreds of clinical prediction models are in everyday use by healthcare providers (see, e.g., 300+ scoring systems on mdcalc.com). These providers decide whether to adopt such models, and then decide how to incorporate the model's predictions into their own decision-making about specific clinical cases.

When the model is a black box, it provides no information to help the human decide which predictions to rely on and which to ignore. Supervising such a model can be a daunting task. To alleviate this challenge, researchers established the burgeoning field of AI interpretability and explainability [8, 3, 73, 5, 10, 1, 70, 40], aiming to make the model's logic comprehensible to people. This field spawned thousands of papers defining interpretability [42, 79, 21] or developing models aimed at achieving it—either through direct design [73, 70, 57] or through explanation methods that approximate complex models through simpler ones on a subset of predictions [55, 8, 3, 5, 10, 1, 40].

The common belief is that presenting the logic employed by AI models to users can positively impact their trust and reliance on these systems. In principle, interpretability *should* help humans spot model errors or potentially unsafe model behavior—paving the way to "debug" the model, improve it, stop it from deployment, or simply better judge when to accept or override its recommendations.

The issue is that we cannot currently tell if making AI interpretable actually helps humans make better decisions. So far, empirical evidence investigating this topic is mixed. Initial studies found that interpretability can increase acceptance of AI predictions regardless of correctness, leading to an effect called "overreliance" [9, 16, 75, 11, 12, 15, 52]. Newer studies demonstrate that interpretability can also exert a beneficial influence on decision-making [62, 77, 37, 20]. Still others found null effects [76, 2, 67]. These contradictory findings stem from fundamental challenges in studying

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

human-AI interaction that make the results difficult to generalize. These challenges include designing realistic yet controlled tasks, accounting for participants' prior beliefs, and objectively measuring when reliance becomes overreliance.

Addressing these challenges is not trivial. Firstly, our selection of the task may easily affect the population we should test. For instance, recidivism prediction might require more participants with sophisticated knowledge, but narrowing down the pool of participants too much may lead to context-specific results. Secondly, what participants know about the task environment affects their subsequent decisions, e.g., two doctors making different diagnoses for the same patient based on variations in their patient record. Thirdly, we need a principled manner of defining when reliance on the model really means *over*reliance. For instance, a simplistic definition of overreliance might consider accepting a prediction from a 99% accurate model as overreliance, just because the prediction turned out to be incorrect. In reality, however, it is likely a perfectly rational decision.

In this work, we develop a novel experimental framework to tackle these challenges. Our framework consists of a task and an experimental protocol that enables us to account for a wide range of potentially confounding factors. We control for factors stemming from interpretability (we use extremely simple models), individual variation (we elicit prior beliefs), and the task (our task does not require expertise, can be run online, and resembles decision-making in practice). We also rigorously quantify overreliance by comparing human behavior to the behavior of a rational agent. This framework enables us to show that people fall prey to the *interpretability trap*—the general tendency to overrely on interpretable models with the degree of overreliance controlled by model format. Specifically, we find that (1) making blackbox AI models interpretable increases overreliance and (2) overreliance can change if we change how the model's logic is represented (!)

Our results suggest that even full understanding of the model's logic is insufficient to counter overreliance. To the contrary, our results show that even small perturbations to how the logic is presented may generate systematic differences in reliance. This is concerning because it means that building transparent models may not only be insufficient, but also counterproductive for promoting appropriate reliance on AI. If this is the case, the use of interpretable or explainable models [23] may have already led to erroneous, unsafe, or unjustified decisions by the humans using these models. For instance, models on the previously-mentioned website for doctors, i.e., mdcalc.com, are actually all interpretable in that they are linear combinations of human-understandable features. In the worst case, this interpretability could have led to a decision violating clinical standards just because the model used by a doctor used non-integer values instead of integer ones.

Our main contributions include:

1. Designing an experimental framework (Section 2) where we control for interpretability-related, user-related, and task-related confounding factors; this allows us to isolate the fundamental effects of interpretability on human behavior.

2. Quantifying overreliance with respect to an agent that models rational behavior (Section 2.1); this allows us to measure it objectively.

3. Running a batch of experiments that showed people overrely on interpretable models (Section 3.1), this effect is robust to model accuracy (Section 3.2.1), but changes depending on how the models are presented (Section 3.2.2),

4. Providing a discussion in Section 4 that points to the need to re-think how we can promote appropriate reliance.

2 Experimental Design

Our research considers decision support with simple prediction models that people can understand. The goal is to examine whether, when, and how interpretability affects humans' reliance on AI models using carefully controlled experiments. By manipulating the aspects of the decision environment and the AI model, we strive to disentangle different confounding factors (e.g., strong prior beliefs) and characterize overreliance in more detail.

Studying overreliance on AI models in an experimental setting is difficult due to several factors:

C1) People are biased towards information that confirms their beliefs [43, 45, 30, 34, 31]. This confirmation bias could affect reliance and cause overreliance. However, any experimental study works with individuals that have prior beliefs. Those beliefs are (initially) unknown.

C2) Overreliance can come from different sources: seeing the model's logic (explanation bias [25]) or seeing its predictions [36, 13]. People may also favor AI [32] or exhibit algorithmic aversion towards AI as a construct in general [59]. It is thus necessary to compute the effects of interpretability and model predictions separately.

C3) People tend to attach to the first information they see [61]. This anchoring bias may cause them to stick to the AI's prediction if it is shown first and, possibly, cause overreliance [12].

C4) Deployed models often make predictions in complex environments. Recreating that complexity to make the task challenging while retaining interpretability is non-trivial.

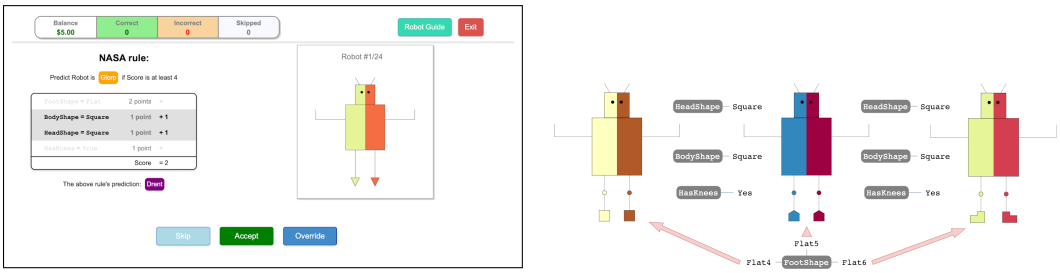
We address these challenges through a carefully designed robot classification experiment described in detail in Appendix B. Building on an existing psychological categorization task [69, 68, 18], participants classify fictional robots as either a "Glorp" or a "Drent" through four stages:

1. *Anchoring Stage*: We present participants with images of labeled Glorps and Drents in an attempt to anchor their beliefs to a "mental model" f^{user}—one of a subset of possible linearly separable mappings from robot features to robot type that is the easiest to spot (C1). The robots differ in terms of categorical attributes, e.g., $\texttt{BodyShape} \in \{\textsf{Round}, \textsf{Square}\}$, and are individualized by spurious features and color (see Fig. 1b)

2. *Probing Stage*: We elicit a complete specification of participants' post-anchoring beliefs by asking them to specify whether each (new) robot example is a Glorp or a Drent. In this way, we confirm their mental model f^{user} (C1).

3. *Model Selection Stage*: We present participants with an interpretable classification model f^{shown} to predict if a robot is a Glorp or a Drent, e.g., Predict $\texttt{Glorp}$ if $\texttt{BodyShape} = \textsf{Round}$. The model is chosen to satisfy a given property with respect to the mental model, such as adding N conditions, removing K conditions, or both. We also select the ground truth f^{true} to ensure that the mental model f^{user} and the shown model f^{shown} achieve a specific level of true performance (C4).

4. *Deployment Stage*: We show participants new robots and measure their performance. Participants see the new robot and categorize it as a Glorp or a Drent without the assistance of a model (C3). Then, they see the f^{shown} model's prediction with or without the model prediction logic visible (C2), and decide whether to accept the predicted robot type as their final decision, override the predicted robot type with a different answer, or abstain from predicting if they are sufficiently uncertain (see Fig. 1a; C4). These decisions are motivated by rewards and punishments participants may receive based on their performance. We chose Score Functions as the main model representation, as they are simple, sparse linear classification models, validated through years of practice in clinical or judicial decision-making [6, 7, 66, 33, 58].



(a) The interface during the Deployment stage.

(b) Subtypes of $\texttt{FootShape}$.

Figure 1: Left: Participants are asked to determine if a robot is a Glorp or a Drent with the assistance of predictions from a model f^{shown}. For each robot, people see an image of a robot, the model, its prediction for the robot, and a summary of their previous decisions and performance. Right: Three robots with the same unique feature pattern which differ in terms of the subtype.

2.1 Measuring Inappropriate Reliance

After deployment, we gather four key sequences: robot feature patterns, ground truth robot types, model predictions, and user decisions. Using these sequences, we quantify user behavior through four metrics: Reliance (agreement with model; Pr(Accept)), Overreliance (excessive acceptance of predictions; Pr(Accept | Should not Accept)), Earnings Deviation (performance gap relative to the rational behavior), and Correctness (decision accuracy; Pr(correct prediction)).

To define inappropriate reliance and the measures above, we first formalize a rational agent's behavior. This agent makes decisions based on (1) the stakes in the decision problem (rewards: 5 for correct override, 1 for correct accept, -3 for mistakes); (2) accumulated knowledge about model performance and (3) known robot types. The rational agent updates its belief about model accuracy (p_t) and ground-truth robot types based on observed outcomes. It makes decisions that maximize the expected reward. It selects the ground-truth robot type after encountering a robot with the same set of features it has seen before. In the case of a new robot, it overrides predictions when $p_t \leq \frac{5}{8}$, skips when $p_t \in (\frac{5}{8}, \frac{3}{4})$, and accepts when $p_t \geq \frac{3}{4}$. The measures are formalized using the rational agent's policy π as the baseline.

3 Interpretability Does not Lead to Better Decisions

We examined interpretability in a scenario where humans must evaluate model logic that in some way differs from their understanding. This is a common problem [78, 44, 24, 57, 54, 63] that can arise e.g. in clinical decision-making [35, 14]. It is common because machine learning models very often find patterns that experts didn't anticipate [56, 4, 47, 41, 19, 14]. In cases like these, people may have prior beliefs about important features and their relationship with the outcome. A model may include some of these features and respect some of their relationships. Interpretable models would show people this overlap between their beliefs and the outcome – i.e., given the model, people would be able to tell if the model is using the patterns they think are important (or not), and whether it is using them in the right way (or not). This understanding should enable people to evaluate the model prior and in deployment and better understand when the model makes wrong decisions.

Contrary to this common-sense assumption, however, in Section 3.1 we find that interpretability leads to overreliance. We call that effect the *interpretabiltiy trap*. Our analysis reveals that overreliance is robust to traditional factors we consider in machine learning and Human Computer Interaction such as accuracy (see Section 3.2.1), and instead changes with factors such as model format (see Section 3.2.2) that has not been previously controlled. This result is particularly important because current work focuses on making models more interpretable [57, 8, 3, 73, 5, 10, 1, 70, 40] without confirming if interpretable models lead to more appropriate reliance and better decision-making.

3.1 Experiment 1: People overrely on interpretable models

In this experiment, we studied **how people's reliance on models changes when interpretability reveals model uses new prediction patterns**. Such models could extend human knowledge and decision-making capabilities by uncovering patterns that were previously unknown or overlooked [57, 41, 19, 14]. They could also be wrong and signal overfitting [14, 55].

Setup Our selection of the experimental parameters was aimed at presenting a model that is sufficiently complex but still trivial enough to be stored in working memory. To do so, we anchored towards mental model f^{user} with 2 features (see Fig. 5) – $\texttt{BodyShape} = \textsf{Square} \wedge \texttt{FootShape} = \textsf{Pointy}$ – and picked f^{shown} that is as similar to f^{user} in terms of a Score Function as possible, but makes at least 2 different predictions. There was only one linearly separable model meeting these constraints (see Fig. 2). To model a realistic AI-assisted decision-making scenario, we set the error of f^{shown} to $\texttt{E} = 3$ (81% accuracy). The deployment error of both models was set to $\texttt{E}_{real} = 6$ to compute Overreliance reliably, make both models equally usable, and not misrepresent the true model error too much. We additionally controlled for the distribution of errors given by $\texttt{E}_{real}$ by ensuring that the 6 mistakes arise at the same time for all participants. Our experiment compared two conditions: the $\texttt{Blackbox\_E} = 3$ condition, where participants made decisions without seeing the model, and the $\texttt{Interpretable\_E} = 3$ condition, where they could additionally see it.

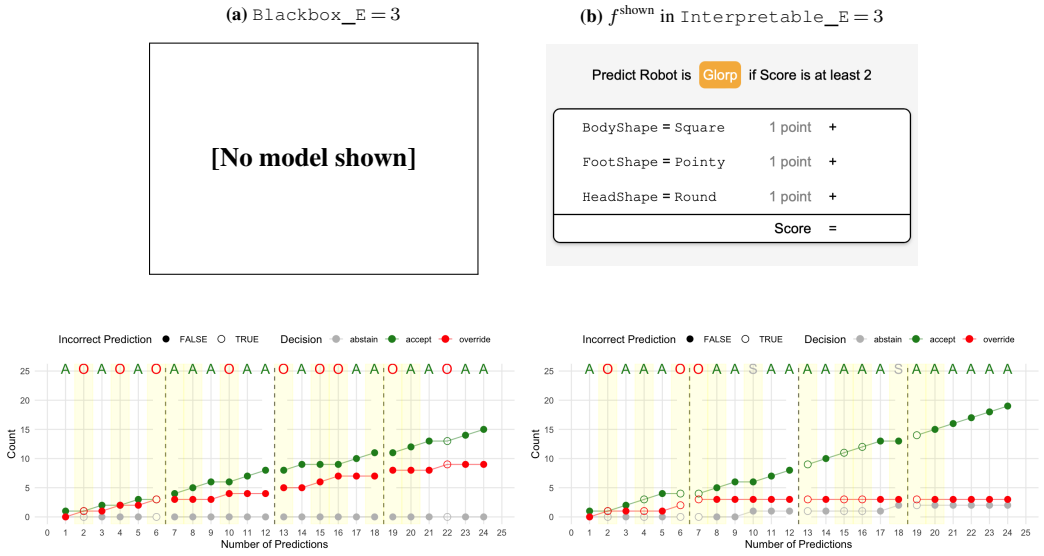


Figure 2: Archetypal examples of participant decisions in different experimental conditions. Yellow glow indicates points where f^{shown} and f^{user} differed. We can see that when participants did not see the model, they almost often overrode model's predictions, most often when it disagreed with their mental model. Once the model was made interpretable, however, they only overrode initially, making few mistakes, and subsequently relied on the model even if it was irrational.

Hypothesis Our assumption was that interpretability would lead to more cautious decision-making and reduce Reliance (**H1.1**) as well as prevent Overreliance (**H1.2**). The reasoning was that seeing the model fail would be more easily attributed to a culprit of a novel predictive pattern, whereas for blackbox the reasons for failure would be tied to specific robots. Because of this caution, we also posited that interpretability will lead to higher Correctness (**H1.3**), due to more correct decisions early in deployment.

Participants and Procedure We recruited 56 English-speaking participants on Prolific (age 19-73, avg. 36), requiring them to pass two tests: (1) demonstrate anchoring towards f^{user} during Probing (max 1 disagreeing prediction) and (2) pass a comprehension check on robot characteristics, model operation, and score function understanding. We accepted 48 submissions (11% exclusion rate), including 6 with one non-anchored prediction. Rejections were primarily due to failed comprehension checks (7) rather than non-anchored models (1). The 35-minute experiment offered $4 base pay plus performance bonus (avg. $3, max $6). Participants were randomly assigned to one of the mentioned 2 conditions with a task to supervise model f^{shown} knowing its error $\texttt{E} = 3$. They played the role of NASA advisors who determined if a model for predicting a particular robot class (a "Glorp") was correct. Each participant went through the same stages of the experiment.

3.1.1 Results

Our results reveal that interpretability hinders achieving appropriate reliance in a realistic AI-assisted decision-making scenario. Contrary to our hypotheses, we found that interpretability makes people rely significantly more on models with novel prediction patterns (**H1.1**), up to the point of Overreliance (**H1.2**).[1] The increased Reliance also lowered Correctness in the task (**H1.3**). These effects are shown Fig. 3, and can be also noticed in individual behavior in Fig. 2 (right).

Reliance: reached 75% ($Mean = 69\%, SD = 16\%$) for participants in the Interpretable_E = 3 condition, but dropped to 58% ($Mean = 62\%, SD = 15\%$) for participants in the Blackbox_E = 3 condition ($U = 205; p < 0.05$).

[1] In this and further experiments, our variables were not normally distributed and hence we used one-sided Man-Whitney U-tests and reported condition medians.

Override Rate: This drop was mainly driven by more Overrides in the `Blackbox_E=3` condition with 25% ($Mean = 25\%, SD = 13\%$), compared to 17% ($Mean = 17\%, SD = 10\%$) for the `Interpretable_E=3` condition ($U = 401; p < 0.01$)

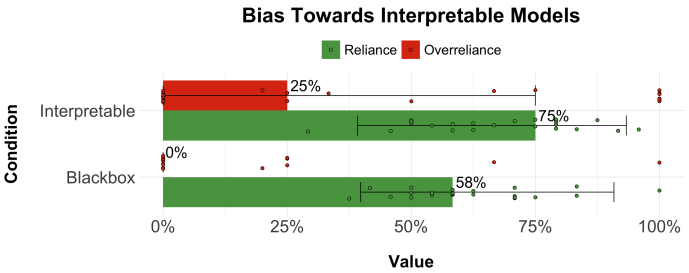


Figure 3: Reliance and Overreliance is significantly higher when seeing the Interpretable model.

Overreliance: was as high as 25% ($Mean = 40\%, SD = 44\%$) for participants in the `Interpretable_E=3` condition, but virtually non-existent at 0% ($Mean = 11\%, SD = 25\%$) for participants in the `Blackbox_E=3` condition ($U = 182; p < 0.01$)

Correctness: Participants in the `Interpretable_E=3` condition achieved Correctness of 78% ($Mean = 76\%, SD = 16\%$), roughly matching the model's accuracy. For `Blackbox_E=3` condition Correctness reached 85% ($Mean = 85\%, SD = 8\%$) which marked a significant increase ($U = 387.5; p < 0.05$).

Earnings Deviation: This reduced performance is also seen when comparing the rewards participants gathered at the end of the experiment. In the `Interpretable_E=3` condition the Earnings Deviation compared to the rational behavior was -9 ($Mean = -11.4, SD = 15.9$) but it was merely -2 ($Mean = -1.5, SD = 11.3$) in the `Blackbox_E=3` condition ($U = 385.5; p < 0.05$).

3.2 Post-Hoc: Overreliance on interpretable models can vary but is never eliminated

We performed additional experimentation to corroborate the effect we observed in Experiment 1 and to check what are its main drivers. We found that it appears irrespective of model accuracy, a traditional measure of trust in Human-Computer Interaction research, but changes with model format – a factor that was not previously considered. Experiments summarized in this section are described in more detail in the Appendix.

3.2.1 Experiment 2: Overreliance on interpretable models remains as we vary accuracy

Even when the model has substantially high accuracy or substantially low accuracy, we still observe overreliance on interpretable models. We repeated our experiment with an addition of 49 participants in two more conditions: `Interpretable_E=1` condition, where the model accuracy was 94%, and `Interpretable_E=6` condition, where the model accuracy was 63%. In the latter case we also increased the number of errors on the deployment set $\mathtt{E}_{real}$ to 9. The error cut-off was based on the logic that $6/16 = 40\%$ is the highest reasonable error level we would expect from a classifier or someone's internal model [72]. Despite different accuracy conveyed to participants, we observed no positive effect exerted on Overreliance. When the error was low in the `Interpretable_E=1` condition, Overreliance hit 40% ($Mean = 41\%, SD = 35\%$) even though Reliance itself decreased to compared to what we observed in Experiment 1 and was 65%. This means that participants experienced a form of an algorithmic aversion and were seemingly let down by the poorer accuracy of the model compared to the advertised one. Still, that decrease in Reliance did not eliminate Overreliance. When error was high (in the `Interpretable_E=6` condition), Reliance once again roughly matched the model's accuracy and reached 67%. But like in the former case, participant's Overreliance was as high as 47% ($Mean = 54\%, SD = 16\%$), meaning that participant's acceptances were misguided.

3.2.2 Experiment 3 and 4: Overreliance on interpretable models changes if we vary format

Surprisingly, models that are traditionally considered "equally interpretable"—those depending on the same number of features, printable on a sheet, and easy to simulate—can induce different degrees of Overreliance. While some variation in decisions could be attributed to individual beliefs, overall

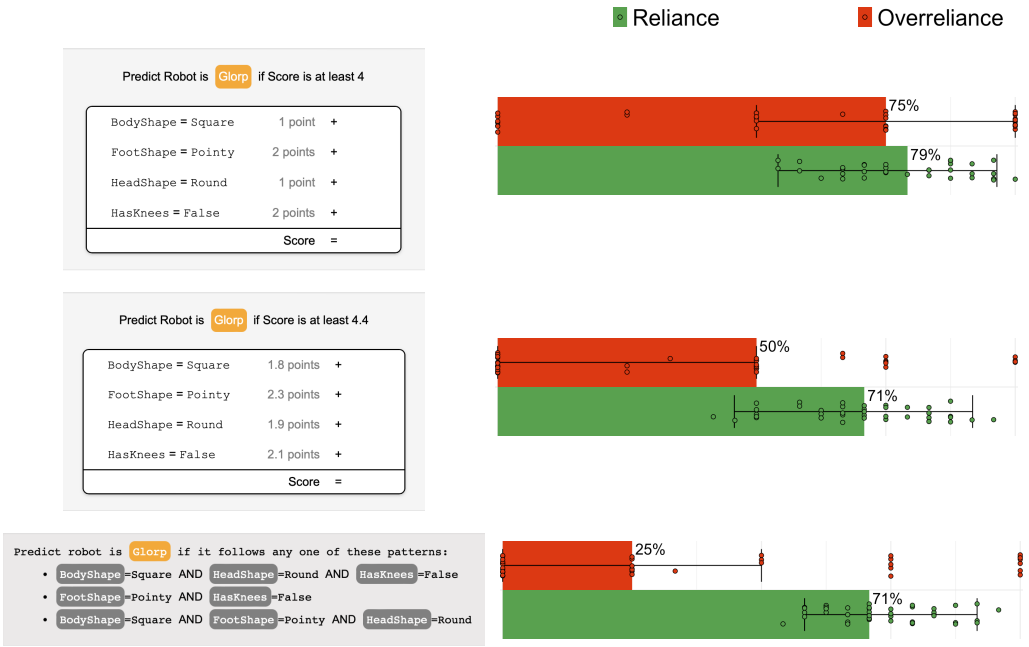


Figure 4: Different representations of the same model f^{shown} and their effect on Reliance and Overreliance.

model transparency, decision-making context, or user's cognitive resources [38], we would expect similar behaviors across a random population when these variables are fixed. Contrary to that expectation, we found that seemingly equivalent interpretable models can induce significantly different levels of Overreliance. To test this, we presented participants (N=116) with the same underlying model with 81% accuracy through 3 different representations: Score Function, DNF formula (OR of ANDs), and Non-integer Score Function (see Fig. 4). We found that participants' perceived complexity varied between conditions (26% rated Score Functions as "too complex", compared to 40% for Non-integer Scores and 46% for DNF) and correlated with behavioral differences:

Reliance: was highest with Score Functions ($Mean = 79\%$, $SD = 13\%$), significantly decreasing for DNF ($Mean = 74\%$, $SD = 11\%$; $U = 878; p < 0.05$) and Non-integer Scores ($Mean = 71\%$, $SD = 14\%$; $U = 474.5; p < 0.01$)

Overreliance showed even larger differences: Score Functions ($Mean = 61\%$, $SD = 38\%$, $Median = 75\%$) induced significantly more overreliance than both DNF ($Mean = 35\%$, $SD = 39\%$, $Median = 25\%$; $U = 970.5; p < 0.01$) and Non-integer Scores ($Mean = 37\%$, $SD = 35\%$, $Median = 50\%$; $U = 445.5; p < 0.01$).

4 Discussion

Our findings reveal the existence of an interpretability trap: people's propensity to overrely on interpretable models across different accuracy levels and model formats, to an extent as high as 75%. The controlling factor seems to concern perceived complexity of the interpretable model, rather than its clarity. We hypothesize that this effect aligns with opponent heuristics theory [29]: people judge simpler representations (like Score Functions) as having higher prior probability while assigning higher likelihood to complex ones. This would lead to optimal posterior probability estimates for moderately complex representations that consequently induce the highest trust and overreliance. This has concerning implications for AI-assisted decision-making, particularly in critical domains. There, interpretability trap may lead to accept clearly ill-motivated model predictions (like a medical diagnosis) simply because the model seemed expertly designed.

This result does not come without limitations. First, our Prolific participants may have had lower cognitive resources than real-world professionals, and the robot classification task may not have generated beliefs as strong as domain expertise. Both of these factors could have made following

the model a default strategy. Second, we didn't explicitly examine scenarios where users' mental models outperform AI and the effect of interpretability could be positive. Still, a realistic scenario reveals that the interpretability trap may occur. This underscores the importance to develop strategies that mitigate overreliance while preserving interpretability's benefits. Such strategies might include presenting competing models with similar accuracy but different logic, or finding ways to facilitate learning from the model logic without amplifying biases. We hope that our work lays the groundwork to explore these areas, and will eventually lead to the development of effective strategies that can harness interpretability – but avoid the *interpretability trap*.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] Daehwan Ahn, Abdullah Almaatouq, Monisha Gulabani, and Kartik Hosanagar. Impact of model interpretability and outcome feedback on trust in ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2024.

[3] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805, 2023.

[4] Genevera I Allen, Luqin Gan, and Lili Zheng. Interpretable machine learning for discovery: Statistical challenges and opportunities. *Annual Review of Statistics and Its Application*, 11, 2023.

[5] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9, 2020.

[6] Joel T Andrade et al. *Handbook of violence risk assessment and treatment: New approaches for mental health professionals*. Springer Publishing Company, 2009.

[7] Elliott M Antman, Marc Cohen, Peter JLM Bernink, Carolyn H McCabe, Thomas Horacek, Gary Papuchis, Branco Mautner, Ramon Corbalan, David Radley, and Eugene Braunwald. The timi risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making. *Jama*, 284(7):835–842, 2000.

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

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[10] Vaishak Belle and Ioannis Papantonis. Principles and practice of explainable machine learning. *Frontiers in big Data*, 4:688969, 2021.

[11] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[13] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

[14] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[15] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Effects of ai and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–42, 2023.

[16] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *arXiv preprint arXiv:2301.07255*, 2023.

[17] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58 (6):697–718, 2003.

[18] Brian J Edwards, Joseph J Williams, Dedre Gentner, and Tania Lombrozo. Explanation recruits comparison in a category-learning task. *Cognition*, 185:21–38, 2019.

[19] Pascal Friederich, Mario Krenn, Isaac Tamblyn, and Alan Aspuru-Guzik. Scientific intuition inspired by machine learning-generated hypotheses. *Machine Learning: Science and Technology*, 2(2):025027, 2021.

[20] Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. Do explanations help users detect errors in open-domain qa? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116, 2021.

[21] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, et al. A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. *Artificial intelligence review*, 56(4):3473–3504, 2023.

[22] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. A decision theoretic framework for measuring ai reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 221–236, 2024.

[23] Marisela Gutierrez Lopez and Susan Halford. Explaining machine learning practice: findings from an engaged science and technology studies project. *Information, Communication & Society*, pages 1–17, 2024.

[24] Tessa Han, Yasha Ektefaie, Maha Farhat, Marinka Zitnik, and Himabindu Lakkaraju. Is ignorance bliss? the role of post hoc explanation faithfulness and alignment in model trust. *arXiv preprint arXiv:2312.05690*, 2023.

[25] Edward R Hirt and Keith D Markman. Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of personality and social psychology*, 69(6):1069, 1995.

[26] Charles A Holt and Susan K Laury. Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655, 2002.

[27] Hsieh-Hong Huang, Jack Shih-Chieh Hsu, and Cheng-Yuan Ku. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decision Support Systems*, 53(3):438–447, 2012.

[28] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. Designing ai for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pages 1–14, 2021.

[29] Samuel GB Johnson, JJ Valenti, and Frank C Keil. Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cognitive psychology*, 113:101222, 2019.

[30] Martin Jones and Robert Sugden. Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50:59–99, 2001.

[31] Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *arXiv preprint arXiv:1804.02969*, 2018.

[32] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. Trust and reliance on ai—an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, 160:108352, 2024.

[33] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[34] Deanna Kuhn, Eric Amsel, Michael O'Loughlin, Leona Schauble, Bonnie Leadbeater, and William Yotive. *The development of scientific thinking skills.* Academic Press, 1988.

[35] Colton Ladbury, Reza Zarinshenas, Hemal Semwal, Andrew Tam, Nagarajan Vaidehi, Andrei S Rodin, An Liu, Scott Glaser, Ravi Salgia, and Arya Amini. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Translational Cancer Research*, 11(10):3853, 2022.

[36] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

[37] Olesja Lammert, Birte Richter, Christian Schütze, Kirsten Thommes, and Britta Wrede. Humans in xai: increased reliance in decision-making under uncertainty by using explanation strategies. *Frontiers in Behavioral Economics*, 3:1377075, 2024.

[38] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.

[39] Jennifer Marie Logg. Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*, 2017.

[40] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

[41] Nathan A Mahynski, Jared M Ragland, Stacy S Schuur, and Vincent K Shen. Building interpretable machine learning models to identify chemometric trends in seabirds of the north pacific ocean. *Environmental Science & Technology*, 56(20):14361–14374, 2022.

[42] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113:103655, 2021.

[43] Clifford R Mynatt, Michael E Doherty, and Ryan D Tweney. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29(1):85–95, 1977.

[44] Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Do machine learning models learn statistical rules inferred from data? In *International Conference on Machine Learning*, pages 25677–25693. PMLR, 2023.

[45] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[46] Josue Obregon and Jae-Yoon Jung. Rulecosi+: Rule extraction for interpreting classification tree ensembles. *Information Fusion*, 89:355–381, 2023.

[47] Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith T Butler. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research*, 3(6):597–607, 2022.

[48] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust in ai. 2019. URL https://sites.google.com/view/xai2019/home. IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019, XAI 2019 ; Conference date: 11-08-2019 Through 11-08-2019.

[49] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. It's complicated: The relationship between user trust, model accuracy and explanations in ai. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4):1–33, 2022.

[50] Michael J Pazzani. Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(3):416, 1991.

[51] Emmanuel Pintelas, Ioannis E Livieris, and Panagiotis Pintelas. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. *Algorithms*, 13(1):17, 2020.

[52] Forough Poursabzi-Sangdeh, Dan Goldstein, Jake Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *CHI 2021*, May 2021. URL https://www.microsoft.com/en-us/research/publication/manipulating-and-measuring-model-interpretability/.

[53] Amy Rechkemmer and Ming Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*, pages 1–14, 2022.

[54] José Ribeiro, Níkolas Carneiro, and Ronnie Alves. Black box model explanations and the human interpretability expectations–an analysis in the context of homicide prediction. *arXiv preprint arXiv:2210.10849*, 2022.

[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[56] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

[57] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[58] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

[59] Gabi Schaap, Tibor Bosse, and Paul Hendriks Vettehen. The abc of algorithmic aversion: not agent, but benefits and control determine the acceptance of automated decision-making. *AI & SOCIETY*, 39(4):1947–1960, 2024.

[60] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. Trust and reliance in xai–distinguishing between attitudinal and behavioral measures. *arXiv preprint arXiv:2203.12318*, 2022.

[61] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157): 1124–1131, 1974.

[62] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38, 2023.

[63] Charles Wan, Rodrigo Belo, Leid Zejnilović, and Susana Lavado. The duet of representations and how explanations exacerbate it. In *World Conference on Explainable Artificial Intelligence*, pages 181–197. Springer, 2023.

[64] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*, pages 318–328, 2021.

[65] Xinru Wang, Zhuoran Lu, and Ming Yin. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM Web Conference 2022*, pages 1697–1708, 2022.

[66] Frank W Weathers, Brett T Litz, Terence M Keane, Patrick A Palmieri, Brian P Marx, Paula P Schnurr, et al. The ptsd checklist for dsm-5 (pcl-5). 2013.

[67] Monika Westphal, Michael Vössing, Gerhard Satzger, Galit B Yom-Tov, and Anat Rafaeli. Decision control and explanations in human-ai collaboration: Improving user perceptions and compliance. *Computers in Human Behavior*, 144:107714, 2023.

[68] Joseph J Williams and Tania Lombrozo. The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive science*, 34(5):776–806, 2010.

[69] Joseph Jay Williams and Tania Lombrozo. Explanation constrains learning, and prior knowledge constrains explanation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

[70] Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, and Aijing Luo. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. *Journal of healthcare engineering*, 2023(1):9919269, 2023.

[71] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570. IEEE, 2018.

[72] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[73] Chang Ho Yoon, Robert Torrance, and Naomi Scheinerman. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, 48(9): 581–585, 2022.

[74] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 307–317, 2017.

[75] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020.

[76] Zelun Tony Zhang, Sven Tong, Yuanting Liu, and Andreas Butz. Is overreliance on ai provoked by study design? In *IFIP Conference on Human-Computer Interaction*, pages 49–58. Springer, 2023.

[77] Zelun Tony Zhang, Felicitas Buchner, Yuanting Liu, and Andreas Butz. You can only verify when you know the answer: Feature-based explanations reduce overreliance on ai for easy decisions, but not for hard ones. In *Proceedings of Mensch und Computer 2024*, pages 156–170. 2024.

[78] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW):1–23, 2019.

[79] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

A Related Work

Measuring Reliance and Inappropriate Reliance The most basic way to measure reliance on AI systems is to calculate the percentage of AI decisions that humans follow. While this provides a straightforward metric, it doesn't help us identify cases where the AI's presence inappropriately influences human decision-making. Neither does "weight of advice", that is a measure of how much people adjust their initial estimates toward AI suggestions [60, 52, 39]. This is because weight of advice only quantifies deviation from one's original judgment without assessing whether such adjustments were appropriate.

Surprisingly, however, there seems to be no one established definition of inappropriate reliance, and most importantly – overreliance, that is the excessive acceptance of the model's predictions. The majority of researchers define it as the percentage of incorrect AI decisions that users accept but would not have made independently [9, 16, 75, 11, 12]. But a recent work by Vasconcelos et al. [62] considered engaging or not engaging with an explanation as a determinant of overreliance – if users fail to spot incorrect AI solutions when given explanations, this indicates they overrelied. These are conceptually different definitions as even effortful consideration of the model's logic [62] might not lead to better performance [12] – because of other biases, faulty memory, incorrect beliefs about the environment, etc. The reason overreliance is treated differently by different authors might partially explain the discrepancies in the results.

Recently, Guo et al. [22] argued that prior approaches may mischaracterize inappropriate reliance, since observed behavior might actually be rational given the available information. Their framework defines reliance only for cases where human and AI predictions disagree, and models the rational agent as making a binary choice between following either the human or AI prediction based on which gives higher expected payoff. While this provides a principled way to measure appropriate reliance in some settings, it has limitations. Most notably, it cannot capture scenarios where neither the human nor AI prediction is correct and humans exhibit clear overreliance - such as accepting all AI predictions even when told the AI's accuracy is very low (e.g., 40%). To alleviate that problem, we propose altering the rational agent framework to model how the agent with full task information would behave when given three possible actions: accepting the AI's prediction, overriding it, or abstaining from decision. More details about our approach are provided in Section 2.1.

Effects of Interpretability on Reliance A growing body of work concerns defining properties of "whitebox" models [e.g., scrutinizability, faithfulness 42, 79, 21]), and shows how to create such models from scratch (see ensemble methods as in [51]) or how to obtain them via post-hoc transformation of blackbox models (e.g. [71, 46]). An even larger volume of work focuses on *explanations* – interpretable models or summaries that approximate blackbox model's logic on a subset of predictions (e.g. the popular LIME by Ribeiro et al. [55]). But to date, there is no clear evidence on whether interpretability calibrates or exacerbates appropriate reliance on AI models. One of the first papers to study the issue was by Poursabzi-Sangdeh et al. [52] where the authors found that people are less accurate when using interpretable models and often fail at spotting model errors. Some additional hints are given by research on explanations but even there, the results are inconclusive. On one hand, Bansal et al. [9] found that pairing humans with AI and explanations merely increased the chance AI predictions will get accepted irrespective of their correctness. These results could mean transparency – and interpretability – inevitably leads to overreliance because its existence somehow makes people trust the models more, also found in [16, 75, 11, 12]. But contrary to this claim, a notable work by Vasconcelos et al. [62] showed that overreliance may be controlled by a cost-benefit user analysis. The authors claim it is possible to see a reduction in overreliance in situations when studying an explanation is considerably easier than solving the problem itself. Reduction in overreliance was also seen in [77, 37, 16]. Yet another perspective is that there is possibly no effect, for instance because long time series of decision-making make people progressively complacent [76]. Similar null effects were reported in [2, 67].

To make things even less decisive, it is questionable if we eliminated confounding factors and learned anything about interpretability per se. Firstly, an extensive amount of prior work studied overreliance on interpretability under the guise of explanations. Participants saw images [62], feature-importance metrics [9, 12] or examples and prototypes [13, 16]. But as Rudin [57] points out in her seminal work, the scope and the validity of information conveyed by different explanations might affect or even mislead participant behavior. Secondly, we echo the design concern of [32] or Vasconcelos

et al. [62] who point out that most if not all existing work does not offer generalizable conclusions. It focuses on highly-specialized tasks such as recidivism prediction [64], tasks where participants have different prior beliefs like apartment-pricing [52], or applied tasks where participants have different capacities like time to examine the whitebox [28]. Lack of knowledge, strong convictions about one's beliefs, or an unmotivated approach might have all affected the final results of the known studies. We make an attempt to alleviate these issues by controlling for most aspects of the decision problem, as described in Section 2.

Our results are in line with research on explanations that show they increase overreliance. We go even further, providing evidence that even simple, transparent models make people overrely. This result opposes some of the remaining research, however. We believe this is due to how overreliance was operationalized in those studies (i.e. differently than irrational acceptance, c.f. [62]) and how the tasks depended on prior knowledge (c.f. [16, 12]) or how the model's logic was unequivocally incorrect (c.f. [62]) – contrary to the real-case scenario where things are rarely black-and-white.

Effects of Model Accuracy on Reliance Research consistently shows that model accuracy significantly influences how much people rely on AI systems. In a foundational study, Yin et al. [72] demonstrated that reported model accuracy affects both behavioral reliance (how often people follow model predictions) and self-reported trust, particularly when accuracy exceeds a minimal threshold (around 55%). Beyond this seminal work, numerous studies have confirmed that reliance generally increases with model accuracy, whether measured through behavioral metrics or self-reported trust [36, 74, 17, 48, 49].

The relationship between accuracy and reliance appears to be moderated by user experience and feedback, however. People tend to increase their reliance after observing that a model outperforms their own predictions, but this effect diminishes if the model's actual performance falls significantly below its reported accuracy [72]. While some researchers have explored whether model confidence or explanations might mediate this relationship, evidence suggests that observed accuracy remains the primary driver of reliance [53], while confidence levels or initial accuracy claims come secondary [48, 49].

A gap in the knowledge that we fill with out current study is how accuracy influences inappropriate reliance, especially overreliance, when paired with interpretability. We find that as much as accuracy remains an important signal for reliance, interpretability does not help in deciding when to rely on the model, and systematically leads to high overreliance.

Cognitive Biases as Potential Drivers of Reliance Several cognitive biases may confound studies of human reliance on AI systems. Confirmation bias - the tendency to favor information confirming existing beliefs - has been well documented in information search and interpretation [43, 45, 30, 34]. In human-AI interaction, this bias could lead participants to more readily accept AI predictions that align with their preconceptions while rejecting those that don't, regardless of the AI's actual reliability. Thus, the results of prior studies that focused on real-world environments and did not control for prior beliefs may be compromised [50, 65, 75, 52, 16, 27, 72]. For example, Poursabzi-Sangdeh et al. [52] found that participants in human-subject studies adhered to both simple and complex models equally often. This was taken as an indication that individuals don't necessarily favor simpler models over complex ones. Yet, it is plausible that the simpler models clashed with the participants' prior beliefs, leading to reluctance to follow the more straightforward model. Other studies estimated prior knowledge for each individual through questionnaires or experimental tasks at the beginning of the study, and compared subsequent behavior changes to this baseline [65, 75, 16]. When the baseline is not estimated on an individual-by-individual level, however, this could potentially invalidate the findings as well. We control for confirmation by using a novel prediction task (robot classification) where participants have no prior domain expertise or established beliefs.

Anchoring bias - the tendency to rely heavily on the first piece of information encountered - is another potential confounder [61]. When AI predictions are presented before human judgment, participants may unduly anchor on these predictions rather than engage in independent reasoning. To mitigate this, our experimental design requires participants to make their own predictions before seeing the AI's recommendations (see Section 2).

Explanation bias - the phenomenon where the act of generating an explanation for a hypothesis increases its perceived likelihood [25] - is particularly relevant to interpretability and explainability

research. When participants see model logic, they might engage in generating their own explanations of how the model works, increasing their perceived likelihood of these explanations. This could lead to strengthening reliance if the explanation process sought to confirm one's beliefs because of confirmation bias. Previous studies might thus conflate the effects of seeing the model logic with the effects of trusting own generated explanation for the model logic. In our study, we control for this by having all participants generate the same hypothesis during the Anchoring phase. We then isolate the effects of interpretability by comparing the decisions made by participants who saw the model versus those who did not see it (see Section 2).

Beyond these biases, model complexity itself may confound reliance measurements. Prior work suggesting mistrust of interpretable AI systems may have been affected by overly complex model presentations (e.g. [52]). We address this potential confounder by using maximally simple interpretable models - Score Functions and DNF formulas with at most 4 features and 3 conditions. These models fully disclose their decision logic while remaining easily comprehensible to non-experts (e.g. see Fig. 4).

B Experimental Design in Detail

B.1 Robot Classification Task

Robot Features We consider a simple binary classification task where participants label robots as Glorps or Drents. Our task is adapted from a task proposed by Williams and Lombrozo [69] that investigated how explaining the differences between categories (Glorps and Drents) affects employing prior knowledge in the discovery and use of classification patterns.

We define 4 categorical attributes $\boldsymbol{c} = (\texttt{HeadShape}, \texttt{BodyShape}, \texttt{FootShape}, \texttt{HasKnees})$ encoded as binary variables $\boldsymbol{x} = \big(\mathbb{1}[\texttt{HeadShape} = \textsf{Square}], \mathbb{1}[\texttt{BodyShape} = \textsf{Square}], \mathbb{1}[\texttt{FootShape} = \textsf{Pointy}], \mathbb{1}[\texttt{HasKnees} = \textsf{True}]\big)$. This gives us $|\mathcal{X}| = 2^4 = 16$ feature patterns (see Table 3 in the Appendix). We assume that each pattern specifies a Glorp or a Drent, $\mathcal{Y} = \{G, D\}$, deterministically – i.e., all robots with a fixed $\boldsymbol{x}$ are predicted the same $y \in \mathcal{Y}$. We create 96 robots to show to participants by re-using feature patterns and changing the visual display of the robot. We introduce 6 "subcategories" of $\texttt{FootShape}$ to increase the number of robots from 16 to $2 \times 2 \times 2 \times 6 = 48$ (see Fig. 1b). We then induce multiple color patterns and effectively increase the number of robots from 48 to 96.

Classification Models Models in our setting are mappings from robot characteristics to robot type, denoted as $f : \mathcal{X} \to \mathcal{Y}$. We define the participant mental model f^{user}, the model shown to the participants f^{shown}, and the true mapping f^{true}. Given that we consider 16 different robots, we have $2^{2^4} = 65,536$ possible functions f. We assume that the mental model f^{user} is linearly separable, which limits us to 1,886 functions $f \in \mathcal{F}$. We chose linearly separable models because they can be expressed both as a Disjunctive Normal Form formula (shortened DNF; an OR of ANDs of conditions on features) or a "score function", a simple linear model, e.g. $\text{sign}(2 * \texttt{BodyShape} = \textsf{Square} + 2 * \texttt{FootShape} = \textsf{Pointy} - 1)$, and change predictions monotonically with each feature. The latter means that switching the value of one feature always results in the same effect for the model prediction. As you can see in Table 1, XOR is an example non-linearly separable function that does not have a score function representation. Also, changing the value of $\texttt{FootShape}$ to Pointy may turn a Square $\texttt{BodyShape}$ robot into a Drent, but will turn a Round $\texttt{BodyShape}$ robot into a Glorp, having a non-monotonic effect for the feature change. We prioritize separable functions because we believe people more naturally form monotonic beliefs.

	Linearly Separable Model (OR)	**Linearly Inseparable Model (XOR)**
DNFs	$\texttt{BodyShape} = \textsf{Square}$ $\vee$ $\texttt{FootShape} = \textsf{Pointy}$	$\texttt{BodyShape}$=Square $\wedge$ $\texttt{FootShape}$=Flat $\vee$ $\texttt{BodyShape}$=Round $\wedge$ $\texttt{FootShape}$=Pointy
Score Functions	$\text{sign}(2 * \texttt{BodyShape} = \textsf{Square} +$ $2 * \texttt{FootShape} = \textsf{Pointy} - 1)$	–

Table 1: Classification models that are linearly separable (left) and linearly inseparable (right). As shown, the linearly separable function can be expressed as a score function with integer coefficients

Features				Models		
HeadShape	BodyShape	FootShape	HasKnees	f^{user}	f^{true}	f^{shown}
Round	Round	Flat	False	u_1	y_1	$\hat{y}_1$
Round	Round	Flat	True	u_2	y_2	$\hat{y}_2$
Round	Round	Pointy	False	u_3	y_3	$\hat{y}_3$
Round	Round	Pointy	True	u_4	y_4	$\hat{y}_4$
Round	Square	Flat	False	u_5	y_5	$\hat{y}_5$
Round	Square	Flat	True	u_6	y_6	$\hat{y}_6$
Round	Square	Pointy	False	u_7	y_7	$\hat{y}_7$
Round	Square	Pointy	True	u_8	y_8	$\hat{y}_8$
Square	Round	Flat	False	u_9	y_9	$\hat{y}_9$
Square	Round	Flat	True	u_{10}	y_{10}	$\hat{y}_{10}$
Square	Round	Pointy	False	u_{11}	y_{11}	$\hat{y}_{11}$
Square	Round	Pointy	True	u_{12}	y_{12}	$\hat{y}_{12}$
Square	Square	Flat	False	u_{13}	y_{13}	$\hat{y}_{13}$
Square	Square	Flat	True	u_{14}	y_{14}	$\hat{y}_{14}$
Square	Square	Pointy	False	u_{15}	y_{15}	$\hat{y}_{15}$
Square	Square	Pointy	True	u_{16}	y_{16}	$\hat{y}_{16}$

Table 2: Overview of the feature space and models. We consider 16 robots defined by four binary attributes: HeadShape, BodyShape, FootShape, HasKnees. Each combination of characteristics (row) maps to a Glorp or a Drent. We elicit the prior beliefs of a participant by querying the labels for f^{user}. Given f^{user} we present a model f^{shown} to and evaluate participant reliance through a simulation. We control the error of f^{user} and f^{shown} by specifying f^{true} as described in Appendix B.2.

B.2 Study Description

Participants go through four stages of the experiment: anchoring (to anchor their beliefs), probing (to capture their mental model), model selection (to select the ground truth and the shown model), and deployment (to compute participant's reliance and correctness on new data). We visualize what the models' predictions and the participant's decisions may be in each of the stages in Table 3.

Anchoring We show participants a subset of robots to anchor their beliefs to a certain model, e.g. BodyShape = Square $\wedge$ FootShape = Pointy (see Fig. 5). We then require them to correctly write the model down as a logical rule in DNF. Although the anchoring model is one of many possible models that classify the subset without error, it uses the fewest conditions. Thus, we expect participants to notice it first.

Probing During probing, participants specify the labels (Glorp or Drent) of 16 unique feature patterns represented by different robots (see Fig. 6). We use their responses to specify all 16 predictions for f^{user}. Since we aim to elicit one particular model, we restrict our analyses to participants who make identical or almost identical (1 incompatible choice) predictions compared to that one model.

Model Selection In the model selection stage, we use a set of parameters and participant's responses from the Probing Stage to specify the predictions of two classification models: f^{true} and f^{shown}.

We chose f^{shown} from a subset of models that are bound to correctly label robots in the anchoring set. We can restrict that subset according to several input parameters:

- edit distance in the feature space compared to mental model f^{user} (i.e. the minimum number of conditions to add or remove to transform one model into another)
- minimum number of different predictions on unique robots compared to the mental model f^{user}
- number of conditions in the DNF formula representing f^{shown}
- whether f^{shown} should use the same features as f^{user} or not

After Anchoring			After Probing			After Model Selection			After 5 Robots in Deployment				
f^{user}	f^{true}	f^{shown}	f^{user}	f^{true}	f^{shown}	f^{user}	f^{true}	f^{shown}	f^{user}	f^{true}	f^{shown}	d_i^{init}	d_i^{final}
G	G		G	G		G	G	G	G	G	G	G	G
G	G		G	G		G	G	G	G	G	G	G	G
G	G		G	G		G	G	G	G	G	G	G	G
D	D		D	D		D	D	D	D	D	D	G	D
D	D		D	D		D	D	D	D	D	D	$\perp$	D
D	D		D	D		D	D	D	D	D	D		
			G			G	D	G	G	D	G		
			G			G	D	D	G	D	D		
			G			G	D	G	G	D	G		
			G			G	D	D	G	D	D		
			D			D	G	G	D	G	G		
			D			D	G	D	D	G	D		
			G			G	D	G	G	D	G		
			D			D	G	D	D	G	D		
			D			D	G	G	D	G	G		
			D			D	G	D	D	G	G		

Table 3: The output of models and the participant's decisions for each robot type at key stages of the experiment. We use $\perp$ to indicate abstention from predicting. Here, f^{user} is the model that outputs the labels specified by the participant during the probing stage. f^{true} is the ground truth model that outputs the true types of each robot. f^{shown} is the model that we show the participant in the deployment stage. We choose f^{true} and f^{shown} once the participant has completed the probing stage and thus specified all of the outputs for f^{user} – so that we can present the participant with f^{shown} with a desired level of error and disagreement with the participant's prior beliefs. d^{init} and d^{final} are the participant's initial robot labels during deployment, and final robot labels during deployment, elicited before and after seeing the model prediction, respectively.

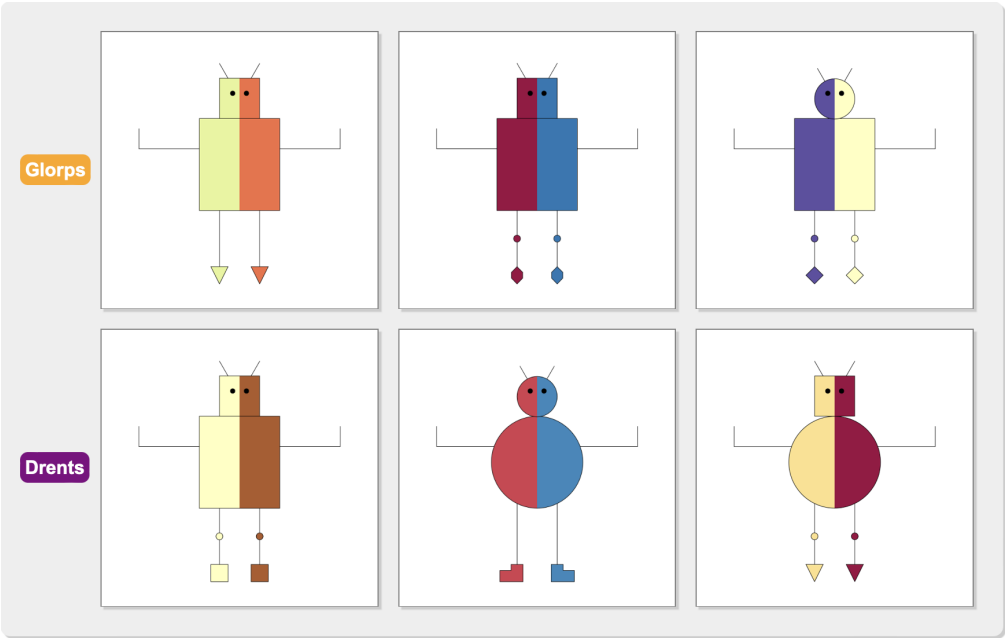


Figure 5: Sample interface of the anchoring stage, with the anchoring set of 3 Glorps and 3 Drents that cover 6 feature patterns. The robots are chosen so that all Glorps had Square `BodyShape` and Pointy `FootShape`, but neither of Drents did, the robots used the largest number of distinct features possible, and the set supported f^{shown} with provided input properties.

Such a restriction allows us to find models that are similar or dissimilar to the mental model f^{user} using precisely controllable parameters. After applying the restriction, we limit the set of viable models to K, out of which we select one model randomly.

Given the mental model f^{user} and the shown model f^{shown}, we specify f^{true} so that f^{shown} made e errors on the 16 unique feature patterns and achieved a desired level of error $\delta = e/16$, and the error

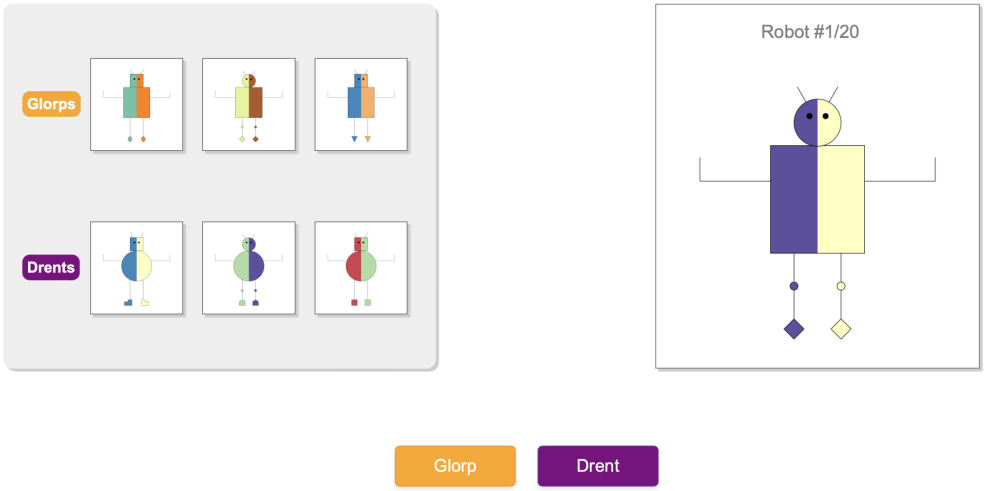


Figure 6: Experimental interface during the Probing stage. Participants see the anchoring set they know from the Anchoring stage and an image of a robot. Their task is to specify the robot type by clicking on a button (Glorp or Drent).

of f^{user} was as close to δ as possible. Formally, f^{true} is the solution to the optimization problem:

$$\min_{f \in \mathcal{F}} \quad \Pr(f^{\text{user}}(\boldsymbol{x}) \neq f(\boldsymbol{x})) - \Pr(f^{\text{shown}}(\boldsymbol{x}) \neq f(\boldsymbol{x})) \tag{1}$$

$$\text{s.t.} \quad \Pr(f^{\text{shown}}(\boldsymbol{x}) \neq f(\boldsymbol{x})) = \delta \tag{2}$$

Deployment We use the Deployment stage to measure participant's reliance and other metrics on a subset of 24 previously unseen robots. The robots are divided into 4 rounds, 6 robots each. For each robot, the participant is first asked for their prediction, then shown the prediction from f^{shown}, then asked whether they wish to accept the prediction of f^{shown}, override it, or *abstain* (skip predicting).

To measure reliance well, we chose deployment robots such that they contain 12 robots where f^{user} and f^{shown} disagree and 12 robots where they agree. These robots are distributed equally between the rounds, and positioned randomly within the rounds. We also make sure that the error of the shown model f^{shown} on the deployment robots is equal to a parameter, and the error of f^{user} is as close to the error of f^{shown} as possible.

We incentivize participants to respond correctly and thoughtfully by specifying a virtual reward and time penalty for each response. Specifically, each participant receives:

- \$1 if they accept a correct prediction:
- \$5 if they override an incorrect prediction
- -\$3 if they override a correct prediction + a 10 second time penalty
- \$0 if they choose to skip

We designed the interface of the Deployment stage to ensure that participants are informed in a way that is not overbearing (see Fig. 1a). They are shown the robot, the model, the model's prediction with the conditions that are met for the current robot, and a summary of their performance to date. They also have access to the anchoring set of robots. In addition, participants see a summary of their performance after each round, both globally and on the last 6 robots. These measures make sure that participants are provided with as much information as possible to make an informed decision and understand what their decision entails.

Deployment is also when participants answer a number of survey questions. After model selection and right before deployment, participants are asked over 20 questions about their understanding of f^{shown} and its similarity to their mental model f^{user}. This gives us qualitative data on participant's perceptions of the model.

B.3 Measuring Inappropriate Reliance

After deployment, we gather the following data:

- the sequence of robot feature patterns $\mathbb{x} = (\mathbf{x_1}, \ldots, \mathbf{x_{24}})$, where $\mathbb{x}[i] = \mathbf{x_i} \in \mathcal{X}$ is a 4-tuple of robot features
- the sequence of ground truth robot types $\mathbf{y} = (y_i, \ldots, y_{24})$ with $\mathbf{y}[i] = y_i \in \mathcal{Y}$ is either G(lorp) or D(rent)
- the sequence of the shown model's predictions $\mathbf{f} = (f^{\text{shown}}(\mathbf{x_1}), \ldots, f^{\text{shown}}(\mathbf{x_{24}})$ where again $\mathbf{f}[i] = f_i = f^{\text{shown}}(\mathbf{x_i}) \in \mathcal{Y}$
- the sequence of user deployment decisions $\boldsymbol{a} = (a_1, \ldots, a_{24})$, where $a_i \in \{\text{Accept}, \text{Override}, \text{Skip}\}$

The convention is that by adding superscript t to any of the sequences, e.g. $\mathbf{f}^t$, we refer to the sequence up to point t. By using these sequences we quantify user behavior by computing the following decision metrics:

- Reliance, which expresses agreement with the model,
- Overreliance, which expresses the tendency to accept model predictions too often,
- Earnings Deviation, which expresses the performance gap between the final reward and the expected reward
- Correctness, which quantifies overall decision accuracy

To define people's tendency to perform certain actions excessively, we turn to the approach suggested by Guo et al. [22], and define the *rational agent* first. The rational agent formalizes behavior expected when considering the stakes in the decision problem and the learned information.

B.3.1 Rational Agent

Intuitively, the rational agent should make decisions invariant to robot color and robot subtype (visual features introduced to increase the number of different robots for the experiment) and try to maximize its reward. To do so, the agent needs to recognize each robot on the basis of its unique feature pattern $\mathbf{x}$ (see Table 2). It also needs to keep a tally of known true labels of the patterns it has uncovered and prioritize actions specifying the ground truth label of the robot beginning from the second encounter onwards. For robots whose ground truth label is unknown, the agent should make decision based on the expected reward.

To properly define the rational agent as specified above we need to introduce the concept of states, actions and rewards. States represent all information available to the agent when making a decision in round t: current robot, model prediction and historical data. Formally, using the superscript notation we set $s \in \mathcal{S} = (\boldsymbol{x}, f^{\text{shown}}(\boldsymbol{x}), (\mathbb{x}^t, \boldsymbol{f}^t, \boldsymbol{y}^t))$. Actions in this understanding are simply decisions regarding each robot: $a \in \mathcal{A} = \{\text{Accept}, \text{Override}, \text{Skip}\}$. The reward for action a in state s is finally given by the incentives specified in the experiment. The numbers were chosen to incentivize overriding:

$$r(a, s) = \begin{cases} 0 & \text{if } a = \text{Skip} \\ 1 & \text{if } a = \text{Accept and } f^{\text{true}}(\boldsymbol{x}) = f^{\text{shown}}(\boldsymbol{x}) \\ 5 & \text{if } a = \text{Override and } f^{\text{true}}(\boldsymbol{x}) \neq f^{\text{shown}}(\boldsymbol{x}) \\ -3 & \text{otherwise} \end{cases} \tag{3}$$

To simplify the notation, let us denote the reward under correct and incorrect model prediction as

$$r_T(a, s) := r(a, s \mid f^{\text{shown}}(\boldsymbol{x}) = f^{\text{true}}(\boldsymbol{x})) \qquad \text{(shown model is correct)}$$
$$r_F(a, s) := r(a, s \mid f^{\text{shown}}(\boldsymbol{x}) \neq f^{\text{true}}(\boldsymbol{x})) \qquad \text{(shown model incorrect)}$$

The expected reward is the probability the model is correct or incorrect times the reward associated with each outcome. However since the agent accumulates knowledge, the said probability changes. We will refer to it as the expected accuracy in round t, p_t. To define p_t, we need to note the predictions that the algorithm got right and wrong so far. Initially, participants are told the accuracy of the model

is δ and that means it correctly labeled R out of N robots they have seen during the Probing stage. With that prior knowledge, p_t is an update to these numbers:

$$p_t = \Pr(f^{\text{shown}}(\mathbf{x}) = f^{\text{true}}(\mathbf{x}) \mid (\mathbb{x}^t, \boldsymbol{f}^t, \boldsymbol{y}^t)) =$$
$$= \begin{cases} 0 & \text{if } \exists i: \ \mathbf{x} = \mathbb{x}[i] \wedge \mathbf{f}[i] \neq \mathbf{y}[i] \\ 1 & \text{if } \exists i: \ \mathbf{x} = \mathbb{x}[i] \wedge \mathbf{f}[i] = \mathbf{y}[i] \\ \frac{R + \left|\{i \in [t]: f^{\text{shown}}(\boldsymbol{x}_i) = y_i\}\right|}{t+N} & \text{otherwise} \end{cases}$$

Finally, we define the rational agent through its decision-making policy $\pi : \mathcal{S} \to \mathcal{A}$. $\pi(s_t)$ is the action that maximizes the expected reward:

$$\begin{aligned} \pi(s_t) &\in \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[r(a, s)\right] \\ &= \Pr(f^{\text{shown}} = f^{\text{true}} \mid (\mathbb{x}^t, \boldsymbol{f}^t, \boldsymbol{y}^t)) \cdot r_T(a, s) \\ &\quad + (1 - \Pr(f^{\text{shown}} = f^{\text{true}} \mid (\mathbb{x}^t, \boldsymbol{f}^t, \boldsymbol{y}^t))) \cdot r_F(a, s)) \\ &= p_t \cdot r_T(a, s) + (1 - p_t) \cdot r_F(a, s) \\ &= p_t \cdot (r_T(a, s) - r_F(a, s)) + r_F(a, s) \end{aligned}$$

According to the rewards we set in the Deployment stage, the incentives for each action at time t are

- Accept : $4p_t - 3$
- Override : $5 - 8p_t$,
- Skip : 0.

Thus, the agent is incentivized to Override for $p_t \leq \frac{5}{8}$, it is incentivized to Skip when $p_t \in (\frac{5}{8}, \frac{3}{4})$, and it favors Accept when $p_t \geq \frac{3}{4}$. For example, if the model's initial accuracy is 50%, the agent would be incentivized to Override in the first round and subsequent rounds (when seeing new unique robots), until realizing the model's accuracy may actually be at least 62.5%. The agent would keep skipping whenever the model's accuracy was said to be between 62.5% and 75%, and accepting when the accuracy was above 75%.

The rational agent becomes an oracle agent – i.e., an agent enacting the ground truth – after observing labels of all 16 unique labels. We denote the policy of the oracle agent as π^*.

B.3.2 Metrics Formalization

Having the rational's agent policy π we define overrreliance as the proportion of irrational Accepts, i.e. Accepts among all cases when the agent would Override or Abstain. Reliance remains the Accept rate. All the metrics are defined in round t:

$$\text{Reliance}(\boldsymbol{a}^t) = \Pr(\text{Accept}) = \tag{4}$$
$$= \tfrac{1}{t} \left|\{i \in [t] : \ a_i = \text{Accept}\}\right|$$
$$\text{Overreliance}(\boldsymbol{a}^t, \boldsymbol{s}^t) = \Pr(\text{Accept} \mid \text{Should not Accept}) = \tag{5}$$
$$= \tfrac{|\{i \in [t]: \ \text{Accept} = a_i \neq \pi(s_i)\}|}{|\{i \in [t]: \ \pi(s_i) \neq \text{Accept}\}|}$$

We quantify the cost of irrational behavior in terms of correctness, and the change in the earnings. We measure the relative loss in earnings with respect to an agent with policy π as:

$$\text{Earnings Deviation}(\boldsymbol{a}^t, \boldsymbol{s}^t) = e_t(a_t) - e_t(\pi(s_t)) \tag{6}$$

Here $e_{t+1}(a) = \max(0, e_t + r(a, s))$ denotes cumulative earnings in round $t+1$ based on action a in state s and $e_0(a) = 0$. Correctness$(\pi, \boldsymbol{s}^t)$ is the proportion of correct predictions over all decisions other than abstention (skipping). We can formalize it using the oracle agent's policy π^* that sets the ground truth action in each step:

$$\text{Correctness}(\mathbf{a}^t, \boldsymbol{s}^t) = \Pr(\text{correct prediction}) = \tag{7}$$
$$= \tfrac{|\{i \in [t]: \ a_i = \pi^*(s_i)\}|}{|\{i \in [t]: \ a_i \neq \text{Abstain}|}$$

We omit the arguments in our measures to indicate they are computed over all 24 deployment robots.

B.4 Screens

Participants were randomly assigned to one of the conditions, and consequently the shown model f^{shown}, and shown model error $\mathtt{E}$. In each condition, participants played the role of NASA advisors who determined if a model for predicting a particular robot class ("Glorp" instead of "Drent") was correct. Each participant saw the same set of screens in the following order:

1. Consent screen,
2. (Anchoring stage) General info screen,
3. (Anchoring stage) Robot instructions screen explaining robot features and introducing the NASA setup,
4. (Anchoring stage) Attention quiz (max 3 attempts) about robot features where participants needed to choose feature values being shown a sample robot,
5. (Anchoring stage) Anchoring screen with the anchoring set participants would have access to for the remainder of the study
6. (Anchoring stage) Instructions on what logical conjunctions are (called prediction patterns) and how they can predict Glorps
7. (Anchoring stage) Quiz of understanding prediction patterns
8. (Anchoring stage) Anchoring screen with the anchoring set; participants needed to explicitly state f^{user} by choosing features, values, and logical connections between the feature-value pairs to create "their rule for deciding when a robot is a Glorp"
9. (Probing stage) A series of 24 robot images for which participants were asked to use the rule they developed in the Anchoring stage by labeling robots as either "Glorps" or "Drents"
10. (Probing stage) Probing summary screen, where participants could move images robots around to finalize their division into "Glorps" and "Drents"
11. (Probing stage) Probing finale screen informing that participants got over half of the labels correctly or not (for participants who were rejected due to incorrect labeling)
12. (Deployment stage) Three rule instructions screens that explained the premise of classification models and introduced score functions
13. (Deployment stage) Comprehension quiz (max 3 attempts) that asked participants to apply score functions to robots and create a score function for f^{user} (by copying the format of another score function)
14. (Deployment stage) Model selection screen that showed participants the rule developed by NASA they would supervise – f^{shown} – and its accuracy $1 - \mathtt{E}$
15. (Deployment stage) Questionnaire page with a list of 30 multiple-choice questions on model understanding, its complexity, perceived similarity to f^{user}, etc.
16. (Deployment stage) Two model questionnaires that asked about participants' attitudes towards the model
17. (Deployment stage) Three deployment instructions screens that explained the deployment stage and the reward structure; each participant started with $5 virtual dollars and received $1 for correct acceptance, $5 for correct override or lost $3 for taking an incorrect action other than abstention
18. (Deployment stage) A series of 24 deployment screens that asked participants to predict robot type and then decide whether to Accept, Override or Abstain from prediction
19. (Deployment stage) Deployment summaries with the number of Accepts, Overrides and Abstains as well as correct and incorrect decisions and rewards every 6 robots
20. (Deployment stage) Questionnaires about participant's changing attitude towards the model every 6 robots
21. (Deployment stage) Questionnaire on risk aversion taken from [26] where participants were choosing between bets
22. (Deployment stage) Questionnaire that asked about participants' attitude towards f^{user}, if not previously completed
23. Reward and thank you screen

C IRB Approval

This research was conducted under the exemption of the Institutional Review Board at University of California, San Diego, Protocol #806253, granted on 02, Feb, 2023. All participants provided informed consent before participating in the study.

D Experiment 2: Overreliance on interpretable models remains even as we vary model accuracy

The experiment was identical to Experiment 1 but for the fact that $\texttt{E}_{real} = 9$ when $\texttt{E} = 6$ to better match the model real accuracy during deployment. We recruited 63 participants (23-26 per condition, age 18-66, English speaking) and accepted 49 submissions (22% exclusion rate). Among the excluded participants, 4 failed comprehension checks (6%) and 10 were not anchored to the mental model f^{user} (16%). We also accepted 4 participants (6%) whose decision during the Probing stage made 1 prediction that disagreed with the planned f^{user}. The experiment lasted an average of 39 minutes. Participants were given a base pay of $\$4$ and a performance bonus of up to $\$6$. On average participants received a bonus of $\$3$.

E Experiment 3 and 4: Overreliance on interpretable models changes if we vary model format

Firstly, we strategically compared two highly interpretable representations: Score Functions and DNF formulas, asking **whether there are differences in reliance and overreliance when the same model is represented in two different interpretable formats**. We chose Score Functions as they are simple, sparse linear classification models, validated though years of practice in clinical or judicial decision-making [6, 7, 66, 33]. We chose DNF formulas, as they are logical alternatives of prediction patterns that, arguably, offer the most straightforward representation of the model logic due to their most reduced, atomic form. We hypothesized that irrespective of the high model interpretability, we would observe differences in Reliance and Overreliance. The reasoning was that participants seeing the Score Function would focus more on how single features validate their mental model, leading to higher acceptance. In contrast, participants seeing the DNF would understand the differences better, making them more cautious.

E.1 Experiment 3: Overreliance can vary with different model representations

E.1.1 Setup

We strove to make f^{shown} as complex as possible to maximize the differences between the two tested representations. This would enable us to more easily extract the effect of model representation. To do this, but still maintain intelligibility, we picked f^{user} as the 2-feature model and set $\texttt{S} = 3$ to have 3 terms in the DNF. We also wanted to maximize Alignment Distance $\texttt{D}$, and found that the upper limit is 6 for DNFs. We also selected disagreement of at least $\delta = 4$ compared to the mental model f^{user}. This procedure limited the set of models to 2, out of which we selected one model randomly:

$$\texttt{BodyShape} = \textsf{Square} \wedge \texttt{HeadShape} = \textsf{Round} \wedge \texttt{HasKnees} = \textsf{False}$$
$$\vee$$
$$\texttt{FootShape} = \textsf{Pointy} \wedge \texttt{HasKnees} = \textsf{False}$$
$$\vee$$
$$\texttt{BodyShape} = \textsf{Square} \wedge \texttt{FootShape} = \textsf{Pointy} \wedge \texttt{HeadShape} = \textsf{Round}$$

To consider a representative case in terms of model performance, we focused on $\texttt{E} = 3/16$ and $\texttt{E}_{real} = 6/24$, to be able to compute Overreliance reliably. Importantly, the deployment error of f^{user} was equal to the value of $\texttt{E}_{real}$, allowing us to reduce the effect of one of the models being superior. We also needed to control an additional source of variation, namely the distribution of errors given by $\texttt{E}_{real}$. To that end, we ensured that the 6 mistakes arise at the same time. We named our conditions `DNF_Complex` and `Score_Complex`.

Hypothesis We hypothesized that the Score Function representation would elicit higher Reliance and Overreliance compared to the DNF representation. This hypothesis was based on two premises.

Firstly, we assumed that participants would perceive the Score Function as more similar to their own mental model, leading to increased trust and reliance. Secondly, we assumed that participants would focus more on the prediction patterns in the DNF and that will enable them to make more informed decisions.

H3.1: *Score Function representation elicits higher Reliance on the tested interpretable model compared to the DNF representation*

H3.2: *Score Function representation elicits higher Overreliance on the tested interpretable model compared to the DNF representation*

Procedure Participants were randomly assigned to one of the 2 conditions: `DNF_Complex`, where they've seen the model as a DNF formula, and `Score_Complex`, where they've seen it as a Score Function. In both conditions, participants were asked to supervise model f^{shown} knowing its error $\texttt{E} = e$. They played the role of NASA advisors who determined if a model for predicting a particular robot class (a "Glorp") was correct. Each participant went through the Anchoring, Probing, Model Selection, and Deployment stages as described previously, and saw the same set of screens (see more details in the Appendix).

Participants We recruited 112 participants on Prolific (56 participants per condition, age 18-70, avg. 35, English speaking) among which we eventually accepted 76 submissions (exclusion rate 32%). Our acceptance criteria required participants to pass 2 tests as in Experiment 1. In the second comprehension check we additionally required participants to pass a quiz on prediction patterns, (logical conjunctions).

Although the exclusion rate is high, most of it was driven by participants whose prior beliefs we couldn't control, i.e. 16 participants (14%) used a model different than the anchored one during the Probing stage. The remaining 20 participants (18%) were excluded because they failed to pass a comprehension check on understanding the robot characteristics, prediction patterns (or terms) in the DNF, and the ability to simulate simple model predictions. Among the accepted participants, 9 participants (8%) made 1 prediction that disagreed with the anchored f^{user} during the Probing stage and who were accepted nevertheless. The experiment lasted an average of 43 minutes. Participants were given a base pay of $6 and a performance bonus of up to $6. On average participants received a bonus of $3.

E.1.2 Results

Generally, we found that participants in the `DNF_Complex` condition had lower Reliance and Overreliance than their counterparts in the `Score_Complex` condition (see Fig. 4).[2] This led to higher Correctness and better Earnings Deviation. Our analysis on the main mechanism of this change was inconclusive, however: it was either a better grasp of the model logic or dismissing the model logic whatsoever the `DNF_Complex` condition. We tackled this problem in Experiment 2

On Decisions Our results supported both hypotheses H1.1 and H1.2 regarding Reliance and Overreliance. We found that seeing the model as a DNF reduces both measures.

- Reliance reached 79% ($Mean = 79\%, SD = 13\%$) for participants in the `Score_Complex` condition, but dropped to 71% ($Mean = 74\%, SD = 11\%$) for participants in the `DNF_Complex` condition ($U = 878; p < 0.05$)
- Overreliance was as high as 75% ($Mean = 61\%, SD = 38\%$) for participants in the `Score_Complex` condition, but only 25% ($Mean = 35\%, SD = 39\%$) for participants in the `DNF_Complex` condition ($U = 970.5; p < 0.01$)

Interestingly, we observed more Overrides in the `DNF_Complex` condition at 17% ($Mean = 18\%, SD = 10\%$), compared to 12% ($Mean = 14\%, SD = 12\%$) for the `Score_Complex` condition ($U = 511.5; p < 0.05$). These results suggest that the difference between the participants in both conditions is making fewer accepts at the cost of more overrides.

[2] In this and further experiments, our variables were not normally distributed and hence we used one-sided Man-Whitney U-tests and reported condition medians.

On Correctness and Earnings While not statistically significant ($p = 0.06$), we observed a trend towards higher Correctness in the DNF_Complex condition (81% against 74% for Score_Complex condition). This trend, combined with the previous results, indicates that participants in the Score_Complex condition erroneously relied on the model when it was incorrect.

We also found significantly lower Earnings Deviation $= -26$ in the Score_Complex condition ($Mean = -18.7, SD = 11.7$) compared to the DNF_Complex condition with Earnings Deviation $= -9$ ($Mean = -10.8, SD = 14$), $U = 458.5; p < 0.01$. These sizeable differences indicate that the mistakes made by participants in the Score_Complex condition were more costly, meaning they did not only pertain to incorrect acceptances but also incorrect overrides, indicating reduced ability to reason about model and own errors.

On In-depth Analysis Participants' perceptions of the models revealed interesting insights:

- 46% of participants in the DNF_Complex condition found the model too complex, compared to 26% in the Score_Complex condition.
- DNF_Complex condition participants reported higher confidence in their decisions, despite perceiving the model as more complex (75% to 60%).
- Contrary to our expectations, perceived similarity to participants' own mental models was equivalent between conditions (35% found the models similar or identical).

This led us to investigate the reasons for the differences in Overreliance in both conditions as we hypothesized complexity might have discouraged participants from using the DNF representation. Our post-hoc analysis revealed distinct patterns in how participants adapted to observed model errors across conditions. After excluding approximately 10% of participants who exhibited perfect performance (suggesting they developed their own prediction strategies), we found significant differences in error adaptation between conditions.

Firstly, we found that our Overreliance measures focused on four specific robots during deployment: Robots 14 and 16, that both followed the prediction pattern HasKnees $=$ False $\wedge$ FootShape $=$ Pointy. This pattern caused f^{shown} to make 2 erroneous predictions and 1 correct prediction on 3 different robots. Then it appeared 4 more times during deployment, repeating Robots 14 and 16. The last of these 3 partially-correct predictions was also the cut-off point when the rational agent had gathered sufficient information to avoid making mistakes on this pattern. What happens if participants see the model logic however and presumably, incorporate it to reason about the correctness of the predictions? Participants in the Score_Complex condition showed a concerning behavior: their performance after the cut-off point actually deteriorated. Specifically, participants in the Score_Complex condition, having observed enough errors on robots with HasKnees $=$ False $\wedge$ FootShape $=$ Pointy, went from 67% incorrectness (3 accepts where 2 were incorrect and 1 was correct) to 100% (4 incorrect accepts) – the change was $-33\%(Mean = -9\%)$. Participants in the DNF_Complex condition, while showing no significant improvement ($Median = 0\%, Mean = 17\%$), avoided this deterioration ($U = 352, p < 0.05$). They went from 67% incorrectness to around 50% incorrectness. This means that they made 2 correct predictions among the 4 robots where we computed Overreliance. This is what happened for an average participant. When it comes to individual participants, analysis of error adaptation approaches (see Fig. 7) revealed two dominant strategies across conditions: no improvement or complete adaptation. While most participants in the Score_Complex condition showed no improvement, the majority in DNF_Complex condition achieved perfect adaptation. Some participants did exhibit intermediate learning patterns, though these were less common. This can be found in Fig. 7.

A closer examination of these cases provided further insights into adaptation patterns. For Robot 14, participants in the DNF_Complex condition showed substantial improvement, reducing incorrect predictions by 50% (i.e. still made 1 wrong prediction after seeing the error, mirroring the finding about the overall incorrectness). Robot 16 showed a similar but weaker pattern, likely due to the earlier occurrence of model errors for this robot affecting learning. Among participants who did adapt their behavior in the DNF_Complex condition, the adaptation was notably strong (100% drop in incorrectness for Robot 16, 50% for Robot 14), suggesting that when participants chose to modify their strategy, they did so decisively.

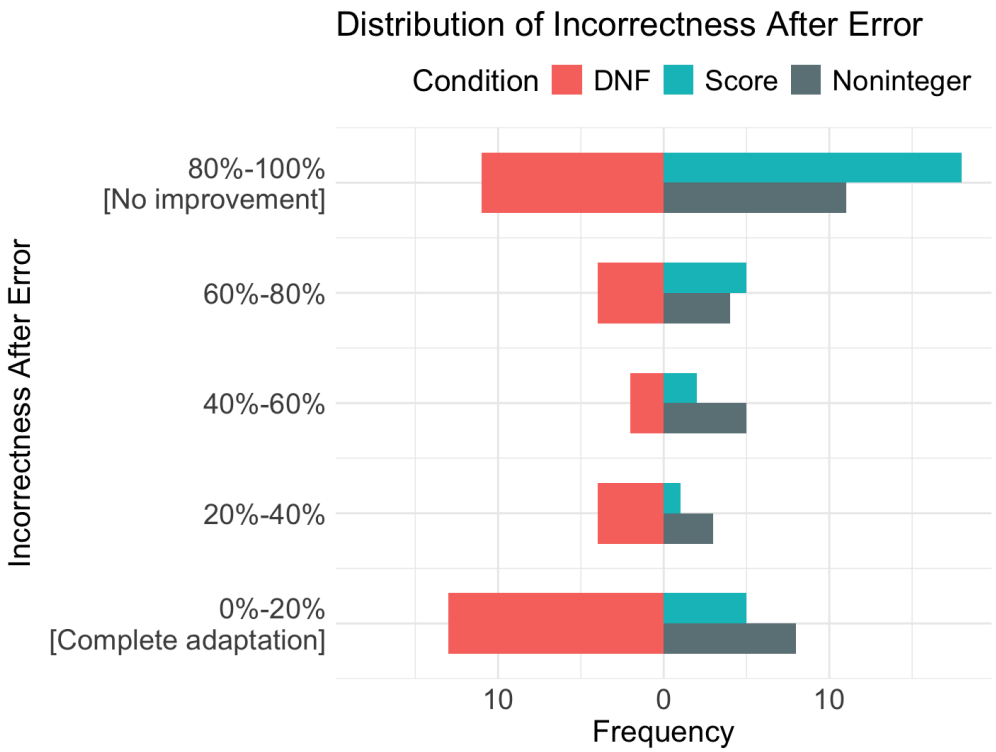


Figure 7: Histogram of participant incorrectness after observing enough model errors to start making informed overrides. The majority of the participants in the `DNF_Complex` showed complete adaptation, and maximally improved their correctness. On the other hand, participants in the `Noninteger_Complex` condition showed learning patterns, with the least learning curve observed for participants in the `DNF_Complex` condition, who predominantly showed no improvement in correctness whatsoever.

E.1.3 Discussion

Our results revealed significant differences in Reliance and Overreliance for participants what saw a model represented as a DNF formula versus participants who saw it as a Score Function. Most strikingly, after seeing model errors, participants showed markedly different adaptation strategies. While participants in the `Score_Complex` condition consistently maintained their initial acceptance strategy even after seeing model errors, participants in the `DNF_Complex` condition split more visibly: the majority achieved perfect adaptation, while some showed no improvement. This distribution suggests two possible mechanisms.

First, the DNF representation might better expose prediction patterns and their failures, helping users track when specific patterns lead to errors. This would align with our initial hypothesis about DNF making model logic more accessible. However, a second possibility is that the DNF's perceived complexity (reported by 46% of the participants) encouraged some users to disengage from model logic entirely and focus on learning from individual robot cases. Both mechanisms could explain why most DNF users adapted perfectly after seeing errors while Score Function users maintained their reliance on the model. We examined which mechanism is more likely in Experiment 4.

E.2 Experiment 4: Overreliance can vary within the same model representation

The results from Experiment 3 raised an important question: **was the reduction in Overreliance with DNF representation due to better understanding of the model logic, or did the representation's complexity simply discourage participants from engaging with the model?** To investigate this, we designed an experiment comparing Integer and Non-integer Score Functions. This modification allowed us to increase model complexity while maintaining the same basic representation format, providing insight into whether complexity alone could drive changes in reliance.

E.2.1 Setup

We chose the exact set of parameters as in Experiment 3. To obtain non-integer coefficients of f^{shown}, we added random $r_i \in [0, 1]$ to each integer coefficient c_i of the original model (see the resulting model in Fig. 4). The procedure followed the steps from Experiment 3. We refer to the new condition as `Noninteger_Complex`

Hypothesis Contrary to our initial assumption in Experiment 3, this time we assumed that it is complexity that plays the main role role in decreasing Overreliance and Reliance. We changed our belief because making participants dismiss the model's logic due to complexity was a simpler solution knowing that almost half of the participants in Stage 1 reported the model as too complex. Hence, we assumed that Reliance and Overreliance would decrease for participants seeing the complex, Non-integer Score Function compared to participants seeing the Integer Score Function.

H4.1: *Non-integer Score Function representation elicits lower Reliance on the tested interpretable model compared to the Integer Score Function representation*

H4.2: *Non-integer Score Function representation elicits lower Overreliance on the tested interpretable model compared to the Integer Score Function representation*

Participants We recruited an additional of 67 participants (age 19-55, avg. 36, English speaking) and accepted 40 out of them (40% rejection rate), including 10 participants (15%) with 1 prediction incompatible with the anchored model. For excluded participants, again, only 12 failed the comprehension check (18%) whereas 15 (22%) displayed non-anchored beliefs. The experiment lasted an average of 39 minutes, and paid $6 with an average bonus of $3.

E.2.2 Results

We found evidence in favor of our hypotheses. We also found more qualitative data corroborating that complexity of the model makes participants disregard model's logic, and move to analyzing data on a case-by-case basis.

On Decisions The comparison between Integer and Non-Integer Score Functions largely replicated the results from Experiment 3:

- Reliance was 79% ($Mean = 79\%, SD = 13\%$) in the `Score_Complex` condition, but in dropped to 71% ($Mean = 71\%, SD = 14\%$) in the `Noninteger_Complex` condition ($U = 474.5; p < 0.01$)
- Overreliance reached a very high 75% ($Mean = 61\%, SD = 38\%$) in the `Score_Complex` condition, but was reduced to 50% ($Mean = 37\%, SD = 35\%$) in the `Noninteger_Complex` condition ($U = 445.5; p < 0.01$)

We found no other significant differences in any reliance-oriented metric, although there was more variance in the responses.

On Correctness and Earnings While we only found marginally significant differences in Correctness between the conditions ($p = 0.08$), we did observe a significant difference in Earnings Deviation, where participants in the `Score_Complex` condition earned 26 points less than the rational agent ($Mean = -18, SD = 11.6$), while participants in the `Noninteger_Complex` condition earned 16 points less ($Mean = -13, SD = 12.2$), $U = 932; p < 0.01$. This pattern mirrors the findings from Experiment 1.

On Differences with DNF According to our hypothesized cause that governs changes in Reliance – perceived complexity – we found no significant differences between the DNF and Non-integer Score Function representations.

On In-depth Analysis Our post-hoc analysis yielded similar adaptation strategies to those found in Experiment 3. Participants in the `Noninteger_Complex` condition demonstrated significantly better adaptation to error but only by not deteriorating their performance. The improvement on individual robots was comparable (50% drop in incorrectness for Robot 14). However, we observed

a subtle shift in adaptation strategies: while the DNF_Complex condition showed a strong split between full adapters and non-adapters, the Noninteger_Complex condition had more participants showing no adaptation and a larger group exhibiting gradual learning (see Fig. 7). Participants' responses also revealed that they found the model complex, with 40% reporting they thought it was too complex, only 6% less than in the DNF_Complex condition.

E.2.3 Discussion

The results of Experiment 4 aligned with our hypothesis that perceived model complexity may cause a drop in Reliance and Overreliance. The results of the Noninteger_Complex condition were not statistically different from the results of the DNF_Complex condition. By and large, neither were the error-adaptation strategies. However, fewer participants claiming the model is too complex in the Non-integer case caused slight differences in the distribution of adaptation behavior. In total, these results suggest that it is the perceived complexity that promotes disengaging from the model and using single predictions to guide decision-making.