Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance

Gagan Bansal **Besmira Nushi**[†] **Ece Kamar**[†] **Walter S. Lasecki**[‡]
Daniel S. Weld **Eric Horvitz**[†]
University of Washington, Seattle [†]Microsoft Research, Redmond [‡]University of Michigan, Ann Arbor

Abstract

Decisions made by human-AI teams (*e.g.*, AI-advised humans) are increasingly common in high-stakes domains such as healthcare, criminal justice, and finance. Achieving high *team* performance depends on more than just the accuracy of the AI system: Since the human and the AI may have different expertise, the highest team performance is often reached when they both know how and when to complement one another. We focus on a factor that is crucial to supporting such complementary: the human's mental model of the AI capabilities, specifically the AI system's *error boundary* (*i.e.* knowing "When does the AI err?"). Awareness of this lets the human decide when to accept or override the AI's recommendation. We highlight two key properties of an AI's error boundary, *parsimony* and *stochasticity*, and a property of the task, *dimensionality*. We show experimentally how these properties affect humans' mental models of AI capabilities and the resulting team performance. We connect our evaluations to related work and propose goals, beyond accuracy, that merit consideration during model selection and optimization to improve overall human-AI team performance.

1 Introduction

While many AI applications address automation, numerous others aim to team with people to improve joint performance or accomplish tasks that neither the AI nor people can solve alone (Gillies et al. 2016; Kamar 2016; Chakraborti and Kambhampati 2018; Lundberg et al. 2018; Lundgard et al. 2018). In fact, many real-world, high-stakes applications deploy AI inferences to help human experts make better decisions, e.g., with respect to medical diagnoses, recidivism prediction, and credit assessment. Recommendations from AI systems—even if imperfect—can result in human-AI teams that perform better than either the human or the AI system alone (Wang et al. 2016; Jaderberg et al. 2019).

Successfully creating human-AI teams puts additional demands on AI capabilities beyond task-level accuracy (Grosz 1996). Specifically, even though team performance depends on individual human and AI performance, the team performance will not exceed individual performance levels if human and the AI cannot account for each other's strengths and weaknesses. In fact, recent research shows that AI accuracy does not always translate to end-to-end team performance (Yin, Vaughan, and Wallach 2019; Lai and Tan 2018). Despite this, and even for applications where people are involved, most AI techniques continue to optimize solely for accuracy of inferences and ignore team performance.

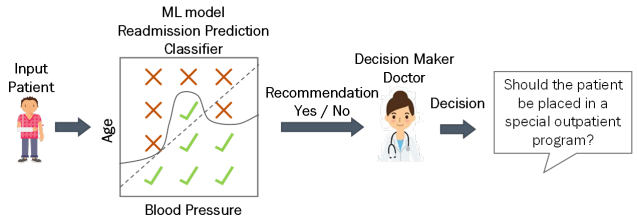


Figure 1: AI-advised human decision making for readmission prediction: The doctor makes final decisions using the classifier's recommendations. Check marks denote cases where the AI system renders a correct prediction, and crosses denote instances where the AI inference is erroneous. The solid line represents the AI error boundary, while the dashed line shows a potential human mental model of the error boundary.

While many factors influence team performance, we study one critical factor in this work: humans' mental model of the AI system that they are working with. In settings where the human is tasked with deciding when and how to make use of the AI system's recommendation, extracting benefits from the collaboration requires the human to build insights (i.e., a mental model) about multiple aspects of the capabilities of AI systems. A fundamental attribute is recognition of whether the AI system will succeed or fail for a particular input or set of inputs. For situations where the human uses the AI's output to make decisions, this mental model of the AI's error boundary—which describes the regions where it is correct versus incorrect—enables the human to predict when the AI will err and decide when to override the automated inference.

We focus here on *AI-advised human decision making*, a simple but widespread form of human-AI team, for example,

in domains like medical diagnosis, candidate screening for hiring, and loan approvals. Figure 1 illustrates an example of AI-advised human decision making for healthcare (Wiens, Guttag, and Horvitz 2016; Caruana et al. 2015). A doctor, using advice from a binary classifier, must decide whether to place a patient in a special (but costly) outpatient program. For a given input, the AI first recommends an action and the human then decides whether to trust or override it to make a final decision. This kind of human-AI collaboration is one formulation of teaming, where there is a binary trust model where the human either trusts or distrusts (and discards) the output of the AI system's influences. We consider the binary trust model instead of situations where the output of the AI system can have varying influence on human decision makers. Team performance in AI-advised human decision making depends on how well the human understands the AI's error boundary. A mismatch between the human's mental model and the true error boundary can lead to suboptimal decisions, such as: (1) the human may trust the AI when it makes an erroneous recommendation, (2) the human may not trust the AI when it makes a correct recommendation. These decision can lower productivity and/or accuracy.

We define properties of an AI's error boundary that affect human's ability to form an accurate mental model, such as *parsimony* and *non-stochasticity*. Intuitively, an error boundary is parsimonious if it is simple to represent. For example, an error boundary that can be described via a minimal number of features or conjunctive expressions on those features is considered to be parsimonious. A non-stochastic error boundary can be modeled with a small set of features that reliably and cleanly distinguishes successes from errors without uncertainty. Another factor that relates to a humans' ability to create a mental model of the error boundary is the task dimensionality, which we characterize by the number of features defining each instance.

We investigate the effect of these properties by conducting controlled user studies using CAJA, which is an open-source and configurable platform that implements an abstract version of AI-advised human decision making (Bansal et al. 2019). Our results demonstrate that parsimony and non-stochasticity of error boundaries improve people's ability to create a mental model. Moreover, the experiments characterize traits of how people create and update mental models over time, highlighting the need for potential guidance in this process. Given the importance of mental models for the ultimate goal of *team* performance, this work advocates for increased attention to properties necessary for effective human-centered AI. We make the following contributions:

1. We highlight an under-explored but significant research challenge at the intersection of AI and human computation research—the role of humans' mental models in team performance in AI-advised human decision making.
2. We identify two attributes of AI systems, parsimony and non-stochasticity of error boundaries, that may help humans learn better mental models of AI competence and therefore improve team performance.
3. Using an open-source, game-centric platform, we show that humans' mental models of AI competence are a critical component of achieving high team performance, provide insights into how humans build mental models in different settings, and demonstrate the desirability of parsimonious and non-stochastic error boundaries.
4. We integrate these results with those of previous work to create a new set of guidelines to help developers maximize the team performance of human-centered AI systems that provide advice to people.

In Section 2 we formally define various concepts: AI-advised human decision making, error boundaries of AI, and mental models of error boundaries. In Section 3, we formulate desirable properties of error boundaries. In Section 4 we study their effect on mental models. We conclude with a discussion of recommendations for developing more human-centered ML.

2 Background

AI-advised human decision making

Following Bansal et al. (2019), we focus on a simple form of human-AI teamwork that is common in many real-world settings, such as a 30-day readmission classifier supporting a doctor (Bayati et al. 2014) or a recidivism predictor supporting judges in courts (Angwin et al. 2016). We refer to situations where an AI system provides a *recommendation* but the human makes the final *decision* as *AI-advised human decision making* (Figure 1). The team solves a sequence of tasks, repeating the following cycle for each time, t.

S1: The environment provides an input, x^t.

S2: The AI (possibly mistaken) suggests an action, $h(x^t)$.

S3: Based on this input, the human makes a decision, u^t.

S4: The environment returns a reward, r^t, which is a function of the user's action, the (hidden) best action, and other costs of the human's decision (e.g., time taken).

The reward feedback in S4 lets the human learn when to trust the AI's recommendation. The cumulative reward R over T cycles is the team's performance. Throughout this paper, we will assume that the AI system is a machine learning (ML) classifier that maps an input $x \in \mathrm{X}$ to an action y from the set of actions Y.

Error boundaries of ML models

The error boundary of model h is a function f that describes for each input x whether model output $h(x)$ is the correct action for that input: $f : (x, h(x)) \rightarrow \{T, F\}$. In other words, the *error boundary* defines the instances for which the model is correct. Note that this is not to be confused with the model's decision boundary, which outputs model predictions. The success of teamwork hinges on the human's recognizing whether to trust the AI model, making error boundaries a critical component of AI-advised human decision making. In fact, appropriate trust in automation is a topic that has received early attention (Lee and See 2004) as determinant factor for designing systems that require people to manage and intervene during imperfect automation.

Human mental models of error boundaries

People create mental models for any system they interact with (Norman 1988), including AI agents (Kulesza et al. 2012). In AI-advised human decision making, a simple definition for such a model would be $m : x' \rightarrow \{T, F\}$, indicating which inputs the human trusts the AI to solve correctly. Here, x' indicates the features that are available to the human. A more complex model might compute a probability and include additional arguments, such as the AI's output and its confidence. Further, there may exist a *representation mismatch*– the human may create a mental model in terms of features that are not identical to the ones used by the ML model. In fact, in real-world deployments, different team members may have access to different features. For example, a doctor may know information about a patient that is missing from electronic health records (*e.g.*, patient's compliance with taking medications), while an AI system may have access to the most recent results and trends in physiological state. However, mental models can be challenging to develop. Even when working within the same feature space, they may not be perfect because users develop them through a limited number of interactions, and humans have memory and computation limitations. To illustrate, the solid line in Figure 1 represents the AI error boundary, while the dashed line shows a possible human mental model of the error boundary.

3 Characterizing AI Error Boundaries

We now define properties that may influence peoples' ability to create a mental model of an AI's error boundary. The first two are the properties of the error boundary itself, while the third is a property of the task.

Parsimony

The parsimony of an error boundary f is inversely related to it representational complexity. For example, in Figure 1 parsimony corresponds to the geometric complexity of the error boundary (solid line). For AI error boundaries formulated in mathematical logic using disjunctive normal form, complexity depends on the number of conjuncts and literals in f. For example, a hypothetical model may yield incorrect recommendations for older patients with high blood pressure or younger patients with low blood pressure. In this case, the error boundary f would be expressed as $\{(age = \texttt{old} \wedge \textit{bloodPressure} = \texttt{high}) \vee (age = \texttt{young} \wedge \textit{bloodPressure} = \texttt{low})\}$, which has two conjunctions with two literals each. This error boundary is more complex and less parsimonious than one that instead uses only one conjunction and two literals.

In reality, an error boundary f may belong to any arbitrary function class. In this work, we choose to express f as a disjunction of conjunctions, where literals are pairs of features and values, so that in our controlled experiments we can vary the complexity of the error boundary and measure how it affects the accuracy of humans modeling the true error boundary. Any other choice of f would make it harder to do such a comparison and would make additional assumptions about the human representation.

	Marvin Correct	Marvin Wrong
Accept	\$0.04	-\$0.16
Compute	0	0

Table 1: Payoff matrix for the studies. As in high-stakes decisions, workers get 4 cents if they accept Marvin when it is correct, and lose 16 cents if they accept Marvin when wrong.

Stochasticity

An error boundary f is non-stochastic if it separates all mistakes from correct predictions. For example, suppose that for the application in Figure 1, the error boundary $f_1 : \{age = \texttt{young} \wedge \textit{blood pressure} = \texttt{low}\}$ is non-stochastic; this means that the readmission classifier always errs for young patients with low blood pressure and is always correct for other patients. In contrast, consider another boundary, f_2, that separates only 90% of the inputs that satisfy f_1. That is, the model will now be correct for 10% of the young patients with low blood pressure, making f_2 a more stochastic error boundary than f_1.

In practice, an error boundary of a given model might be stochastic for three different reasons: generalization, representation mismatch between the AI and human, and inherent stochasticity in the outcome being predicted. Generalization may avoid overfitting by sacrificing instances close to the decision boundary for the sake of using a less complex, and hence more parsimonious, model (*e.g.* a polynomial of a lower degree). However, this may lead to a more stochastic error boundaries. Representation mismatch, for example, may result in a case where many instances that differ for the model appear equal to the human, who cannot understand why the model occasionally fails or succeeds. Finally, the learning model itself might also not be able to completely model the real-world phenomenon due to missing features or imperfect understanding of feature interactions.

In addition to the properties of the error boundary, the dimensionality of the task itself may affect the human's discoverability of the error boundary.

Task dimensionality

We quantify task dimensionality using the number of features defining each instance. With larger numbers of defining features, the search space of all possible error boundaries increases, which may affect how humans create mental models about error boundaries. In practice, using a larger number of features may improve AI accuracy but adversely affect mental models and thus team performance.

4 Experiments

Setup

We now present user studies we performed to build insights about the factors that may affect peoples' abilities to create a mental model of the AI. The studies were conducted using CAJA, an open-source, game-like platform that mimics AI-advised human decision making (Bansal et al. 2019). CAJA is set up in an assembly line scenario, where the task

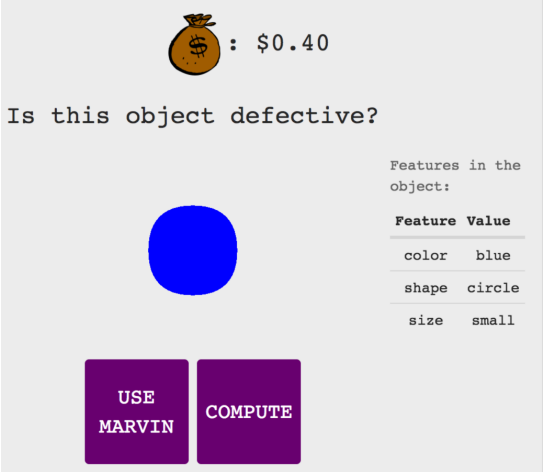


Figure 2: For each object, a subject can either choose to use Marvin's recommendation or perform the task independently.

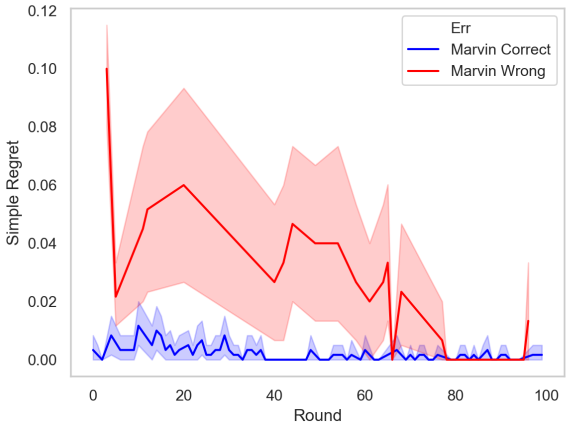


Figure 3: With more rounds of interaction, users perform closer to the optimal policy. Blue indicates the rounds when the AI system (Marvin) is correct and red indicates rounds when the AI makes an error. As mistakes are more costly (Table 1), in the beginning and when Marvin makes a mistake the difference between the optimal reward and reward earned by average worker is higher because the users have an incorrect mental model and fail to override the AI.

of human subjects is to decide whether or not the objects going over the pipeline are defective (Figure 2). To decide on these labels, for each instance, subjects take a recommendation from an AI system called Marvin and, based on their mental model of Marvin, decide whether they should accept the AI recommendation or override it by clicking the compute button. After submitting a choice, the human receives feedback and monetary reward based on her final decision. Table 1 shows the payoff scheme used across these experiments, which aims to simulate high-stake decision making (*i.e.*, the penalty for an incorrect action is much higher than the reward for a correct one). In this game, subjects are not supposed to learn how to solve the task. In fact, the decision boundary is generated randomly and the only way for a participant to earn a high score is by learning the error boundary and relying on the Compute button to acquire the right prediction if Marvin is not to be trusted. This abstracts away human expertise in problem solving so the focus remains on the ability to learn the error boundary.

The CAJA platform enables control over many parameters relevant to AI-advised human decision making: task dimensionality, AI performance, length of interaction, parsimony and stochasticity of the error boundary, cost of mistakes, etc. In the human studies, we systematically vary these parameters and measure team performance to study the factors that affect humans' ability in developing mental models. Note that game parameters also distinguish between features that the machine reasons about and those that the human has access to. More specifically, the platform currently allows configurations where machine-visible features are a superset of human-visible ones, which is also the type of configuration we use in the next experiments.

All studies were conducted on Amazon Mechanical Turk. For every condition we hired 25 workers and on average workers were paid an hourly wage of $20. To remove spam, we removed observations from workers whose performance was in the bottom quartile. We explain results in a question and answer format.

Results

Q1: *Do people create mental models of the error boundary? How do mental models evolve with interaction?*

We visualize action logs collected by CAJA to understand the evolution of mental models with more rounds of interaction. Figure 3 shows the average simple regret (*i.e.*, difference between the optimal and observed reward) of workers' actions over time (*i.e.*, rounds). The optimal policy is an oracle with access to Marvin's true error boundary that can thus always correctly choose when to trust Marvin. We observe that the simple regret decreases with more interactions, indicating that, on average, workers gradually learn the correct mental model and perform closer to the optimal policy.

Figure 4 shows the evolution of the mental model for one particular worker when Marvin's true error boundary is non-stochastic, uses one conjunction and two literals, and task dimensionality is three, *i.e.*, three features describe the problem space visible to the human. In the beginning, the worker makes more mistakes (more red crosses) because the mental model thus far is only partially correct. Eventually, the worker learns the correct model and successfully compensate for the AI (more red checks). Note that a mental model may be partially correct for two reasons: it is either over-generalized or over-fitted. An *over-generalized mental model* includes more cases in the error boundary than it should, for example, when the true error boundary is small circles, and the over-generalized mental model includes all small shapes. In contrast, an *over-fitted mental model* misses cases that the true error boundary contains. For example, point (c) of Figure 4 shows where the worker over-fit to

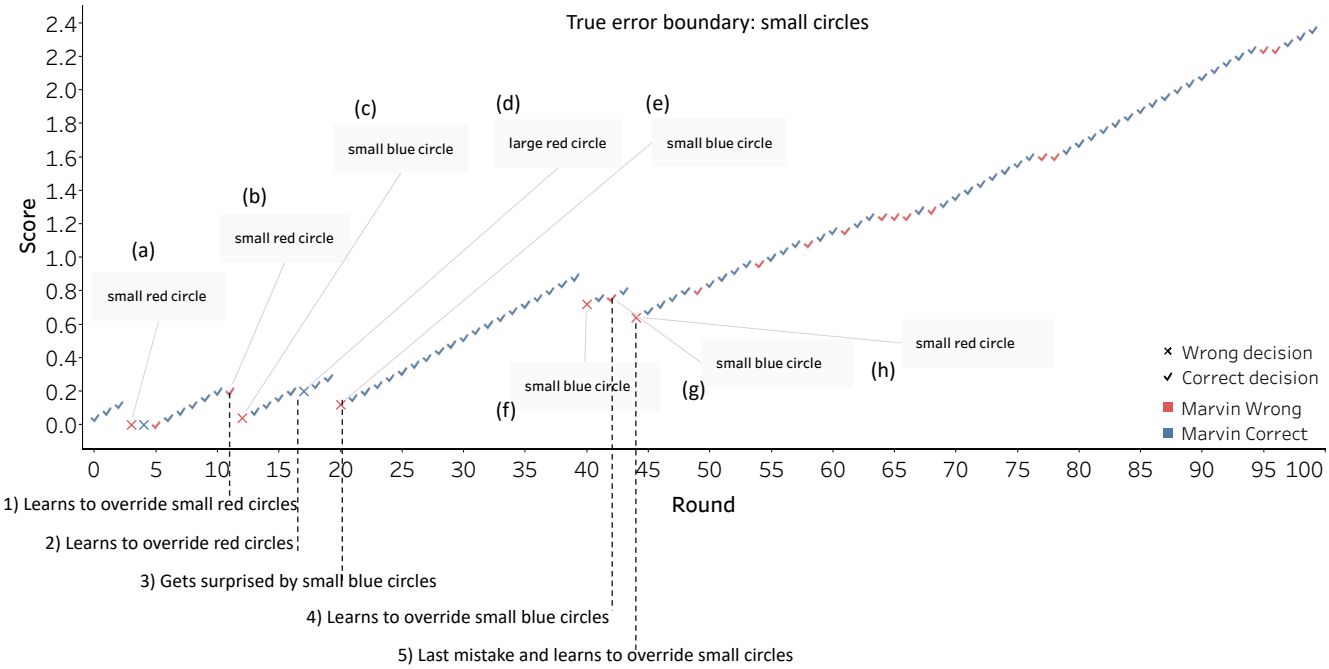


Figure 4: A visualization of a worker's behavior that shows how their mental model is refined with continuing interaction. Here, the score indicates the cumulative reward, and Marvin makes mistakes whenever the object is a small circle. Red markers indicate such rounds. Cross markers indicate if the worker's final decision was wrong. Hence, red crosses indicate a false accept (e.g., (a), (c), and (e)) and result in large negative reward. On the other hand, blue checks indicate a successful accept and result in a positive reward. Blue crosses indicate a false override and red checks indicate a true override. The figure contains a lot more crosses before round 45 than after. This indicates that the worker makes most of the wrong decisions in the first half of the interaction but eventually learns to act optimally. Annotations 1-5 describe the different stages of the worker's mental model. For example, by (1) the worker learns to override small red circles presumably because she learned from a previous wrong decision (a). However, since this mental model is only partially correct, in subsequent rounds (c, e, f) the worker makes wrong decisions for small blue circles. This causes surprise and confusion at first, but she eventually learns to override small blue circles by (4). But then in subsequent rounds she makes a wrong decision for a small red circle (5). After this mistake, the worker finally ties together lessons from all of her previous mistakes, figures out that small circles are problematic irrespective of the color, and acts optimally thereafter.

small red circles when in fact errors occur for small circles. Clearly, a combination of both impartialities can also occur if the human tries to generalize too early on the incorrect feature literal.

Q2: *Do more parsimonious error boundaries facilitate mental model creation?*

To answer this question, we compare team performance of many conditions that vary parsimony by changing the number of conjunctions and literals. We additionally vary the number of features to study the effect of parsimony for different task dimensionality. Figure 5 shows the overall team performance (cumulative score) for two boundaries of different complexity: a single conjunction with two literals (e.g., red and square), and two conjunctions with two literals each (e.g., red and square or small and circle). Different features may have different salience; therefore, for the same formula, we randomly assign different workers isomorphic error boundaries. For example, the error boundary (red and square) is isomorphic with the error boundary described by (blue and circle). Since error boundary complexity increases with the number of conjunctions, we observe that a more parsimonious error boundary (*i.e.*, a single conjunction) results in a higher team performance. Thus, our results demonstrate the value for learning ML models with parsimonious error boundaries, for example, by minimizing the number of conjunctions. In Figure 6, we observe that team performance generally decreases as the number of human-visible features increases, which is consistent with previous findings on human reasoning about features (Poursabzi-Sangdeh et al. 2018).

Q3: *Do less stochastic error boundaries lead to better mental models?*

In the previous experiments, Marvin's error boundary was non-stochastic (*i.e.*, Marvin made a mistake if and only if the object satisfied the formula). In practice, error boundaries may be fuzzier and not as clean. To understand the effect of stochasticity, we vary two parameters: $P(err|\neg f)$, the conditional probability of error if the object does not satisfy

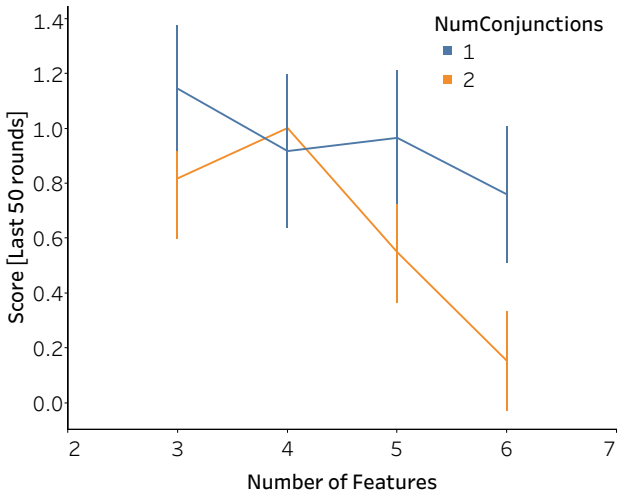


Figure 5: Team performance decreases as the number of conjuncts in the error boundary is increased. Number of literals were fixed to 2.

the formula, and $P(err|f)$, the conditional probability of error if the object satisfies the formula. In the non-stochastic experiments, we use a $P((err|\neg f) = 0, P(err|f) = 1))$ configuration. The other conditions that we experiment with are $(0, 0.85)$, $(0, 0.7)$, and $(0.15, 0.85)$. Of these, only the last condition is two-sided, meaning that errors can occur on both sides of the boundary although with less probability when the formula is not satisfied. All other conditions are one-sided.

Figure 7 shows that for the one-sided error case, the percentage of workers who make the correct decision (vertical axis) increases over time. In contrast, for two-sided error boundaries, workers find it challenging to learn the true error boundary. In addition, we observe that even for one-sided error case, increased stochasticity makes it difficult for participants to trust Marvin and learn a correct mental model. For example, the $(0, 0.7)$ condition has clearly more rounds where Marvin was correct (indicated by circles) and the percentage of people who trusted Marvin is less than 50%.

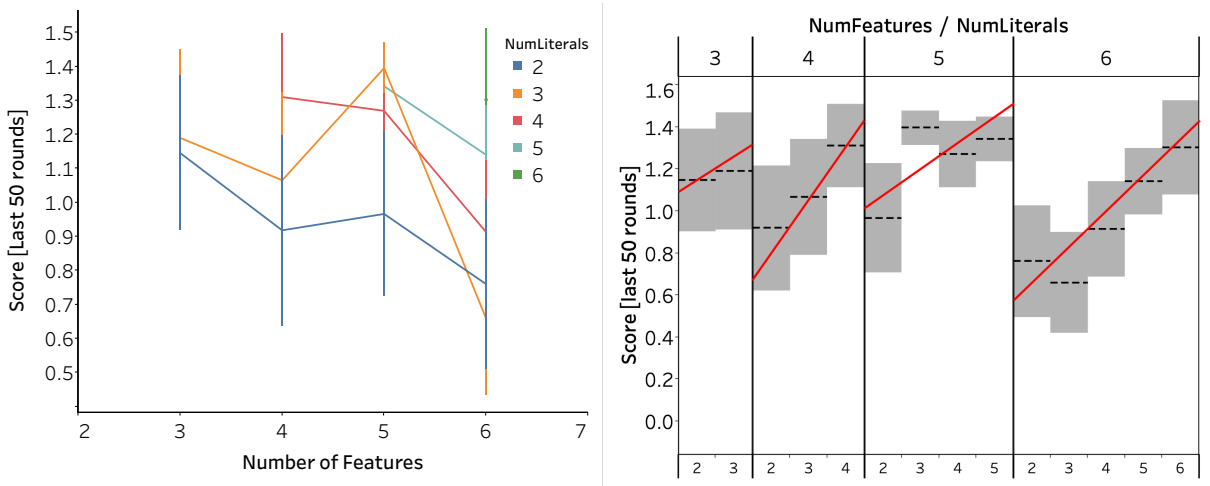
5 Related Work

Mental models for collaboration. Early work explored the importance of mental models for achieving high performance in group work (Grosz and Kraus 1999; Mohammed, Ferzandi, and Hamilton 2010), human-system collaboration (Rouse, Cannon-Bowers, and Salas 1992), and interface design (Carroll and Olson 1988). More recently, the impact of mental models has been revisited for better understanding human-in-the-loop systems (Chakraborti and Kambhampati 2018) and for grounding human-AI collaboration within traditional HCI work (Kaur, Williams, and Lasecki 2019). Our work builds upon these foundations and studies the problem for AI-advised human decision making. While there exist many forms of mental modeling (*i.e.*, How does the system work?) and they are relevant for collaboration, this work focuses particularly on mental models about system performance (*i.e.*, When does the system err?), which are learned upon context and past experiences.

Backward compatibility. The closest relevant work to this study that also operates on mental models about error boundaries is presented in (Bansal et al. 2019) and focuses on the usefulness of such models during AI updates highlighting the importance of remaining backward compatible while deploying a new model. Backward compatibility is measured through comparing the errors of the previous and the updated version of the model and quantifying the percentage of all input instances that were correct in the previous version that remain correct in the updated one. The work showed that error boundaries that are not backward compatible with previous versions of the model breaks mental models human have created in the process of collaboration, and showed that updates to a more accurate model that is not backward compatible can hurt team performance.

In traditional software design, backward compatibility is a well-studied software property (Bosch 2009; Spring 2005), used to denote software that remains compatible with a larger legacy ecosystem even after an update. In the field of AI/ML, a related notion to backward compatibility is catastrophic forgetting (Kirkpatrick et al. 2017; Goodfellow et al. 2013; McCloskey and Cohen 1989), which is an anomalous behavior of neural network models that occurs when they are sequentially trained on more instances and forget to solve earlier instances over time. While forgetting in sequential learning is an important problem, backward compatibility is applicable to a larger set of update scenarios that do not necessarily require more data (*e.g.* different architecture or the same architecture but with different parameters).

Interpretability for decision-making. As learning models are being deployed to assist humans in taking high-stake decisions, the explainability of machine predictions is crucial for facilitating human understanding. Ongoing and prior research has contributed to improving the interpretability of such predictions either by building more interpretable models (Caruana et al. 2015; Rudin 2018; Lage et al. 2018) or by imitating complex models via simpler but more explainable ones (Lakkaraju, Bach, and Leskovec 2016; Tan et al. 2018). However, while explanations help with understanding, it is not yet clear under which conditions they improve collaboration and human productivity (Doshi-Velez and Kim 2017; Poursabzi-Sangdeh et al. 2018; Feng and Boyd-Graber 2019). For example, some explanations may describe how the system works but they do clearly disclose when it will fail and needs human intervention. In other cases, inspecting an explanation might take just as much time as solving the task from scratch (*i.e.*, high cognitive load). Both challenges motivate the need for predictable and easy-to-learn error boundaries, properties of which we study in our experimental evaluation. A promising direction is generating explanations of error boundaries themselves as a tool for users to quickly learn and remember various failure conditions. Recent work (Nushi, Kamar, and Horvitz 2018), uses decision trees to predict and visualize error boundaries for the purpose of debugging ML models



(a) (b)

Figure 6: a) Team performance decreases as the task dimensionality increases (i.e., number of features). b) Re-visualization of a) that shows that, for a given number of features, team performance increases with the number of literals in the error boundary, because the errors become more specific. The solid red lines show this trend. Number of conjuncts was fixed to 1.

but more work is needed on deploying and evaluating such tools for decision-making.

Modeling and communicating uncertainty in ML. Confidence calibration has been a topic of extensive research, especially for embracing and communicating uncertainty in models that are inherently non-probabilistic. Foundational work in this space has proposed techniques for calibrating output scores for support vector machines (Platt and others 1999), decision trees and Naïve Bayes models (Zadrozny and Elkan 2001), and deep neural networks (Gal and Ghahramani 2016; Gal 2016; Guo et al. 2017). Later work has proposed data collection algorithms for addressing overconfident predictions (Lakkaraju et al. 2017; Bansal and Weld 2018). While confidence estimation and reporting is informative for decision-making, research in human-centered machine learning (Gillies et al. 2016) and HCI shows that people have difficulties with correctly interpreting probabilistic statements (Handmer and Proudley 2007) and even system accuracy itself (Yin, Vaughan, and Wallach 2019). Research in the intersection of AI and HCI has found that interaction improves when setting expectations right about what the system can do and how well it performs (Kocielnik, Amershi, and Bennett 2019; Amershi et al. 2019). This paper takes a step forward by proposing properties of ML models that can assist with setting the right expectations and evaluating them through controlled user studies. Moreover, we envision this line of work on making error boundaries predictable as complementary but also valuable for designing better confidence models as the defined properties. For example, parsimonious error boundaries are easier to generalize also from a statistical learning point of view, which would help with calibration.

6 Recommendations for Human-Centered AI

When developing ML models current practices solely target AI accuracy, even in contexts where models support human decision making. Our experiments reveal that error boundaries and task complexity can influence the success of teamwork. The user studies presented suggest the following considerations when developing ML models to be used in AI-advised human decision making:

1. Build AI systems with parsimonious error boundaries.
2. Minimize the stochasticity of system errors.
3. Reduce task dimensionality when possible either by eliminating features that are irrelevant for both machine and human reasoning or most importantly by analyzing the trade-off between the marginal gain of machine performance per added feature and the marginal loss of the accuracy of human mental models per added feature.
4. Based on results from Bansal et al. (2019), during model updates, deploy models whose error boundaries are *b*ackward compatible, *i.e.* by regularizing in order to minimize the introduction of new errors on instances where the user has learned to trust the system.

Given the importance of these properties on overall team performance, and potentially of other properties to be discovered in future work, it is essential to make such properties a part of considerations during model selection. For example, if a practitioner is presented with two different models, h_1 and h_2, of similar accuracy (e.g., such a situation could arise as a result of a grid search for hyper-parameter selection), and the error boundary f_1 is more stochastic than f_2, clearly h_2 would be the better choice. In a more complex situation, where h_2's accuracy is slightly inferior to h_1's, the practitioner must carefully estimate the potential loss in team accuracy attributed to human mistakes (*i.e.*, trusting the

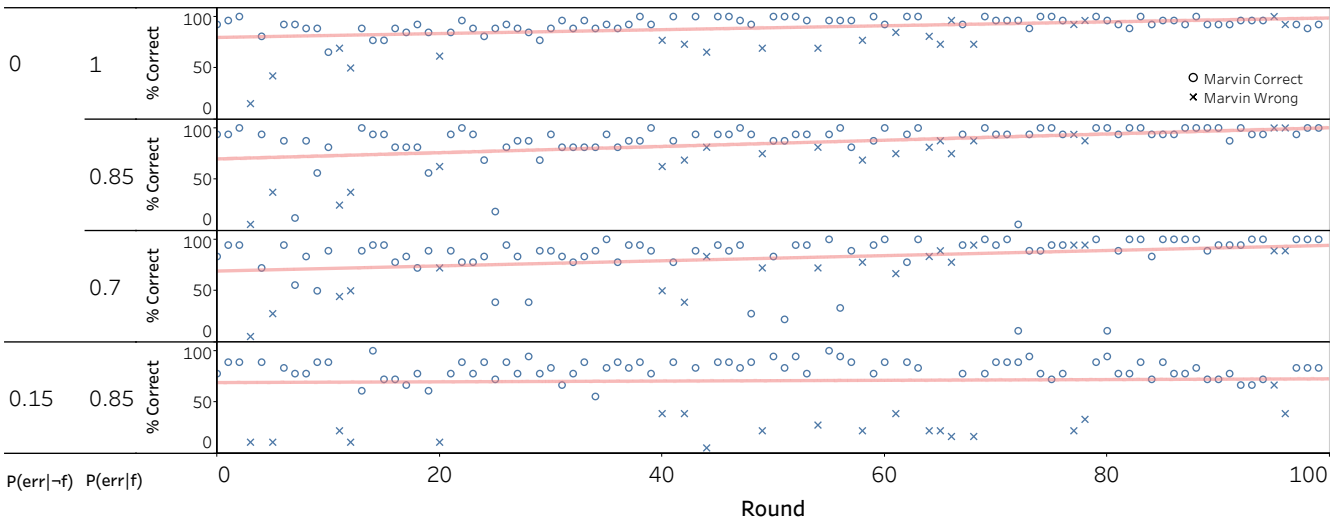


Figure 7: For one-sided error boundaries (the top three rows), the percentage of workers who choose the optimal action improves with time and reaches 100% – the positive slope of the best fit line shows this increasing trend. For the two-sided stochastic boundary (bottom row), the improvement is minimal and stays close to 50% – the slope of the best fit line is close to 0.

model when it is incorrect) due to stochasticity and compare this loss to the difference in accuracy between the two candidate models. Often, a negligible compromise in ML accuracy, can lead to for a higher gain in accuracy of overall teamwork. The same analysis could be employed to appraise the optimal tradeoffs when one human-aware property may be at odds with another (*e.g.*, making an error boundary more parsimonious might also make it more stochastic).

Model selection decisions also depend on the type of tools made available to users for learning and remembering error boundaries. For example, if users can access a scratchpad that records and summarizes the observed error boundary in real time, then they might be able to afford a slightly more complex error boundary.

Human-aware model selection should also be supported by making the presented properties part of the optimization problem is formulation while training either by including human-aware considerations in loss functions or by posing additional optimization constraints. The former technique has been used to combine backward compatibility in the loss function (Bansal et al. 2019) and to combine tree-based regularization to learn a more interpretable model (Wu et al. 2018); the latter has found application in domains like fair classification and healtcare (Dwork et al. 2012; Zafar et al. 2017; Ustun and Rudin 2017). More effort is needed to algorithmically ensure error boundary parsimony and non-stochasticity and combine such efforts for generating actionable confidence scores. This would reshape learning techniques to optimize for both the human in the loop or any other part of the ecosystem that requires reliable trust contracts to cooperate with the AI.

Finally, as human-AI collaboration becomes more pervasive, we foresee further opportunities to study human-AI team behavior in the open world, and for richer and more general forms of human-AI teams, for example, cases where AI recommendation directly updates human's belief in the final decision in contrast to our simplified notion of accept or override. Other opportunities include making interaction more natural by building computational models about what users have learned and by simplifying mental model creation using explanatory tools.

7 Conclusion

We studied the role of human mental models on the human-AI team performance for AI-advised human decision making for situations where people either rely upon or reject AI inferences. Our results revealed important properties that describe the error boundaries of inferences that can influence how well people can collaborate with an AI system and how efficiently they can override the AI when it fails. We find that systems with exactly the same accuracy can lead to different team performance depending upon the parsimony, non-stochasticity, and dimensionality of error boundaries. Future research opportunities include developing methods for integrating these considerations into algorithmic optimization techniques. While AI accuracy has been traditionally considered a convenient proxy for predicting human-AI team performance, our findings motivate investing effort to understand how to develop AI systems to support teamwork, in particular, in making properties of error boundaries more understandable and learnable when selecting an AI model for deployment.

Acknowledgements

We thank M. Ribeiro, M. Czerwinski, Q. Chen, and anonymous reviewers for helpful feedback. This work was supported by Microsoft Research, ONR grant N00014-18-1-2193, the Future of Life Foundation, the WRF/Cable Professorship, and a DARPA Young Faculty Award.

References

Amershi, S.; Weld, D. S.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S. T.; Bennett, P. N.; Quinn, K. I.; Teevan, J.; Kikin-Gil, R.; and Horvitz, E. 2019. Guidelines for human-ai interaction. In *CHI*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software across the country to predict future criminals and its biased against blacks. *ProPublica*.

Bansal, G., and Weld, D. S. 2018. A coverage-based utility model for identifying unknown unknowns. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D.; Lasecki, W. S.; and Horvitz, E. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the 33rd AAAI conference on Artificial Intelligence. AAAI*.

Bayati, M.; Braverman, M.; Gillam, M.; Mack, K. M.; Ruiz, G.; Smith, M. S.; and Horvitz, E. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one* 9(10).

Bosch, J. 2009. From software product lines to software ecosystems. In *SPLC*, 111–119.

Carroll, J. M., and Olson, J. R. 1988. Mental models in human-computer interaction. In *Handbook of human-computer interaction*. Elsevier. 45–65.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*, 1721–1730. ACM.

Chakraborti, T., and Kambhampati, S. 2018. Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-ai collaboration. *arXiv preprint arXiv:1801.09854*.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

Feng, S., and Boyd-Graber, J. 2019. What can ai do for me?: evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 229–239. ACM.

Gal, Y., and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059.

Gal, Y. 2016. *Uncertainty in deep learning*. Ph.D. Dissertation, PhD thesis, University of Cambridge.

Gillies, M.; Fiebrink, R.; Tanaka, A.; Garcia, J.; Bevilacqua, F.; Heloir, A.; Nunnari, F.; Mackay, W.; Amershi, S.; Lee, B.; et al. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 3558–3565. ACM.

Goodfellow, I. J.; Mirza, M.; Xiao, D.; Courville, A.; and Bengio, Y. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Grosz, B. J., and Kraus, S. 1999. The evolution of shared-plans. In *Foundations of rational agency*. Springer.

Grosz, B. J. 1996. Collaborative systems (AAAI-94 presidential address). *AI magazine* 17(2):67.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1321–1330. JMLR. org.

Handmer, J., and Proudley, B. 2007. Communicating uncertainty via probabilities: The case of weather forecasts. *Environmental Hazards* 7(2):79–87.

Jaderberg, M.; Czarnecki, W. M.; Dunning, I.; Marris, L.; Lever, G.; Castañeda, A. G.; Beattie, C.; Rabinowitz, N. C.; Morcos, A. S.; Ruderman, A.; Sonnerat, N.; Green, T.; Deason, L.; Leibo, J. Z.; Silver, D.; Hassabis, D.; Kavukcuoglu, K.; and Graepel, T. 2019. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science* 364:859–865.

Kamar, E. 2016. Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *IJCAI*.

Kaur, H.; Williams, A.; and Lasecki, W. S. 2019. Building shared mental models between humans and ai for effective collaboration.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 201611835.

Kocielnik, R.; Amershi, S.; and Bennett, P. N. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems.

Kulesza, T.; Stumpf, S.; Burnett, M.; and Kwan, I. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *CHI*, 1–10. ACM.

Lage, I.; Ross, A. S.; Kim, B.; Gershman, S. J.; and Doshi-Velez, F. 2018. Human-in-the-loop interpretability prior. *arXiv preprint arXiv:1805.11571*.

Lai, V., and Tan, C. 2018. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *ACM Conference on Fairness, Accountability, and Transparency*.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, 1675–1684. ACM.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Horvitz, E. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Lee, J. D., and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46(1):50–80.

Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.-W.; Newman, S.-F.; Kim, J.; et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2(10):749.

Lundgard, A.; Yang, Y.; Foster, M. L.; and Lasecki, W. S. 2018. Bolt: Instantaneous crowdsourcing via just-in-time training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 467. ACM.

McCloskey, M., and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24. Elsevier. 109–165.

Mohammed, S.; Ferzandi, L.; and Hamilton, K. 2010. Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management* 36(4):876–910.

Norman, D. 1988. *The psychology of everyday things*. Basic Books.

Nushi, B.; Kamar, E.; and Horvitz, E. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. *HCOMP*.

Platt, J., et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10(3):61–74.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Rouse, W. B.; Cannon-Bowers, J. A.; and Salas, E. 1992. The role of mental models in team performance in complex systems. *IEEE Transactions on SMC* 22(6):1296–1308.

Rudin, C. 2018. Please stop explaining black box models for high stakes decisions. *CoRR* abs/1811.10154.

Spring, M. J. 2005. Techniques for maintaining compatibility of a software core module and an interacting module. US Patent 6,971,093.

Tan, S.; Caruana, R.; Hooker, G.; Koch, P.; and Gordo, A. 2018. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640*.

Ustun, B., and Rudin, C. 2017. Optimized risk scores. In *KDD*, 1125–1134. ACM.

Wang, D.; Khosla, A.; Gargeya, R.; Irshad, H.; and Beck, A. H. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Wiens, J.; Guttag, J.; and Horvitz, E. 2016. Patient risk stratification with time-varying parameters: a multitask learning approach. *JMLR* 17(1):2797–2819.

Wu, M.; Hughes, M. C.; Parbhoo, S.; Zazzi, M.; Roth, V.; and Doshi-Velez, F. 2018. Beyond sparsity: Tree regularization of deep models for interpretability. In *AAAI*.

Yin, M.; Vaughan, J. W.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models.

Zadrozny, B., and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, 609–616. Citeseer.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970.