Understanding non-adoption of "optimal" algorithmic tips for problem-solving: how people view and use tips and the barriers they encounter

DAVID LEE*, UC Santa Cruz
WICHINPONG PARK SINCHAISRI*, UC Berkeley

Designing effective human-AI systems requires developing a deeper understanding of how humans use or choose not to use AI systems. In this paper, we ask: how do humans view and use algorithmic tips in multi-step problem solving contexts? What barriers keep them from trusting or adopting tips to solve problems? We conducted a qualitative analysis of participant responses to "optimal" algorithmic tips provided to them for managing a virtual kitchen after disruptions necessitated a change of strategy. We found that tips were viewed or used by participants in four different ways: as rules, directional principles, options to try, and highlights. Even in cases when workers rejected tips outright, tips could still be useful through creating focal points in the solution space for worker sense-making. We also found that in problem solving contexts, the challenge of operationalizing tips can lead to diverse problems. Some related to participants not trusting the tip (due to it being counterintuitive or resulting in bad outcomes), but others related to tip usability (lacking clarity, being difficult to implement, or being difficult to track whether they were implementing) and to broader environment factors (misaligned incentives) that also challenge a simple definition of what it means for a tip to be optimal.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: human-ai collaboration, algorithmic tips, barriers to adoption of tips

ACM Reference Format:
David Lee and Wichinpong Park Sinchaisri. 2018. Understanding non-adoption of "optimal" algorithmic tips for problem-solving: how people view and use tips and the barriers they encounter. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Human-AI interfaces are increasingly used to aid humans in the real world, including in consequential domains from healthcare [8] to legal decision-making [1]. However, these interfaces have not been fully adopted for many reasons such as the black-box nature of underlying algorithms resulting in a lack of transparency, accountability, or interpretability. The lack of human understanding of how algorithms work could pose serious problems to the society, from prisoners incorrectly denied parole to polluted air mistakenly identified as safe [13], and lead to aversion to the machine-generated recommendation among humans [4, 7]. For example, Dietvorst, Simmons, and Massey [7] showed that, in a forecasting task, humans prefer to follow the suggestion made by another human forecaster rather than by an algorithm and that their *confidence* in the algorithm declines at a faster rate when a flawed suggestion was made. Castelo, Bos, and Lehmann [4] further demonstrated that this is particularly true for subjective tasks where humans *incorrectly assume*

*Both authors contributed equally to this research.

Manuscript submitted to ACM

that algorithms were only able to perform objective tasks. On the other hand, in tasks in which a ground truth exists, humans can make the opposite mistake of automation bias, where they trust algorithms when they should not [10].

To address these challenges, researchers have sought to understand more deeply when humans are or are not able to detect and override errors [6]. For example, displaying confidence scores can help calibrate trust in AI models, but showing how much different features contribute to a prediction does not [5, 9, 14, 16]. More transparent models can actually make participants less able to detect mistakes due to information overload [12], unless cognitive forcing functions are used to force engagement with AI explanations [3]. Much of this work has focused on prediction problems, in which the primary problem is whether to accept or not accept the AI-recommended decision.

In a recent paper [2], Bastani, Bastani, and Sinchaisri studied algorithmic tips in a problem solving context in which humans need to make a sequence of interdependent decisions with the aid of tips that provide users with guidance (e.g. "server should cook twice") but are not complete solutions that only requires deciding whether one should "adopt" or "not adopt". They introduced an approach to determining "optimal" tips in that they would bridge the largest gap between human strategies and optimal ones. However, they found that despite the algorithmic tip performing much better than alternatives in helping participants improve their score on average, the counterintuitive nature of the tip made it much less likely for people to adopt as compared to a more intuitive human-provided (but non-optimal) tip.

In this paper, we analyze the qualitative responses from this large-scale behavioral experiment to develop a richer understanding of human non-adoption of machine-generated "optimal" tips. We found that tips were viewed or used by participants in four different ways: as rules, directional principles, options to try, and highlights. Even in cases when workers rejected tips outright, tips could still be useful through creating focal points in the solution space for worker sense-making. We also found that in problem solving contexts, the challenge of operationalizing tips can lead to diverse barriers. Some related to participants not trusting the tip (due to it being counterintuitive or resulting in bad outcomes), but others related to tip usability (lacking clarity, being difficult to implement, or being difficult to track whether they were implementing) and to broader environment factors (misaligned incentives) that also challenge a simple definition of what it means for a tip to be optimal. We end by discussing how this richer view of human interactions with tips suggest implications for design, raise questions about what it means for a tip to be "optimal", and point to directions for future research for understanding AI-assisted decision-making in more complex problem solving contexts.

2 RESEARCH SETTING: ALGORITHMIC TIPS FOR MANAGING A VIRTUAL KITCHEN

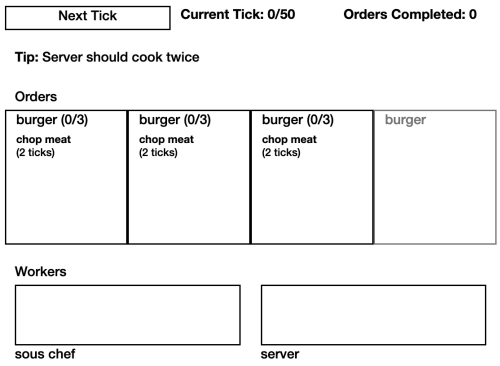
We consider human adoption of tips in a setting where humans receive an advice while performing a sequential decision-making task. In the experimental setting of [2], human participants act as managers for a virtual kitchen in which they must assign various cooking tasks (e.g., chopping, cooking, plating) to virtual kitchen workers (e.g., chef, sous-chef, server) with varying capabilities in a way that minimizes the completion time of all food orders (e.g., four burgers). Participants play the game for two rounds in which they operate a full-capacity kitchen (e.g., with all three virtual workers) and then for four additional rounds where the most capable virtual worker, the chef, is no longer available and the same food orders have to be made by the remaining two virtual workers, the sous-chef and the server. In these four rounds, workers try to reach a known optimal completion time and are provided with different tips (depending on their experimental condition). The key challenges of this game are that (i) human players need to uncover the actual skill level of each virtual worker and make short- and long-term trade-offs when allocating tasks to avoid bottlenecks, and (ii) the disruption caused by the departure of the chef requires the players to adapt to the new environment where a new decision-making strategy is needed.

The setting of the virtual kitchen management game can be considered as a general task allocation and scheduling problem that managers face in practice as it captures the key challenges of managing bottlenecks, learning workers' (initially) unknown skill levels, and adapting to a disruption. In practice, rather than knowing the true optimal value based on a fully-solved optimal policy, managers are often informed of performance benchmarks within the organization and across the industry. For example, according to the Bureau of Transportation Statistics, an American regional airline Endeavor Air has the highest on-time arrival performance of 89.16% compared to the industry's average of 84.63% [11]. Operations managers are constantly seeking to scheduling tasks to minimize delays and meet industry benchmarks.

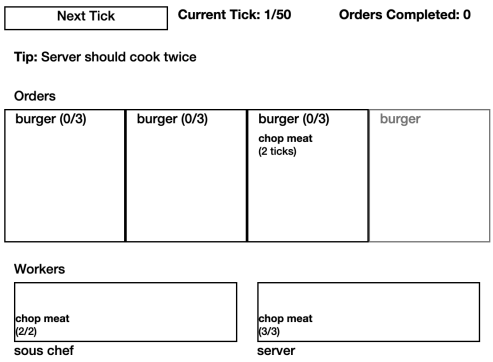
An extensive behavioral study was conducted on Amazon Mechanical Turk in two phases. In the first phase ($N = 172$), participants played the game without receiving a tip, and their sequences of decisions were recorded. This data was then fed into a novel machine learning algorithm that can extract the decision-making strategies employed by the participants and generate an interpretable advice ("tip") by comparing the actions taken by the humans with the optimal policy. Due to the sequential nature of the decision-making task, the optimal policy is a complex sequence of state-action pairs. For interpretability, the algorithm then chooses only a snapshot of the optimal policy, or one state-action pair to provide to the participants. Therefore, even though the tip is based on the optimal policy, it is not guaranteed that the participants would be able to recover the remaining state-action pairs in the optimal policy. In the second phase ($N = 1,011$), a new set of participants were randomly assigned into one of the four conditions: *control* (e.g., not receiving any tip), *algorithm* (e.g., receiving a tip chosen by the novel algorithm), *human* (e.g., receiving a tip that received the most votes from participants in the first phase as "best tip to help future players"), and *baseline* (e.g., receiving a tip from a naive, frequency-based algorithm). Then, they proceeded to play the same virtual kitchen-management game where the tip (if any) was shown during the game in every round. Figure 1 shows the example screenshots of the game interface for a participant randomly assigned into the *algorithm* condition. Performance was measured in terms of the completion time in each round of the game and the fraction of participants who achieved the optimal solution. The tips shown in this phase are:

- *Algorithm:* Server should cook twice (out of four burgers). This tip is part of the optimal policy.
- *Human:* Server should cook once (out of four burgers). This tip is not part of the optimal policy, but it is closer to the optimal policy under the full-capacity scenario (e.g., the server takes the longest time to cooking among all virtual workers, so cooking should be assigned primarily to the chef) than the algorithmic tip.
- *Baseline:* Sous-chef should plate twice (out of four burgers). This tip is also part of the optimal policy.

The experimental results of [2] demonstrated that the algorithmic tip enabled human participants to substantially improve their performance compared to counterparts that were not shown the tip or were shown alternative tips. Specifically, participants in the *Algorithm* condition completed the final round of the game in 37.1 steps, outperforming those in other conditions: 37.9 (*Control*, $t(243) = -4.361$, $p = 0.00000806$), 37.5 (*Human*, $t(246) = -2.52$, $p = 0.00605$), and 38.4 (*Baseline*, $t(246) = -7.348$, $p < 10^{-12}$). Furthermore, 19% of participants in the *Algorithm* condition achieved optimal performance (34 steps) in the final round, compared to less than 1% in all other conditions. Thus, the algorithmic tip can be considered as the "optimal" tip. However, the authors noted that the adoption rate among participants in the *Algorithm* treatment was much lower (24–48%) than the adoption rate among those in the *Human* treatment (83–88%). Although, even with low adoption, the algorithmic tip was found to already be effective at improving performance, increasing adoption of the optimal tip could lead to a greater performance improvement. After the participants played all six rounds of the game, they responded to a survey about their experience, gameplay, and thoughts about the tip. In this paper, we analyze the survey responses to gain a deeper insight into human adoption and nonadoption of tips.



(a) The initial state where users observe available subtasks, median times to completion, and two idle virtual workers. The interface also shows the current time index, time limit, current progress, and potential tip.



(b) The next state after two previously available subtasks were assigned to the virtual workers and the true completion times were realized, revealing different levels of virtual workers' skills

Fig. 1. Example screenshots from the virtual kitchen management game (from [2])

3 METHOD

This paper is a works-in-progress report of our preliminary analysis of worker responses to one of the open-ended questions in the post-study survey: *"What did you think about the tip for these last four rounds and how did you incorporate it in your strategy?"* Thus far, we have only analyzed responses from the workers in the *Algorithm* condition, i.e. those who received the optimal tip *"server cooks twice"*.

One author carried out an inductive coding process, in which we started with open coding anything relevant to characterizing the diverse ways in which workers engaged with provided tips. These open codes were then compared and grouped together to look for common themes, which led us to narrow down our open coding to the following two research questions in subsequent rounds:

RQ1: How did workers view or conceptualize the tips provided to them?
RQ2: What barriers kept people from using tips, either initially or in later rounds?

Comparing and grouping these codes led us to our preliminary codebook of four different ways in which workers view and use tips (*rules*, *directional principles*, *options to try*, and *highlights*) and six different barriers to using tips (*counterintuitive*, *hard to implement*, *bad outcomes*, *lack of clarity*, *hard to track*, and *misaligned incentives*). We plan to conduct a final round of coding with multiple coders using this codebook plus an *other* category for each research question (to account for anything we may have missed in our initial codes) to code all the responses. We will then compute an inter-rater reliability score and make refinements of our codebook based on discussions of any differences. We also plan to perform the same analysis on workers in the other tip conditions. Responses for the *Human* tip will be interesting since the tip is from a non-optimal policy (but moves participants in the same direction as the optimal one). Responses for the *Baseline* tip will be interesting since the tip is from an optimal policy, but does not consider what can help participants the most.

4 HOW WORKERS VIEW AND USE TIPS AND THE BARRIERS THEY ENCOUNTER

In what follows, we discuss our findings for each of these two themes followed by a discussion of their implications in **Section 5**. To help convey the context underlying the participant quotes illustrating our findings, our blockquotes are in the form (PID, $c_1|c_2|c_3|c_4$, $d_1|d_2|d_3|d_4$), where PID is the participant, a_i refers to the number of times they had the server cook in rounds 1 to 4 and d_i refers to the duration that they were able to achieve (with 34 being the minimum duration possible). For example, *(P90, 2|2|2|2, 36|37|36|37)* refers to a quote from participant 90 who had the server cook twice in each of the four rounds and achieved a duration of 36, 37, 36, and 37 in those four rounds.

4.1 How workers view and used tips: rules, directional principles, options to try, and highlights

Of the workers who started off with an optimistic view of the tip, a few viewed the tips as *rules* that they *"had to figure out how to incorporate" (P90)*. For these workers, tips constrained the solution space they had to explore when sensemaking. For example, workers said:

> *"I knew that the server took longer to cook but HAD to cook twice so I had to figure out how to incorporate it" (P90, 2|2|2|2, 36|37|36|37)*

> *"I thought of it as a rule and not a tip, even though it didn't say it was a rule. So, I followed the tip..." (P108, 2|2|2|2, 34|36|34|34)*

Another group of workers also had an optimistic view of the tips, but described them as *directional principles* to focus or be more cognizant of. For these workers, they did not feel like they needed to follow it exactly, but the tip guided them in becoming more aware about using the server to cook. Workers said:

> *"I didn't try to have the serve cook twice, but I was cognizant and more aggressive with having them cook in general- I just didn't track the exact number of times." (P183, 1|2|1|2, 39|39|38|40)*

> *"It was very helpful. It made me focus on making sure the server cooked more even if that was not his obvious strength." (P43, 1|2|2|2, 38|34|34|34)*

Unlike the optimistic view of the previous two groups, others viewed it more neutrally as *options to try*. They tested it in the first round and then acted according to what they found, saying:

> *"I tested it out the first round and found out that it worked, so I repeated it during later rounds." (P214, 2|2|2|2, 36|34|34|34)*

> *"I tried it the first time, but I don't think it was a good tip, so I ignored it the next times." (P128, 2|1|1|1, 41|38|38|38)*

Finally, there were many who were skeptical of the tip (with many choosing not to follow it initially). However, even for these workers, tips played important roles as highlights making those options more salient for the workers' own sensemaking and testing processes, or simply as something to try when nothing else worked. Like Schelling points in game theory [15], they provided focal points that made those options more prominent compared with other options in the environment. Workers said:

> *"I thought that it was kind of suspicious at first but as I was figuring the game out myself, I thought that it was correct." (P195, 2|2|2|2, 35|40|34|34)*

> *"I did not listen to the tip the first two times since he takes more ticks but noticed when I incorporated it, I was more efficient" (P52, 1|1|2|2, 38|38|36|34)*

> *"At first I didn't follow it because it seemed counter intuitive since they're slow. But then I had trouble, so I tried it and came out ahead." (P5, 1|1|2|2, 38|38|34|34)*

4.2 Barriers to adoption: counterintuitive, hard to implement, bad outcomes, lack of clarity, hard to track, and misaligned incentives

Workers described many barriers that kept them from using the tip or to using it successfully. One of the most common barriers expressed was that the tip was *counterintuitive* and did not make sense logically, which caused many to not follow the tip initially (and as will be seen later, the counterintuitive nature of the tip also compounded some of the later barrier):

> *"The first round, I ignored it because I knew the sous chef would do it quicker." (P229, 1|2|2|2, 36|34|34|34)*

> *"At first I didn't follow it because it seemed counter intuitive since they're slow." (P5, 1|1|2|2, 38|38|34|34)*

Other workers talked about how it was *hard to implement*, which seemed to relate to it being counter intuitive. Because tips need to be implemented within the context of a broader strategy, workers had to develop some sense of how it worked to apply it:

> *"I had a difficult time incorporating it and using it to my advantage. It always felt like the server took longer than needed when I could have had them doing other tasks." (P167, 1|2|2|1, 38|35|40|38)*

> *"I tried to incorporate it into my strategy but somewhere along the way I got lost." (P96, 2|1|1|2, 37|38|38|40)*

> *"It wasn't as useful as the tip in the first three rounds. I didn't really know how to implement it into my own strategy, or what it really implied." (P132, 1|2|1|1, 38|40|36|38)*

A third challenge was that incorporating the tip (by having the server cook twice) could result in worse outcomes. These *bad outcomes* caused people to abandon the tip:

> *"I tried it the first time, but I don't think it was a good tip, so I ignored it the next times." (P128, 2|1|1|1, 41|38|38|38)*

> *"I let the server cook twice in the last couple of rounds and it didn't work well. If the game had continued I would have let the server only cook once." (P101, 1|1|2|2, 36|36|38|39)*

The previously described 3 barriers were the most commonly expressed, but there were also other barriers revealed in our analysis. For example, a few people felt there was a *lack of clarity* regarding what the tip actually meant concretely. They said:

> *"I wasn't sure what it meant. Does chopping count as cooking?" (P133, 0|1|1|1, 42|36|38|41)*

> *"I thought it was a little too broad, but maybe I'm just stupid because I could not figure out how to finish in less than 40 ticks." (P27, 1|2|2|2, 40|41|40|41)*

> *"I was really confused about this tip, I wasn't sure what it meant by let the server cook twice. I did this and it did not really help me, but maybe I misinterpreted the tip." (P135, 1|1|2|1, 38|39|36|39)*

Another barrier was that it was sometimes *hard to track* how many times the server had cooked so they did not know whether they had implemented the tip or not. This is a variant of the 'hard to implement' barrier, but unlike those quotes where participants focused on the challenge of getting it to work logically, this participant encountered more of a logistical challenge:

> *"It was confusing, I couldn't keep track of if he cooked or not" (P88, 1|1|1|1, 38|36|36|35)*

Finally, one participant touched on *misaligned incentives* in that trying to figure out how to implement the tip could result in lower short-term compensation:

> *"i didn't like it because i believe it took me longer to finish and i didnt receive any bonuses in those weeks" (P225, 1|2|1|2, 38|48|39|39)*

5 DISCUSSION AND FUTURE DIRECTIONS

Our analysis provides a richer picture of the diverse ways in which workers view and use tips and the barriers the prevent the tips from being useful for workers. We note that some of our findings may be specific to the unique context of problem-solving tasks. Specifically, our study considered a setting in which workers sought to achieve a *known* minimal duration for a sequence of actions (a *search problem*) with tips taking the form of *constraints* that an action sequence should satisfy (rather than, for example, specific actions recommended for specific points in time) and coming from *unknown sources* (workers were not told whether tips were from humans or algorithms). These caveats are important to note, but we also believe that these attributes are similar to many problem solving contexts in industry and. As described in the Introduction, managers are often engaged in optimizing a complex sets of decisions towards known industry benchmarks. With these caveats in mind, we discuss implications and directions for future study.

5.1 Diverse reasons for lack of trust

First, we saw that there were diverse reasons why workers did not trust the algorithmic tip. Workers felt that the tip was counterintuitive or led them to bad outcomes. These suggest different solutions to supporting workers that would be interesting to evaluate in follow-up studies. For example, one might look for ways to provide more explainable tips that are less counterintuitive. But one could also try to increase worker confidence of the tip's value by citing statistics of others who vouch for it or by showing that others also encountered bad outcomes on the path to implementing the tip successfully.

The fact that the tip could lead people down worse paths also raises questions about what it means for a tip to be "optimal". For example, while the server needed to cook twice to achieve the optimal duration of "34", it was also possible to get to a duration of "35" with the server cooking once. If the former case had a much higher likelihood of resulting in long durations than the latter case and may take a much longer time to figure out, then it could be true that the average payout for "server cooks once" could actually be higher, especially in the short-term.

5.2 Trust was not the only obstacle to benefiting from tips

We also saw that lack of trust was not the only obstacle preventing people from benefiting from tips. Several of the barriers we observed related to the "usability" of the tip with workers finding the tip lacked clarity, was hard to implement, or was just hard to even track whether or not they had used it. This partially related to the fact that the tip was a constraint on a complex and interdependent sequence of actions which workers were not always able to foresee in advance. These suggest solutions such as creating feedback or communication mechanisms for raising questions about clarity, creating more actionable tips guiding implementation, and creating tools for workers to track relevant statistics or map complex spaces.

We also saw that broader environmental factors could shape people's use of tips such as the incentives that workers are working within and whether those incentives encourage and support learning and exploration. This is an important

reminder that it is not enough to design at the level of the human-ai interactions themselves. Agent interactions cannot be divorced from the broader organizational, community, or societal contexts in which the interactions are situated.

5.3 Different approaches to incorporating tips

Our findings also seem to reflect different approaches that people take to problem solving. Some people are logicians emphasizing reasoning about why strategies work. Others are experimenters emphasizing trial and error. These different approaches may affect how open people are to trying out counterintuitive tips, the barriers that affect a tips usefulness to them, and the design interventions that would be effective for encouraging tip use. It would be interesting to study this further. For example, do certain personality types predict the ways in which people react to algorithmic tips? How about teams of people with different combinations of these personality types? What implications might this have on forming teams or using collaborative interactions to facilitate more effective use of algorithmic tips? We see this as a particularly interesting direction not just because of potentially different approaches to incorporating tips, but also because collaboration could also help workers reason about counterintuitive tips or to figure out what a tip means and how to implement it.

Finally, we also found it interesting that people who ignored or discounted tips could still benefit from them and that tips could provide value through being highlights for making certain directions more salient. It would be interesting to think about new types of tips that aren't necessarily trying to convince people of a specific action to take, but are simply trying to support by shining a light on information or strategies that may be subtle or counterintuitive.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine Bias. In *Ethics of Data and Analytics*. Auerbach Publications, Boca Raton, 254–264.

[2] Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. 2021. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454* (2021).

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 1–21.

[4] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[5] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19, Paper 559)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[6] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[7] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[8] I Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* 23, 1 (Aug. 2001), 89–109.

[9] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19, Paper 487)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[10] Kathleen L Mosier, Linda J Skitka, Susan Heers, and Mark Burdick. 2017. Automation bias: Decision making and performance in high-tech cockpits. In *Decision Making in Aviation*. Routledge, 271–288.

[11] Bureau of Transportation Statistics. 2021. Airline on-time performance and causes of flight delays. https://www.bts.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays

[12] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 237)*. Association for Computing Machinery, New York, NY, USA, 1–52.

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

[14] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251.

[15] Thomas C Schelling. 1980. *The Strategy of Conflict: With a New Preface by the Author.* Harvard University Press.

[16] Yunfeng Zhang, Q Vera Liao, and Rachel K E Bellamy. 2020. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305.