Overreliance on AI: Literature review

AETHER AI ETHICS AND EFFECTS IN ENGINEERING AND RESEARCH

Overreliance on AI occurs when users start accepting incorrect AI outputs. This can lead to issues and errors that can ultimately make people lose trust in AI systems. This report explains what overreliance on AI is, how it happens, and how we can mitigate it.

An important goal of AI system design is to empower users to develop ***appropriate reliance*** on AI. This is important given that policymakers and practitioners call for greater human oversight—making users the last line of defense against AI failures. This report shows how and why overreliance on AI makes it difficult for users to meaningfully leverage the strengths of AI systems and to oversee their weaknesses. Based on a literature review of ~60 papers from different research areas, this report provides a detailed overview of how overreliance on AI happens, how to measure overreliance, what its consequences are, and how we can minimize its negative effects.

Authors

Samir Passi
User Researcher
v-samirpassi@microsoft.com

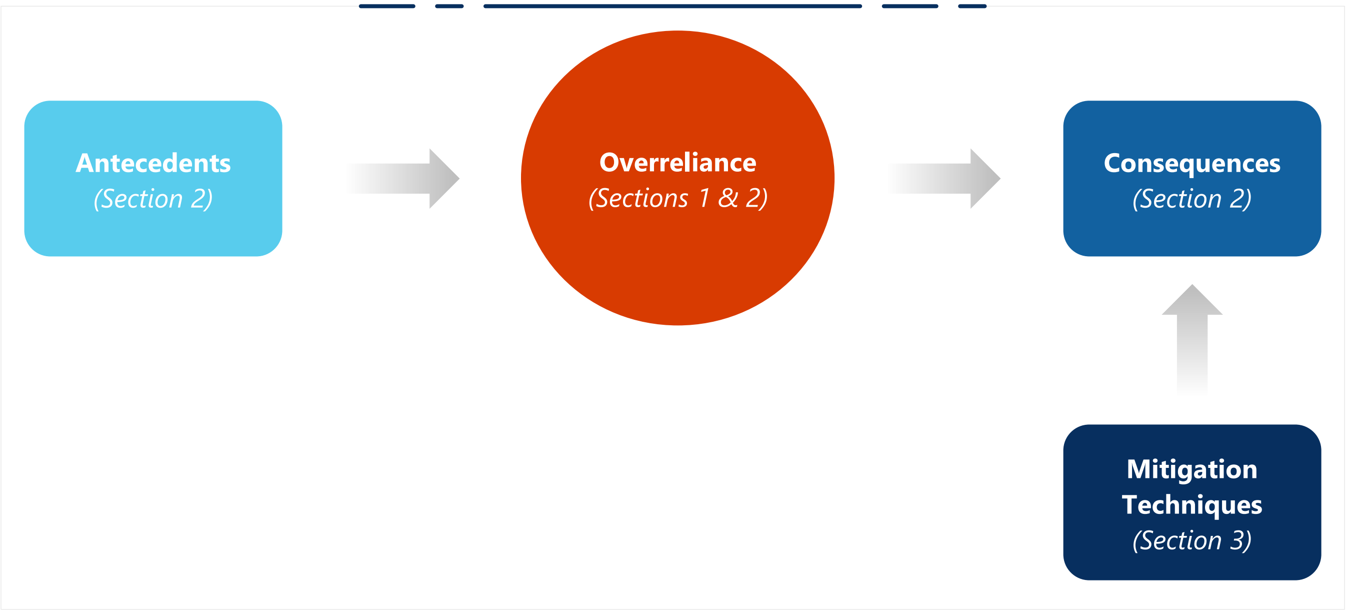
Mihaela Vorvoreanu
Director, Aether UX Research & Education
mihaela.vorvoreanu@microsoft.com

Introduction

This report synthesizes ~60 research papers about overreliance on AI. The papers originate from a variety of disciplines, including Human-Computer Interaction (HCI); Human Factors; Intelligent User Interfaces (IUI); Computer Supported Cooperative Work (CSCW); Organizational Science; and Fairness, Accountability, and Transparency (FAccT).

The report has three sections:

1. **What is overreliance on AI?** – Definition of overreliance and ways to assess and measure overreliance. Skip to Relevant terms and related measures.
2. **Antecedents, mechanisms, and consequences of overreliance on AI** – Overview of how pre-existing conditions affect overreliance on AI, how overreliance on AI manifests in practice, and its negative implications. Skip to Antecedents, mechanisms, and consequences summary.
3. **Techniques to mitigate overreliance on AI** – Ways to reduce overreliance on AI. Skip to Mitigation techniques summary.



1

What is overreliance on AI?

1.1 Definition

Overreliance on AI is defined as users accepting incorrect AI recommendations—i.e., making errors of commission. Overreliance generally happens when users are unable to determine whether or how much they should trust the AI. Users have difficulty determining appropriate levels of trust because they lack awareness of:

- What the AI system can do
- How well it can perform
- How it works

An important goal of AI system design is to empower users to develop *appropriate* reliance on AI. However, how to do so is complicated. Appropriate reliance is a moving target because it is hard to operationalize and depends on context and application domain.

It is not always obvious when users over-rely on AI. Imagine an AI system that does tasks better than humans. Users make more accurate decisions when they over-rely on this system compared to when they work alone. When the human+AI team performs better than the human working alone, the situation seems acceptable and unproblematic. However, all AI systems make mistakes. Sometimes AI systems make mistakes that are different from those humans make. Sometimes AI systems start making new kinds of mistakes after receiving model updates (Bansal et al. 2019). Thus, even when humans perform well using AI, the unpredictability of AI mistakes warrants caution against overreliance.

Policymakers and practitioners call for greater human oversight—i.e., make humans carefully review AI recommendations before making final decisions. While calls for human supervision assume “fluid cooperation” between humans and AI, “the dynamics of shared control between […the two] are more complicated” (Elish 2019: 5). As we describe in this report, humans are unable to mitigate AI shortcomings when they start over-relying on AI systems. Therefore, calls for human oversight can provide a false sense of security (Green 2021; Koulu 2020). For these reasons, understanding, monitoring, and studying user reliance on AI is a big priority.

1.2 Relevant terms and related measures

Here are some measures of overreliance commonly encountered in research literature:

Overreliance measure	Explanation
Agreement with incorrect recommendations	A common way to measure overreliance is to count how often users accept incorrect AI recommendations (out of all incorrect recommendations shown to them) (Buçinca et al. 2021).
Switch fraction	*Switch fraction* measures how often users completely change their answers to match AI recommendations (Lu & Yin 2021; Kim et al. 2021).
Modifying answers to align with AI recommendations	A way to extend the switch fraction is to measure how much users change their answers to align with AI recommendations—e.g., percentage of change (Kim et al. 2021).
Weight of advice (WOA)	WOA measures the importance users give to AI recommendations (Logg et al. 2019). It is a way to quantify the impact of *switch fraction* and the extent to which users modify their answers to align with AI recommendations. A value of WOA=1 indicates overreliance. $WOA = \frac{revised_{prediction} - initial_{prediction}}{\mathrm{AI}_{recommendation} - initial_{prediction}}$

Key terms related to overreliance measures:

a. **Agreements:** How often user predictions are the same as AI recommendations, when users make predictions before seeing AI recommendations (Lu & Yin 2021).

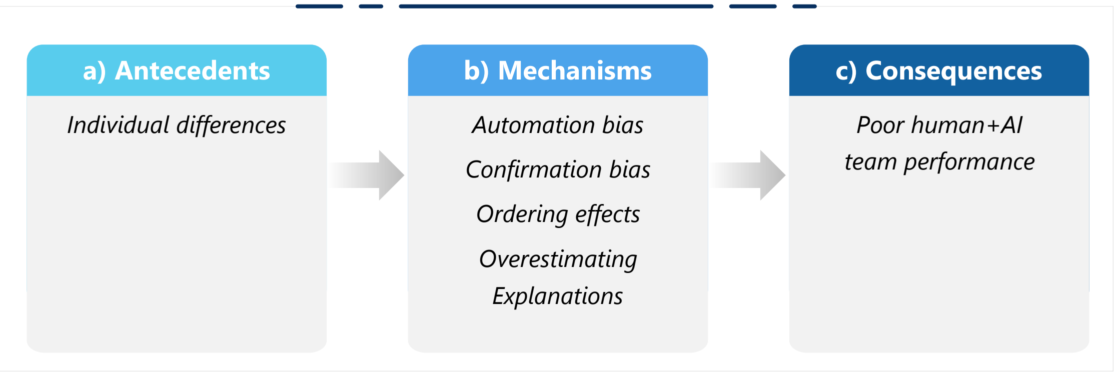
b. **Disagreements:** How often user predictions are different from AI recommendations, when users make predictions before seeing AI recommendations (Ibid.)

c. **Human errors:** How often users make incorrect predictions when user predictions are different from AI recommendations, but AI recommendations are also wrong (Buçinca 2021).

d. **Delegation:** How often users let an AI system fully make decisions on their behalf (Chiang & Yin 2021).

2 Antecedents, mechanisms, and consequences of overreliance on AI

This section describes (a) pre-existing conditions that affect user overreliance, (b) how and why users over-rely, and (c) the negative consequences of overreliance.



a) Antecedents

Individual differences

b) Mechanisms

Automation bias

Confirmation bias

Ordering effects

Overestimating Explanations

c) Consequences

Poor human+AI team performance

2.1 Antecedents of overreliance on AI

Individual differences

Individual differences affect user reliance on AI.

Individual differences refer to differences in users' demographic, social, cultural, and professional characteristics. **Individual differences lead users to develop both over- and under-reliance on AI.**

a. **AI literacy:** AI literacy is the measure of how much users know about AI.[1] AI literacy affects users' attitudes towards AI.[2] Users with and without AI background develop inappropriate reliance in different ways. Users with high AI literacy overestimate the utility of numbers in explanations (e.g., believing that numbers can help debug AI) while users with low AI literacy overestimate the AI's intelligence if it provides numeric explanations (e.g., believing that numbers are a sign of objective logic) (Ehsan et al. 2021). Users with low AI literacy are often most affected by AI recommendations. For instance, in a study involving medical decision-making scenarios, clinicians with low AI literacy were seven times more likely to select medical treatments that aligned with AI recommendations (Jacobs et al. 2021).

b. **Expertise:** Domain expertise is the measure of how much users know about the task domain. Both low- and high-expertise users can develop overreliance on AI (Gaube et al. 2021). Low-expertise users often show *algorithmic susceptibility*—tendency to accept AI recommendations at a high rate. High-expertise users often develop *algorithmic aversion*—self-reported tendency to disregard AI recommendations—but still rely heavily on AI while making decisions.

[1] For more information on AI literacy, including how to measure it, see Long and Magerko (2020).

[2] For more information on user attitudes towards AI, including how to measure them, see Zhang & Dafoe (2019).

c. **Task familiarity:** Closely related to domain expertise is *task familiarity*—the measure of how familiar users are with the task. High task familiarity does not necessarily imply high expertise. For example, a user may have high familiarity with *programming* but have low domain expertise with a new programming language. Users with high task familiarity (a) report more trust in AI but show less adherence to its recommendations and (b) tend to over trust AI systems in the presence of explanations (Schaffer et al. 2019). High task familiarity leads users to become overconfident in their own ability to perform the task. The more confident users are, the less well they perform when working with AI systems (Green & Chen 2019).

Recommendations

Design AI features for variation in user characteristics such as confidence, expertise, task familiarity, AI literacy, and attitudes towards AI.

> *See GenderMag: A Method for Evaluating Software's Gender Inclusiveness for evaluating software in light of different cognitive facets: Computer-self efficacy, information processing, attitude towards risk, and motivation.*

Closely monitor low-expertise users.

> *For example, pay attention to novice users who use the AI system for help with tasks.*

Pay attention to how users over-rely on AI when doing more and less familiar tasks.

> *For example, use telemetry to analyze how users accept different kinds of AI recommendations and develop a taxonomy of different overreliance issues.*

Gauge user confidence in both the AI and their own ability to perform the task.

> *Nudge users to actively reflect on their own work to keep user overconfidence and automation complacency in check.*

2.2 Mechanisms of overreliance on AI

This subsection explains *how* and *why* users over-rely on AI.

Automation bias

Automation bias is the tendency to favor recommendations from automated systems, while disregarding information from non-automated sources. **Users with automation bias often over-rely on AI.**

1. **Users with high automation bias are unable to develop appropriate reliance on AI when its performance changes** (Pop et al. 2015). AI systems that initially work well can later start making mistakes. Users with automation bias over-rely on systems that perform well. But the same users trust the system less after seeing it fail. Later, when the system performs well again, there is no guarantee that it can earn back user trust. User trust in AI goes down by a relatively large amount when system capability decreases but increases by a much smaller amount when system capability increases back again.
2. **Users show high automation bias when working on objective and unfamiliar tasks.** Automation bias causes users to “constantly give more weight to equivalent advice when it is labeled coming from an algorithmic versus human source” (Logg et al. 2019: 92). Users thus over-rely on AI when they do not have enough knowledge and skills to properly evaluate AI recommendations.
3. **Users show less automation bias when working on subjective tasks.** Users rely more on human suggestions when working on subjective tasks, in part because users assume human decision-making is easier to understand (Yeomans et al. 2019). AI often assists decision-making in scenarios that have a mix of objective (those with a singular metric of success) and subjective (those with multiple metrics of success) tasks. Keep in mind that user reliance operates on a spectrum (under- to overreliance) but also depends on the nature of the task (objective vs. subjective). A user may over-rely on AI for unfamiliar parts of their work but under-rely on AI for familiar parts of their work.

Recommendations

Identify ways to assess automation bias from telemetry.

Help users calibrate trust in AI, based on knowledge that automation bias causes overreliance.

Monitor user overreliance post deployment as trust in AI fluctuates over time.

Do further research to understand the differential costs of AI errors.

Confirmation bias

Confirmation bias is the tendency to favor information that aligns with prior assumptions, beliefs, and values. **Users over-rely on AI when its recommendations align with their own predictions.** Confirmation bias leads users to further strengthen the beliefs they already have about AI.

1. When faced with confirmation bias, **users over-rely on AI when they (a) know less about how well AI systems work and (b) are more confident in their own ability to do the task** (Lu & Yin 2021). In such situations, users over-rely on AI *regardless* of the correctness of its recommendations and perceive it as being more accurate, competent, reliable, and understandable. Confirmation bias makes users wrongly assume that the AI uses logic and reasoning similar to their own.
2. *Caveat:* **Confirmation bias can lead users who under-rely on AI to further distrust AI** (Lee & Rich 2021). Users often develop *algorithmic aversion*—a biased negative assessment of algorithmic systems. Users with algorithmic aversion frequently expect AI to give wrong recommendations. If users expect the AI to fail and it does fail, their trust in AI further deteriorates.

Recommendations

Ensure that users have at least a minimum knowledge of how AI features work. See Guidelines for Human-AI Interaction, Guideline 2: Make clear how well the system can do what it can do.

Use onboarding techniques and tutorials to make users aware that overreliance is a common phenomenon. For instance, provide examples of confirmation bias.

Nudge users to engage in meta-cognition.

> *For example, provide session statistics and/or a list of items to review with users at the end of a working session to help them reflect on their work.*

Ordering effects

Ordering effects refer to how changing the order of presented information alters user perceptions and decisions. For AI, ordering effects occur based on whether users see the AI system succeed or fail during early interactions. **The *timing* of AI errors significantly affects user reliance.**

1. **Users over-rely on AI if it does well during initial interactions but under-rely on it if it fails during initial interactions**. Users who see AI perform well early on often develop automation bias and complacency, making significantly more errors due to positive first impressions (Nourani et al. 2021). Users who see the AI fail early on often develop algorithmic aversion (Kim et al. 2020).
2. **User expertise alters the impact of ordering effects** (Nourani et al. 2020). Novice users over-rely on AI regardless of whether it fails or succeeds during early interactions because they do not have sufficient knowledge to identify errors. Expert users show a more complex behavior: When AI fails during early interactions, experts develop under-reliance on AI. The under-reliance never fully goes away, even when the AI starts doing better (see #1 under *automation bias*). When AI does well during early interactions, experts develop overreliance on AI. If the AI starts doing less well later, experts find it relatively easy to appropriately adjust their trust on AI.
3. **Ordering effects are tied to a cognitive bias called *anchoring effect***—relying too much on the first piece of provided information when making decisions. Anchoring effects happen in two ways:
    a. **Anchoring effect happens when users see AI recommendations before making their own decisions.** AI recommendations act as anchors and significantly influence users' decision-making processes (Vaccaro & Waldo 2019). Users often alter their decisions to make them align with AI recommendations.
    b. **Anchoring effect happens when users see information about AI (e.g., accuracy) before interacting with it.** AI's stated accuracy negatively affects user reliance (Yin et al. 2019). If stated accuracy is low, then even when the AI performs well, users continue to distrust it. If stated accuracy is high, then when the AI makes mistakes, users lose trust in it even if its accuracy is better than their own.

Recommendations

Use onboarding techniques and tutorials to influence users' first impressions with AI systems as those are crucial for developing appropriate reliance and trust.

Ensure that users correctly interpret AI performance metrics such as accuracy scores.

> *See Guidelines for Human-AI Interaction, Guideline 2: Make clear how well the system can do what it can do.*

Monitor for anchoring bias in addition to ordering effects.

> *For example, when users choose to see the top AI recommendations, the first recommendation can cause anchoring bias.*

Overestimating explanations

Explanations help users better assess and understand AI recommendations. However, **detailed explanations often lead users to develop overreliance on AI.**

1. **Explanations increase user reliance on all AI recommendations.** Showing explanations to users with high task familiarity leads to automation bias (Schaffer et al. 2019). Increasing the level of detail in explanations leads to more trust in AI but also overreliance on AI (Bussone et al. 2015). In fact, even explanations with no basis in the AI's actual working can make users trust AI more (Lai & Tan 2019; Ehsan et al. 2021). The effects become worse when AI is used in subjective domains. Detailed explanations make users believe that the AI reasons about the task in a manner similar to humans (Bussone et al. 2015). High-fidelity explanations lead users, especially novice users, to trust bad models (Papenmeier et al. 2019).
2. **Explanations increase user reliance on incorrect AI recommendations.** Explanations increase "blind trust" rather than "appropriate reliance" on AI (Bansal et al. 2020). Users make poor decisions when incorrect recommendations are accompanied by explanations (Zhang et al. 2020). For example, a non-informative explanation such as an accuracy score improved user trust in AI even when the claimed accuracy was as low as 50% (Lai & Tan 2019).
3. *Caveat:* **Explanations can also make users lose trust in AI and under-rely on it.** Providing explanations can lead to the problem of *explanation mismatch*—AI providing an explanation that does not align with user expectations (Papenmeier et al. 2019). This can also happen for non-explanation performance measures such as accuracy and confidence scores. For example, the AI gives a recommendation that the user knows is incorrect, but the AI gives that recommendation a high confidence score.

Recommendations

Be careful with providing explanations for recommendations because they increase trust in incorrect recommendations.

> *For example, consider running a study to see how users interact with AI recommendations with and without accompanying explanations. Watch out for potential issues regarding confirmation bias and explanation mismatch.*

Conduct user research to understand the impact of explanation types and patterns on user overreliance. See Guideline for Human-AI Interaction, *Guideline 11: Make clear why the system did what it did.*

> *For example, consider running a study with two types of explanations—one that explains the recommendation (e.g., what it is) and another that explains why the recommendation was generated (e.g., fit with existing context).*

Do not use language that anthropomorphizes AI in the user interface, explanations, and marketing materials.

2.3 Consequences of overreliance

This subsection describes the negative impacts of overreliance on AI.

Poor human+AI team performance

A human+AI team is not guaranteed to perform better than the human or AI working alone. **Overreliance on AI leads users to perform worse on tasks compared to the performance of the user or AI working alone** (Bansal et al. 2020; Buçinca et al. 2021; Green & Chen 2019a, 2019b; Jacobs et al. 2021; Lai & Tan 2019; Zhang et al. 2020).

Poor human+AI team performance happens for several reasons:

1. **Users alter, change, and switch their actions to align with AI recommendations** (Gaube et al. 2021; Green & Chen 2021; Poursabzi-Sangdeh et al. 2021; Suresh et al. 2020). For example, in a study, researchers told users that an *intelligent* algorithm will evaluate the text written by users to indicate whether it has positive or negative tone (Springer et al. 2017). The algorithm, however, merely generated random output. Nearly twice as many users rated the random algorithm as being accurate and placed an excessive amount of trust in it. Misplaced reliance on AI makes it difficult for users to identify and resolve AI errors (Vaccaro & Waldo 2019). Even the most accurate AI does not guarantee the best human+AI team performance (Bansal et al. 2021).
2. **Users find it difficult to evaluate AI's performance and to understand how AI impacts their decisions.** For instance, users often overestimate system accuracy (Nourani et al. 2021) and do not realize when they have ceded control to the AI (Levy et al. 2021). The situation where users let AI make decisions on their behalf can be especially harmful for human+AI team performance and productivity. Incorrect recommendations significantly lower user accuracy on tasks (Jacobs et al. 2021). **In fact, users who receive incorrect recommendations are often slower than users who do the task from scratch without AI's help** (Levy et al. 2021).
3. **Overreliance makes users trust AI in scenarios where they should not.** For example, users get confused when there is a big mismatch between their answers and AI recommendations (Kim et al. 2021). The substantial errors make users incorrectly assume that they have made a mistake when, in fact, the AI is at fault. The scenario gets worse when users work with new data (e.g., out-of-distribution data—data that the AI has not seen during training). Users expect AI to *maintain* its performance on new data but assume that their own performance will worsen on new data (Chaing & Yan 2021). Users thus end up relying more on AI when dealing with out-of-distribution data, **leading them to trust AI more when its performance is questionable and uncertain.**

Recommendations

Compare human+AI performance to the human working alone, especially in the presence of problematic recommendations.

> *AI systems are at high risk for poor human+AI performance if AI's performance is at best similar to or worse than humans.*

Monitor human+AI performance over time for different kinds of users, especially when introducing new AI features.

> *See Guidelines for Human-AI Interaction, Guideline 14: Update and adapt cautiously.*

Summary: Antecedents, mechanisms, and consequences of overreliance on AI

		Short description	Mitigation techniques
Antecedents of overreliance	Individual differences	Differences in users' demographic, professional, social, and cultural traits affect their reliance on AI.	Provide personalized adjustments for users; Effectively onboard users; Give users choice
Mechanisms of overreliance	Automation bias	Tendency to favor recommendations from automated systems, while disregarding information from non-automated sources.	Effectively onboard users; Employ cognitive forcing functions; Provide personalized adjustments to users; Provide real-time feedback
	Confirmation bias	Tendency to favor information that aligns with prior assumptions, beliefs, and values.	Employ cognitive forcing functions; Effectively onboard users; Provide personalized adjustments to users; Provide real-time feedback
	Ordering effects	The order of presented information affects user perceptions and decisions. The *timing* of AI errors significantly affects user reliance.	Effectively onboard users; Provide personalized adjustments to users; Alter speed of interaction;
	Overestimating explanations	High-fidelity explanations can lead users to develop overreliance on AI.	Be transparent with users; Provide real-time feedback; Provide effective explanations
Consequences of overreliance	Poor human+AI performance	Overreliance causes poor human+AI team performance compared to the human or AI working alone.	All

3 Mitigation techniques for overreliance on AI

This section provides a list of mitigation techniques based on existing research to address overreliance on AI.

3.1 During initial interactions

First impressions play a crucial role in shaping user reliance on AI. This subsection outlines mitigation techniques that can be used during a user's initial interactions with AI systems to help develop appropriate reliance.

Effectively onboard users

AI systems should have effective onboarding capabilities and techniques (Chaing & Yin 2021; Lai & Tan 2019; Lu & Yin 2021; Nourani et al. 2021). For instance:

1. **AI systems should show examples of both correct and incorrect recommendations to help users develop appropriate first impressions** that "cover the variability of system capabilities" (Nourani 2020).
    a. Users are more willing to use algorithmic systems when they do not see systems make mistakes (Dietvorst 2015). However, never seeing the system err makes users over-trust its capabilities.

Take care not to overwhelm users with information during onboarding (Suresh et al. 2020). Identify ways to onboard users progressively to different AI features.

Be transparent with users

Providing information about AI models helps users develop appropriate reliance on AI (Yin et al. 2019). Follow the *transparency principle*. Ensure that users understand what you are telling them, they are adjusting their behavior and expectations accordingly, and that those changes survive over time.
For instance:

1. Provide users basic information about global model properties such as accuracy, design objective, as well as strengths and limitations to help them better assess AI recommendations (Cai et al. 2019).
    a. When using the accuracy score, ensure that you properly communicate to the user what the score implies (e.g., binary vs. multi-class classification).
    b. Gather well-known edge cases and report AI's performance on them. This helps users know contexts where they must be more careful while using the AI system.
2. Provide further information about the intended use cases of an AI system to help users better understand when and whether to trust the AI.
    a. Examples include information on use cases anticipated during development, benchmarked model evaluations in different conditions, and relevant training data details (Chiang & Yin 2021).

Provide personalized adjustments for users

AI systems should tailor their onboarding experiences to account for differences in user characteristics. For instance:

1. Devise strategies to assess *automation bias* based on early user interactions during onboarding/tutorials and, accordingly, adjust the level of automation and feedback for both low and high automation-bias users (De-Arteaga et al. 2020; Levy et al. 2021).
2. Devise strategies to assess users' confidence in their own abilities (and predictions) and, accordingly, adjust the user experience (UX) to help under- and over-confident users develop appropriate reliance (Gaube et al. 2020; Lu & Yin 2021; Schaffer et al. 2019).
3. Devise strategies to assess *AI literacy*—how much users know about AI—and adjust the UX to help users with low or high AI literacy to develop appropriate reliance (Chaing & Yin 2021; Jacobs et al. 2021; Wang et al. 2020).
4. Alter the sequence of AI success and failure scenarios during early interactions to mitigate the impact of *ordering effects*. If, for example, it is acceptable to sacrifice accuracy on tasks, show AI strengths first before introducing failure scenarios to help users develop a better mental model of the AI system (Nourani et al. 2021).

3.2 During regular use

This subsection describes mitigation techniques that can be used during a user's routine interactions with an AI system they've been familiar with for some time.

Employ cognitive forcing functions

Cognitive forcing functions (CFFs) are interventions that interrupt a person's routine thought process and make them engage in analytical thinking (Lambe et al. 2016). Over time users get complacent about AI systems; they start using mental shortcuts and spend less effort evaluating AI recommendations. **Use CFFs to shift users from a fast and automatic thinking process to one that is slow and deliberative** (Wason & Evans 1974; Kahneman 2011). Specifically concerning AI:

1. **CFF designs significantly reduce overreliance on incorrect AI recommendations** (Buçinca et al. 2021).
    a. Devise CFF strategies to increase users' motivation to engage with AI recommendations, performance metrics, and explanations. Examples of CFFs include checklists, time-outs, on-demand explanations, and asking users to explicitly rule out alternatives.

    b. *Caveat*: CFFs mitigate overreliance but are often less favored by users because of the added cognitive burden. Do further research to know the applicability of CFFs to specific use-cases.

Provide real-time feedback

Providing real-time feedback to users about their performance and the AI's for better human+AI team performance (Lai et al. 2020). Real-time feedback helps users triangulate their decisions when working with AI recommendations (De-Arteaga et al. 2020). For instance:

1. Provide high-level information about AI, such as accuracy scores, to users (Lu & Yin 2021).
    a. **Do not uncritically present high performance scores because they cause user overreliance**. User reliance is affected by the AI systems' stated accuracy (Lai & Tan 2019). However, users often take accuracy scores at face value and are not made aware that model scores are inherently uncertain. For example, pre-release performance benchmarks are often high because they are calculated on controlled, sanitized datasets.

    b. *Caveat*: Overwhelming users with more information about an AI system's "training data, model architecture, performance, and recommendations all lead to [...users] following both correct and incorrect recommendations more often" (Suresh et al. 2020: 315). **Do further research to know what forms of information users need (and respond correctly to) in different contexts.**

2. Use confidence scores to help users develop appropriate trust in AI (Zhang et al. 2020).
    a. **Develop ways to help users correctly interpret confidence and uncertainty scores.** Users desire confidence scores but often find them difficult to interpret (Gaube et al. 2021).

    b. *Caveat*: Confidence scores can backfire and must be used strategically (Yin et al. 2019). For example, high confidence scores for evidently incorrect recommendations cause users to develop algorithmic aversion.

3. Inform users when they accept potentially problematic or incorrect AI recommendations (Levy et al. 2021). Examples include recommendations with low confidence scores, those based on limited data, and those containing fabricated elements (e.g., AI-generated datasets).
    a. **Train separate models to detect problematic outlier recommendations—**e.g., those based on abnormal or insufficient data (Poursabzi-Sangdeh et al. 2021).

    b. **Examine user attitudes towards algorithmic advice before system use** (e.g., are users prone to automation or confirmation bias) since the incorrectness of recommendations might not be obvious in many cases (Logg et al. 2019).

Provide effective explanations

It is not enough for AI to be accurate; it must also be understood (Yeomans et al. 2019). Explanations help users better assess the correctness of AI recommendations and the working of AI systems. However, detailed explanations often lead users to develop inappropriate reliance. **Explanations should thus not only justify AI recommendations but also ensure they help users develop appropriate reliance on AI.** For instance:

1. **Focus on building *better* explanations.** There is no clear recipe for building effective explanations. Explainable AI is an open research area, and we need further research to assess the efficacy and short-/long-term impact of different explanation types on overreliance and human+AI team performance.
    a. Build informative, not just convincing, explanations (Bansal et al. 2020). The goal of explanations is to increase trust in AI, but also to help users better evaluate AI recommendations. For example, do not just highlight data features, but also explain their importance (Lai et al. 2020).
    b. Pay close attention to the content of explanations (Dodge et al. 2019; Zhang et al. 2020). For example, explanations containing model performance metrics help users develop appropriate trust at the model level (e.g., 'this model performs well'). Explanations containing confidence/uncertainty scores help users develop appropriate trust at the recommendation level (e.g., 'this recommendation is less likely to be correct').
    c. Be careful with providing complex explanations because they may lead to higher response times and lower user satisfaction (Tan et al. 2018). Dense and lengthy explanations often backfire.

2. **Focus on how different explanations interact with other aspects of AI systems** to better understand how and why users may over-rely. (Nourani et al. 2021). For instance, analyze interaction effects between different explanation types (e.g., how vs. why) and the following characteristics:

   a. **User confidence** (high vs. low)
   For example, effects of explanations quickly wear off as user overconfidence increases (Schaffer et al. 2019).

      - Consider running a study to see how users react to explanations over time as they become more comfortable using the system. For instance, do users begin taking explanations for granted?
      - Consider running a study to see how more confident users interact with explanations. Are overconfident users less likely to generate and inspect explanations because they think they already know what the system does?

   b. **User agency** (e.g., can users edit AI recommendations before accepting them?).
   For example, effect of explanations in decision-making tasks is different from those in debugging tasks (Lai et al. 2020).

      - Use telemetry to create overreliance measures such as acceptance of problematic recommendations (with little to no edits post acceptance) and weight of advice (including the extent to which users edit recommendations post acceptance).

   c. **User biases** (e.g., automation vs. confirmation)
   For example, explanations influence the perceived intelligibility and working of AI systems (Bussone et al. 2015).

      - Consider running a study to see if users are more likely to accept AI recommendations with/without accompanying explanations.

   d. **Ordering effects** (e.g., success vs. failure).
   First impressions significantly affect user reliance on AI.

      - Some users see failures first and develop algorithmic aversion; use onboarding to help them see success scenarios and get a balanced perspective on the AI system.
      - Some users are enamored by AI; use onboarding to help them proceed with caution by seeing problematic scenarios (e.g., 'Top 3 things that can go wrong while using the AI system').

   e. **Task difficulty** (e.g., low vs. medium vs. high)
   Easy tasks lead to complacency, while difficult tasks lead to inappropriate reliance.

      - Make sure users do not go on autopilot when working with AI systems. Users should think carefully, slow down, reflect in metacognition, and remain vigilant. Use CFFs to nudge users to actively self-reflect on human+AI team performance.
      - People stay vigilant when there is variety. Identify ways to introduce forms of differences and inconsistency in the user experience of the AI system (e.g., aspects of gamification, checking for errors).

Alter speed of interaction

User reliance is affected by the AI's response time—the time it takes to make recommendations. The relation between response time and user reliance is complicated and depends, in part, on the perceived difficulty of the task and the order in which users see AI recommendations.

1. One group of researchers found that users trust good models more and bad models less if the response time is higher (Park et al. 2019). In this study, users estimated the number of jellybeans in a jar. Users made their predictions before seeing algorithmic recommendations and were not asked to actively reflect on the perceived task difficulty. Researchers found that the waiting time provided users with the opportunity to reflect on the task and estimate their own decision-making process and the AI's. **Identify ways to leverage response time to help users reflect on the human+AI team performance.**
2. Another group of researchers found that slow response times can at times have the opposite effect and make users perceive AI systems as less accurate (Efendić et al. 2020). In this study, users were told that they were either a university admissions officer or a corporate sales officer tasked with predicting the academic success of students or future product sales. Users saw recommendations before making predictions. All users agreed that making future predictions was a difficult task for humans but an easy one for algorithms. **Conduct further research to understand how to effectively use response time to address overreliance on AI.**

Give users choice

Conduct research and devise strategies to better incorporate *collaboration* as a feature in AI design—for example, whether the system will always provide recommendations or only upon request.

1. Research shows that **providing recommendations only upon request helps mitigate overreliance on AI** (Gaube et al. 2021). Regardless of the desired collaboration model, it is prudent to ask users to make their own predictions before seeing AI recommendations (Poursabzi-Sangdeh et al. 2021) or to provide users with the option to enable/disable AI recommendations.
    - Instead of giving users a universal enable/disable toggle for AI recommendations, identify tasks for which users may not want AI recommendations vs. those in which users are okay with recommendations. Use this to provide users with granular choices—for example, disable AI recommendations for an hour or disable AI recommendations for specific tasks.
2. *Caveat*: Further research is required for AI use-cases that are not binary conditions where users either completely accept or reject AI recommendations (Bansal et al. 2021).

Summary: Techniques to mitigate overreliances on AI

Time	Mitigation technique	Short summary	Issue(s) addressed

During initial interactions	Effectively onboard users	Provide both correct and incorrect predictions to help users develop appropriate first impressions.
		Customize tutorials for people with low/high automation bias, low/high AI literacy, and low/high task familiarity.
			Automation bias;
			Ordering effects;
			Poor human+AI performance

	Be transparent with users	Clearly communicate: (a) basic model properties (e.g., known strengths and limitations, overall design objective) and (b) intended use-cases (e.g., cases envisioned during development, benchmarked model evaluations).	Overestimating explanations; Poor human+AI performance

	Provide personalized adjustments for users	Evaluate user susceptibility (from tutorials and early results) to adjust automation accordingly.	Individual differences;
			Ordering effects;
			Poor human+AI performance;

During regular use	Employ cognitive forcing functions	Increase users' cognitive motivation to engage with AI recommendations, using techniques such as confidence and uncertainty information, accuracy scores, and cost of errors.	Automation bias;
			Confirmation bias;
			Poor human+AI performance

	Provide real-time feedback	Real-time feedback on human performance leads to improvement (e.g., alerting the user when they have accepted a risky recommendation).
		Give people ways to triangulate their decisions while working with AI models. Help people reflect on their own decision-making process.
			Automation bias;
			Confirmation bias;
			Overestimating explanations;
			Poor human+AI performance

	Provide effective explanations	Build informative, not just convincing, explanations. Explanations sensitive to model performance help users develop appropriate trust at model level.
		Explanations sensitive to prediction uncertainty help users develop appropriate trust at the recommendation level.
			Overestimating explanations;
			Poor human+AI performance;

	Alter speed of Interaction	Alter the AI system's response time and provide users ways to reflect on the task and estimate their own and the AI's decision-making process while they are waiting for the AI recommendation.	Ordering effects;
			Poor human+AI performance;

	Give users choice	Give AI recommendations only upon request.	Poor human+AI performance

References

Albright, A. (2019). If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. *Harvard John M. Olin Fellow's Discussion Paper 85*. http://www.law.harvard.edu/programs/olin_center/fellows_papers/86_Shadarevian.php

Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 74:1-74:34. https://doi.org/10.1145/3449148

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. (2021, February). Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. *AAAI 2021*. https://www.microsoft.com/en-us/research/publication/is-the-most-accurate-ai-the-best-teammate-optimizing-ai-for-teamwork/

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 19.

Bansal, G., Nushi, B., Kamar, E., Weld, D., Lasecki, W., & Horvitz, E. (2019). A Case for Backward Compatibility for Human-AI Teams. *ArXiv:1906.01148 [Cs, Stat]*. http://arxiv.org/abs/1906.01148

Bansal, G., Tongshuang, W. U., Zhou, J., Raymond, F. O. K., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2020). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery,* New York, NY, USA, Article 81, 1–16. https://doi.org/10.1145/3411764.3445717.

Bentvelzen, M., Niess, J., & Woźniak, P. W. (2021). The Technology-Mediated Reflection Model: Barriers and Assistance in Data-Driven Reflection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445505

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 454–464. https://doi.org/10.1145/3377325.3377498

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). 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), 188:1-188:21. https://doi.org/10.1145/3449287

Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., Peters, A., Jernigan, W. (2016). GenderMag: A Method for Evaluating Software's Gender Inclusiveness. *Interacting with Computers, Volume 28, Issue 6,* 760–787. https://doi.org/10.1093/iwc/iwv046

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *Proceedings of the 2015 International Conference on Healthcare Informatics*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359206

Chiang, C.-W., & Yin, M. (2021). You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. *13th ACM Web Science Conference 2021*, 120–129. https://doi.org/10.1145/3447535.3462487

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376638

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

Efendić, E., Van de Calseyde, P. P. F. M., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes, 157,* 103–114. https://doi.org/10.1016/j.obhdp.2020.01.008

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2021). The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *ArXiv: 2107.13509.*

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *Npj Digital Medicine*, *4*(1), 1–8. https://doi.org/10.1038/s41746-021-00385-9

Green, B. (2021). *The Flaws of Policies Requiring Human Oversight of Government Algorithms* (SSRN Scholarly Paper ID 3921216). Social Science Research Network. https://doi.org/10.2139/ssrn.3921216

Green, B., & Chen, Y. (2019a). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 50:1-50:24. https://doi.org/10.1145/3359152

Green, B., & Chen, Y. (2021). Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. *ArXiv:2012.05370 [Cs]*. https://doi.org/10.1145/3479562

Green, B., & Chen, Y. (2019b). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Translational Psychiatry*, *11*(1), 1–9. https://doi.org/10.1038/s41398-021-01224-x

Kahneman, D. (2011). Thinking, fast and slow. *Macmillan.*

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219

Kim, A., Yang, M., & Zhang, J. (2020). *When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms* (SSRN Scholarly Paper ID 3691575). Social Science Research Network. https://doi.org/10.2139/ssrn.3691575

Koulu, R. (2020). Human Control over Automation: EU Policy and AI Ethics. *European Journal of Legal Studies*, *12*(1), 9–46.

Lai, V., Liu, H., & Tan, C. (2020). "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376873

Lai, V., & Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38. https://doi.org/10.1145/3287560.3287590

Lambe, K. A., O'Reilly, G., Kelly, B.D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety 25,* 10 (2016), 808–820.

Lee, M. K., & Rich, K. (2021). Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445570

Levy, A., Agrawal, M., Satyanarayan, A., & Sontag, D. (2021). Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445522

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445260

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

Lu, Z., & Yin, M. (2021). Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445562

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *ArXiv:1802.00682 [Cs]*. http://arxiv.org/abs/1802.00682

Nourani, M., King, J., & Ragan, E. (2020). The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*, 112–121.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., & Gogate, V. (2021). Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. *26th International Conference on Intelligent User Interfaces*, 340–350. https://doi.org/10.1145/3397481.3450639

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, *15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *ArXiv:1907.12652 [Cs]*. http://arxiv.org/abs/1907.12652

Park, J. S., Barber, R., Kirlik, A., & Karahalios, K. (2019). A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 102:1-102:15. https://doi.org/10.1145/3359204

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. *Human Factors*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–52). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445315

Riveiro, M., & Thill, S. (2021). "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence*, *298*, 103507. https://doi.org/10.1016/j.artint.2021.103507

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 240–251. https://doi.org/10.1145/3301275.3302308

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Springer, A., Hollis, V., & Whittaker, S. (2017, March 20). Dice in the Black Box: User Experiences with an Inscrutable Algorithm. *2017 AAAI Spring Symposium Series*. 2017 AAAI Spring Symposium Series. https://www.aaai.org/ocs/index.php/SSS/SSS17/paper/view/15372

Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. *12th ACM Conference on Web Science*, 315–324. https://doi.org/10.1145/3394231.3397922

Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). Investigating Human + Machine Complementarity for Recidivism Predictions. *ArXiv:1808.09123 [Cs, Stat]*. http://arxiv.org/abs/1808.09123

Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., & Bernstein, A. (2021). Second Chance for a First Impression? Trust Development in Intelligent System Interaction. *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 77–87. https://doi.org/10.1145/3450613.3456817

Vaccaro, M., & Waldo, J. (2019). *The Effects of Mixing Machine Learning and Human Judgment*. https://cacm.acm.org/magazines/2019/11/240386-the-effects-of-mixing-machine-learning-and-human-judgment/fulltext

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021). Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445365

Vodrahalli, K., Gerstenberg, T., & Zou, J. (2021). Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions. *ArXiv:2107.07015 [Cs]*. http://arxiv.org/abs/2107.07015

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *ArXiv*.

Wason, P. C., & Evans, J. (1974). Dual processes in reasoning? *Cognition 3,* 2 (pp. 141–154).

Weisz, J.D., Muller, M., Houde, S., Richards, J., Ross, S.I., Martinez, F., Agarwal, M., & Talamadupula, K. (2021). Perfection Not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 402–412. https://doi.org/10.1145/3397481.3450656

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300509

Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. Tech. rep., Centre for the Governance of AI, University of Oxford. 2 General attitudes toward AI | Artificial Intelligence: American Attitudes and Trends (governanceai.github.io)

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852