

### 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).

Users are more willing to use algorithmic systems when they do not see systems make mistakes (Dietvorst 2015). Never seeing the system err, however, makes users over-trust its capabilities.

Take care not to overwhelm users with information during onboarding (Suresh et al. 2020). Identify ways to progressively onboard users 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, 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, and strengths and limitations help users 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 performance on them. This helps users to know the contexts in which 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 how and when to trust the AI (or not).
  - 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 to accommodate both low/high automation bias users (De-Arteaga et al. 2020; Levy et al. 2021).
2. Devise strategies to assess how *confident* users are in their own abilities (and predictions) and accordingly adjust user experience to help under/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 user experience to help users with low/high AI literacy to develop appropriate reliance (Chiang & 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*. For example, if it is acceptable to sacrifice accuracy on tasks but have a better mental model of AI, show AI strengths first before introducing failure scenarios (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 after they have been acquainted with the system 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).

- 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.

2. *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 and the AI's performance 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 as they cause user overreliance.** User reliance on AI systems is affected by their 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/incorrect AI recommendations (Levy et al. 2021). Examples include recommendations with low confidence scores, those based on limited data, those containing fabricated elements (e.g., AI-generated datasets).

- a. **Train separate models to detect problematic outlier recommendations**—those based on abnormal or insufficient data (Poursabzi-Sangdeh et al. 2021).
- b. **Examine user attitudes towards algorithmic advice prior to 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 as 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 example, analyze interaction effects between different explanation types (e.g., how vs. why) and
  - 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 the AI system. 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 some form of differences and inconsistency in the user experience of the AI system (e.g., aspects of gamification, checking for errors, etc.).



## 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 and the AI's decision-making process. **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 see 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.

For example, instead of providing a universal enable/disable toggle for AI recommendations, identify tasks for which users do not want AI recommendations vs. those in which users are okay with recommendations. Use this to provide users with granular choice—such as, disable AI recommendations for an hour or disable AI recommendations for this task.

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 overreliance on AI

Time	Mitigation technique	Short summary	Issue(s) addressed
During initial interactions	<b>Effectively onboard users</b>	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
	<b>Be transparent with users</b>	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



	<b>Provide personalized adjustments for users</b>	Evaluate user susceptibility (from tutorials and early results) to adjust automation accordingly.	Individual differences; Ordering effects; Poor human+AI performance;
<b>During regular use</b>	<b>Employ cognitive forcing functions</b>	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
	<b>Provide real-time feedback</b>	Real-time feedback on human performance leads to improvement (e.g., alerting 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
	<b>Provide effective explanations</b>	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;

<b>Alter speed of interaction</b>	<p>People trust good models more and bad models less if the system's response time is higher.</p> <p>While waiting, provide user ways to reflect on the task and estimate their own and AI's decision-making process.</p>	<p>Ordering effects; Poor human+AI performance;</p>
<b>Give users choice</b>	<p>Give AI recommendation only upon request.</p>	<p>Poor human+AI performance</p>