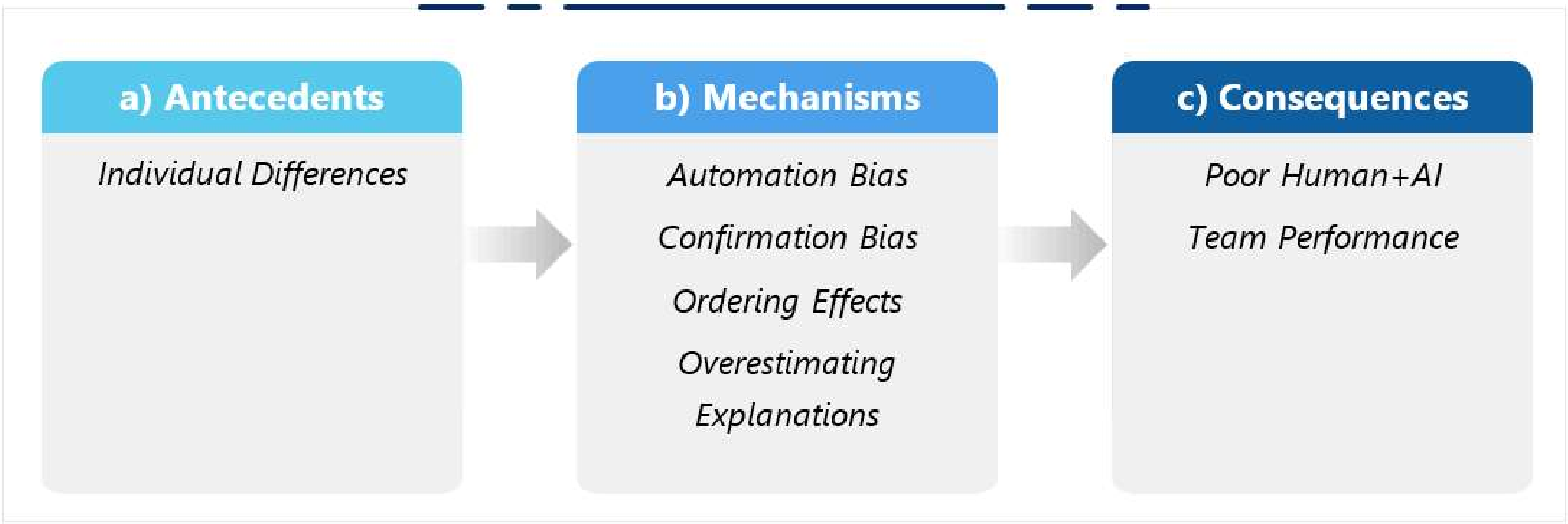
2. Mechanisms and consequences of overreliance on AI

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.



2.1 Antecedents of overreliance on AI

Individual differences

Individual differences lead users to develop both over- and under-reliance on AI.

Individual differences refer to differences in users’ demographic, social, cultural, and professional characteristics.

- **AI literacy:** AI literacy is the measure of how much users know about AI. For more information on AI literacy, including how to measure it, see Long and Magerko (2020).

AI literacy affects users' attitudes towards AI. For more information on user attitudes towards AI, including how to measure them, see Zhang and Dafoe (2019).

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).

- **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.

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 (Gaube et al. 2021).

- **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 makes users 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).

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 close 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

Users with automation bias often over-rely on AI.

Automation bias is the tendency to favor recommendations from automated systems, while disregarding information from non-automated sources.

- **Users with high automation bias are unable to develop appropriate reliance on AI as its performance changes.**

  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 (Pop et al. 2015).

- **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.

💡 **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 (with a singular metric of success) and subjective (with multiple metrics of success) tasks. Keep in mind that user reliance operates on a spectrum (under- to over-reliance) 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.

👍 **Identify ways to assess automation bias from telemetry.**

👍 **Help users to 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 boas leads users to further strengthen the belief they already have about AI.

💡 **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.**

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 (Lu & Yin 2021). Confirmation bias makes users wrongly assume that the AI uses logic and reasoning similar to their own.

💡 **Confirmation bias can lead users who under-rely on AI to further distrust AI.**

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, user trust in AI further deteriorates (Lee & Rich 2021).

- 👍 **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.**

- 💡 **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 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).
- 💡 **User expertise alters the impact of ordering effects.**
  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 (Nourani et al. 2020). When AI fails during early interactions, experts develop under-reliance on AI. The under-reliance never fully goes away even as the AI starts doing better. 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.

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:

Overestimating explanations

- 💡 **Detailed explanations often lead users to develop overreliance on AI.**

2.3 Consequences of overreliance on AI

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:

Summary: Antecedents, mechanisms, and consequences of overreliance

		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