Fostering appropriate reliance on GenAI
Lessons learned from early research

AETHER AI ETHICS AND EFFECTS IN ENGINEERING AND RESEARCH
UXRE

Executive summary

Overreliance on AI is the phenomenon of system users accepting incorrect AI outputs, typically because AI systems make it difficult to spot errors. Overreliance on AI is a barrier to productive human-AI collaboration. Often, people perform tasks worse when using AI than when working alone, or than AI working alone, according to this meta-analysis of 106 experiments. Poor human+AI task performance can lead to costly mistakes (e.g., when medical doctors accept incorrect AI outputs), loss of trust, and product abandonment. Overreliance and its consequential risks are exacerbated by generative AI's (GenAI) impressive but potentially misleading outputs and users' lack of AI literacy.

The goal then, is to mitigate overreliance on AI and to foster appropriate reliance, helping users of GenAI systems accept correct AI outputs and reject incorrect ones. This report summarizes lessons learned about fostering appropriate reliance on GenAI—specifically, applications using retrieval augmented generation (RAG)—from multiple internal research studies and engagements with product teams working on Microsoft Copilot over the past two years. The lessons we learned are of two kinds:

- **UX goals for fostering appropriate reliance:** To foster appropriate reliance, design AI systems to: (1) create useful mental models—help users form realistic mental models of the AI system's capabilities and limitations; (2) signal to users when to verify AI outputs—make it easy to spot mistakes; and (3) facilitate verification—decrease users' cognitive load when verifying AI outputs.
- **Tips for user researchers** to identify and assess overreliance on AI and the effectiveness of mitigations.

According to The International Scientific Report on the Safety of Advanced AI, AI models are probabilistic and will continue to make mistakes. To build reliable user experiences with AI, we need to design for error. Even though fostering appropriate reliance on AI is an open research area that requires ongoing inquiry and innovation, we urge AI builders to develop and test strategies for appropriate reliance.

Authors: Mihaela Vorvoreanu, Samir Passi, Shipi Dhanorkar, Amy Heger, Kathleen Walker

Introduction

Defining overreliance on AI

Overreliance on AI is the phenomenon of AI system users accepting incorrect outputs. Overreliance generally happens when AI system design makes it difficult for users to identify errors in AI outputs. There are many mechanisms that explain how overreliance on AI happens, as seen in an extensive review of literature from different research areas. Consequences of overreliance on AI include mistakes that can lead to severe harms (e.g., when medical doctors accept incorrect AI outputs), loss of trust, and product abandonment.

GenAI increases risk of overreliance and consequential harms

Generative AI (GenAI) refers to AI models that generate text, code, graphics, or audio. GenAI products can quickly create volumes of impressive content and generate different responses for the same prompt, as well as give partially correct responses and fabricate information—including cited sources. The convincing quality of model responses combined with users' lack of AI literacy can lead to overreliance, sometimes with harmful consequences. For example, GenAI can push professionals to make wrong decisions, result in fatally inaccurate identification of mushrooms, give unreliable and biased medical advice, or spread political misinformation.

While there is a long history of research on overreliance on AI—including mechanisms of *how* and *why* users over-rely (e.g., automation bias, confirmation bias)—how to design for appropriate reliance on GenAI is an emerging research area (see our research synthesis).

Appropriate reliance requires UX research

Appropriate reliance—when users accept correct AI outputs and reject incorrect ones—is essential if people are to fully realize the potential benefits and minimize the risks of GenAI. Fostering appropriate reliance for effective human-GenAI collaboration requires different approaches for different contexts and user groups. For example, user expertise matters—novices do not over-rely in the same way as experts. Task type also matters. For example, code writing requires correctness, while story writing involves inventiveness. UX research is essential for testing and prioritizing design approaches for appropriate reliance, based on a system's context, design goals and user needs. There's **no one-size-fits-all strategy for mitigating risks of overreliance.**

Insights from our research

At Microsoft, our group has been focused on research for creating design guidance for appropriate reliance on GenAI. Over the past two years we've examined overreliance on GenAI and the effectiveness of mitigation strategies, through multiple internal user research studies.

In our research studies, we looked at interactions with and perceptions of GenAI and GenAI-powered chatbots such as ChatGPT, Bing Chat, ChatGPT plugins, and Microsoft Copilot. We used a variety of research methods such as in-depth artifact-based interviewing, unmoderated

usability testing, experiments, and analysis of online user-generated content about these technologies.

Overall, we learned that people trust these tools' outputs and are not inclined to verify them. This is explained by a set of misconceptions about how these systems work, what they can and cannot do, and how well. We found that people's mental models of these technologies are similar to Web search, and that people use problematic heuristics to evaluate output trustworthiness – for example, the output's well-written style is considered an indicator of accuracy and trustworthiness. When we explicitly asked users to verify AI outputs by engaging with cited sources, accuracy on information retrieval tasks was low, indicating how current UI paradigms make it difficult to spot errors in AI outputs and to check accuracy and completeness.

Note, we do not claim our research findings generalize to all users and all types of GenAI applications. However, we learned valuable lessons about fostering appropriate reliance on GenAI and evaluating overreliance in research studies. We believe these lessons can be beneficial to builders of GenAI systems, specifically, systems using retrieval augmented generation (RAG). We summarize lessons learned in this report's two sections:

1. **UX goals for appropriate reliance on GenAI:** three main goals for overcoming observed barriers to appropriate reliance.
2. **Tips for user researchers** to identify and assess overreliance on AI and the effectiveness of mitigations.

1 UX goals for appropriate reliance on GenAI: A framework

While there remain many unknowns in this nascent research area, our studies distilled three important UX goals to keep in mind when designing for appropriate reliance on GenAI:

1. **Create useful mental models:** Help users form realistic mental models of the AI's capabilities and limitations.
2. **Signal to users when to verify:** Make it easy to spot mistakes, drawing users' attention when they need to verify outputs more carefully.
3. **Facilitate verification:** Make it easy for users to verify the correctness and completeness of AI-generated content.

1. Create useful mental models

The term *mental model* refers to an abstract or simplified mental representation of how a technology works. A useful mental model can be simple and even inaccurate but must be useful enough to let a person predict what will happen when they use a technology. Useful and realistic mental models are key to users' appropriate reliance on GenAI.

Across our studies, users' incorrect mental models exacerbated overreliance. For example, when interacting with ChatGPT, novice users assumed that GenAI summarization functioned like a search engine that didn't make mistakes. They did not understand the distinction between the model's generative capabilities and search. Participants also missed nuances in GenAI outputs, not realizing these can be wrong in different ways, such as providing partially correct information. Users mistook the well-written style of a ChatGPT response as a signal of substance, or accuracy. When using LLM plugins, users perceived these as "super apps" that would remove the possibility of LLM confabulations.

Recommendations

To help users form useful mental models, consider these two strategies:

- **Be transparent**. Make clear what the system can do and how well it performs (see HAX Guidelines 1 and 2). Provide messaging that informs users of the presence of AI and help them understand AI's role in interaction, as well as the system's capabilities and limitations. (Capabilities can include intended use cases. Limitations include types and frequency of mistakes or model uncertainty.)

  This messaging could be located in first-run experiences, at various entry points to interacting with an AI, during latency, and in uncertainty expressions, tooltips, or disclaimers.

Keep in mind that the mere presence of such messaging is not a guarantee that users understand or even notice it. Assess its effectiveness with **usability testing**.

- **Educate users** about the AI system's workings (see HAX Guideline 11). For example, does easy-to-understand UI messaging let users know that GenAI systems generate content, not merely retrieve it? Or that AI-generated summaries may be incorrect or incomplete? Note that in addition to documentation, educating users through UI messaging can help set realistic expectations about the AI system.

Conduct **user research** to assess the effectiveness of UI messaging and whether it is easily discoverable and appropriately integrated into the user experience.

2. Signal to users when to verify

Useful mental models alone are not enough for facilitating appropriate reliance. We also need to help people spot when something might be wrong, especially with outputs in high-stakes scenarios. UX must encourage vigilance, drawing users' attention when they need to verify outputs more carefully.

However, this is easier said than done. We observed that current GenAI applications' UI paradigms' do not encourage users to verify outputs. Furthermore, attempts to promote user vigilance and avoid overreliance can backfire. We found that the mere presence, variety, and formatting of cited sources added to users' perception of the output's trustworthiness. In fact, participants were skeptical when we pointed out instances of source confabulation, and pushed back with comments such as "I would be very surprised if these [cited sources] weren't real."

Recommendation

Make it easy to spot mistakes. Draw people's attention when they need to verify AI outputs, especially in high-stakes scenarios. Promising directions for UX strategies that encourage users to review, edit, and confirm outputs include:

- Uncertainty expressions (e.g., highlighting tokens that have low output probability; verbal expressions, like "I'm not sure")
- Cognitive forcing functions (e.g., giving users time to think, confirmation dialogues, friction, AI critiques, AI questioning).

Keep in mind that overreliance mitigations can backfire. For example, uncertainty expressions can be a double-edged sword; while first-person uncertainty expressions (e.g., "I'm not sure") can increase user accuracy in a task, they may also lead to lower user confidence in the system and a longer task completion time, according to this study.

It is essential to assess the effectiveness of mitigation strategies with methods such as cognitive walkthroughs and usability testing, asking:

- Do users understand the need to oversee and review AI responses?
- Are there effective ways of alerting users to possible AI mistakes or high-stakes outcomes?
- Are there effective ways of encouraging users to review, edit, and confirm AI outputs before use?

Note: Overreliance risk is higher in automation scenarios where the AI can perform actions without giving the user the opportunity to review them.

3. Facilitate verification

In addition to helping people spot errors or discrepancies between outputs and grounding data, we need to make it easy for users to verify the correctness and completeness of AI-generated content. A useful mental model combined with reminders to be vigilant can encourage users to verify GenAI outputs. But verification can be overwhelming, because it is a difficult and time-consuming task. When verifying outputs is a burden, the risk of overreliance is high.

For example, current UI patterns rely on citing sources that often link to lengthy documents. This burdens the user with the verification task of sifting through the document to find the information used for the AI's output. In studies where we asked people to engage with cited sources, user effort and time on task increased notably with the volume of information, while task performance dropped. As one participant stated, "The amount of manual work it takes to verify is not fair in a way."

Recommendation

Make it easy for users to verify the correctness and completeness of AI-generated content.

Conduct **usability testing** of verification UI. Measure task accuracy, time on task, and satisfaction, asking:

- Are verification aids such as sources and explanations easy to discover and appropriately integrated into the user experience? **Keep in mind overreliance mitigations can backfire.** Explanations can increase user trust even when they are incorrect. The mere presence of sources can make users trust AI outputs more.
- Are verification aids such as sources and explanations effective at helping users verify AI responses?
- Do verification aids have reliability issues (e.g., fabricated sources or inaccurate explanations)?
- Can users with accessibility needs oversee and review AI responses?

2 Tips for UX researchers

As we pointed out in the first section of this report, it is essential to evaluate the effectiveness of overreliance mitigations with user research. Identifying overreliance behaviors specific to a GenAI application is a prerequisite. In this section we share tips for user researchers to identify and assess overreliance on AI and the effectiveness of mitigations.

Defining overreliance risk in your product context: What to watch for

Begin by understanding and defining what overreliance risk looks like in a specific product.

I. Identify problematic user actions

User actions that indicate overreliance differ according to context. For example, in:

- **Summarization,** users often fail to check for factual errors or don't cross-reference summaries with original source material.
- **Information retrieval,** users will often rely solely on GenAI search summaries and not fact-check outputs.
- **Code generation,** over-reliant users will often accept AI-generated code without reviewing for semantic errors or testing for security vulnerabilities.

II. Identify negative consequences of overreliance

Assessing overreliance starts with outlining its impact relevant to your GenAI product and its use cases. Detrimental effects of overreliance can include:

- Increased task completion time.
- Decreased task accuracy.
- Decreased productivity.
- Loss of trust
- Product abandonment

Selecting a research approach

You can either monitor for overreliance by (1) integrating assessment into an existing study, or (2) conducting a dedicated user study focused on overreliance in your GenAI product.

1. How to integrate overreliance assessment into existing user studies

Watch for problematic user behaviors during observational and interview studies:

- **Users making excuses for GenAI mistakes** (e.g., assuming errors are due to faulty source material).
- **Users blaming themselves for GenAI mistakes** (e.g., believing they phrased their prompt incorrectly).
- **Users ignoring contextual information** (e.g., disregarding UI indicators of potential mistakes).
- **Users exhibiting undesirable cognitive biases** (e.g., trusting AI due to automation bias or confirmation bias).
- **Users assuming a GenAI product can identify its own mistakes** (e.g., expecting AI to be able to evaluate itself because it answers user questions about output accuracy, mistakes in outputs, and general system workings.)

Note:

It is important to understand *why* users exhibit problematic attitudes and behaviors. Identifying *what* problematic attitudes and behaviors your users have is just the first step. The key to mitigating overreliance is understanding *why* users default to or exhibit such attitudes and behaviors.

The use of golden paths has limited effectiveness for overreliance assessments. Golden path designs are *unfit* for assessing overreliance risk because they mask mistakes. If users are not exposed to varying levels and types of mistakes in GenAI outputs in user studies, your ability to monitor for such attitudes and behaviors will be limited.

Ask users targeted questions

Use interviews or usability studies to assess overreliance risks, asking questions about:

I. **Mental models**
- What do you think X can/cannot do, and why?
  Helps spot incorrect mental models about system capabilities and limitations.
- How do you think X generated this output?
  Helps spot incorrect mental models about how the system works.
- Does this output seem correct? Why or why not?
  Helps spot the use of incorrect heuristics to assess output correctness.

- What kinds of tasks will you use or avoid using X for, and why?
  Helps spot incorrect mental models about the system's (un)intended use cases.

II. Verification process

- How will you figure out if this output is correct?
  Helps assess users' verification journey, including the heuristics they use.
- Can you easily verify the accuracy of this output?
  Helps assess whether in-built verification methods are easy to find and use.
- Would you check this output differently if you were using X to do a different task?
  Helps identify if and how users' verification behaviors differ by task type.
- Would you verify this output differently based on your familiarity with the topic?
  Helps identify if and how users' verification behaviors differ by topic familiarity.

2. How to conduct dedicated user studies on overreliance

Step 1: Outline your **research goal**

Identify what you want to accomplish with the research study. For example you may want to:

- Learn about user mental models (e.g., *What are novice users' mental models of ChatGPT?*).
- Evaluate effectiveness of overreliance mitigations (e.g., *Do citations mitigate overreliance on AI-generated search results?*).
- Identify human oversight challenges that may exacerbate overreliance (e.g., *What are users' verification experiences in this GenAI product?*).

Note:

Overreliance risk and mitigations can be assessed multiple ways. Select the appropriate methods for your research goal. Examples of research methods used in overreliance studies include observation, think-aloud, interviews, and usability testing. Studies can be moderated or unmoderated. (In our research, we preferred unmoderated usability studies because participants try to engage in desirable behaviors when researchers are present.)

Step 2: Identify appropriate **measures**

Choose appropriate quantitative and/or qualitative measures based on (a) the user behaviors you want to focus on, and/or (b) the overreliance mitigations you want to test.

Example measures to assess overreliance risk and mitigations may be:

- Number of times users accept GenAI outputs containing mistakes (e.g., number of times programmers accept AI-generated code snippets with mistakes).
- Prevalence of incorrect verification heuristics (e.g., the types of incorrect heuristics people use to assess the accuracy of GenAI outputs, such as equating the presence of sources or the formatting of outputs with accuracy).
- User accuracy on tasks, with and without GenAI assistance (e.g., when answering a set of questions).
- User accuracy on GenAI-assisted task, with and without mitigations (e.g., user accuracy when using GenAI to answer a set of questions with and without sources).
- User accuracy on GenAI product-related information (e.g., user accuracy when answering questions about a product's capabilities and limitations).

Note:

Think about what kinds of mistakes you want to focus on. Not all GenAI mistakes are equal. For example, syntactical mistakes in AI-generated code may simply cause the code to not compile, but semantic code mistakes can cause system damage.

When assessing mitigations, note that verification aids such as sources and explanations can have mistakes too. Test for different combinations and types of mistakes in outputs and mitigations.

You must establish ground truth to identify mistakes in GenAI outputs. In use cases such as information retrieval, ground truth is easily available (e.g., correct answers to factual questions). However, for use cases such as summarization, you will need to construct ground truth (e.g., create correct summaries from sets of documents).

Step 3: Select the **target user group**

Account for differences among users. User characteristics affect *how* users over-rely and *to what extent* they are at risk. For example:

- **Low AI literacy increases overreliance risk** (e.g., assuming AI is always correct, automation bias).

- **High overall trust in AI can make mitigations backfire** (e.g., the mere presence of citations can exacerbate overreliance, with users overestimating accuracy and reliability of sources).
- **Lack of domain expertise increases overreliance risk** (Low-expertise users often accept AI outputs at a higher rate than those with high expertise (e.g., novice programmers may accept more AI-generated code than an expert programmer).
- **Lack of domain expertise makes it difficult to verify AI outputs** (e.g., medical students may not have sufficient expertise to verify GenAI-powered diagnoses even when verification aids such as source snippets and explanations are present).
- **Low task familiarity increases overreliance risk (**e.g., a programmer learning a new programming language may lack sufficient syntactical knowledge to catch certain mistakes in AI-generated code).

Note:

Users with high AI literacy, domain expertise, or task familiarity can also over-rely. For example, users with high AI literacy may over-trust explanations of AI outputs; users with high domain expertise may exhibit confirmation bias; and users with high task familiarity may overestimate their ability to catch AI mistakes.

Carefully consider the impact of different user characteristics in the context of your GenAI product and uses.

Step 4: Brainstorm **task design**

Think about what types of tasks you want users to perform in the study because task type affects the nature and extent of overreliance. For instance, tasks can differ in:

- **Variety:** Tasks that require users to have specialized knowledge or expertise increase the risk of overreliance.
  (E.g., finding an answer to a factual question [information retrieval] vs. identifying main points in a documentation [summarization] vs. completing a piece of code [code generation].)
- **Stakes:** The negative impact of overreliance increases with high-stakes tasks.
  (E.g., identifying the main narrative in novels [low stakes], writing code for personal projects [medium stakes], and finding answers to medical or financial questions [high stakes].).
- **Complexity:** Attentional demands and cognitive load of complex tasks increase overreliance risk.

(E.g., finding an answer to a factual question vs. drafting a report that compares and synthesizes an argument using multiple documents.)

Note:

1. **Design tasks that represent real-world usage for your GenAI product.** Representative tasks help assess how users naturally interact with GenAI outputs in realistic situations. Involve subject-matter experts when designing studies in scenarios where you have insufficient domain knowledge (e.g., medical diagnoses, financial advice).

2. **Decide between testing one mistake/mitigation vs. multiple.** For example, when assessing overreliance risk, choose whether to test one or multiple mistakes in a single GenAI output. And when assessing overreliance mitigations, choose whether to test one mitigation or multiple in tandem.

Step 5: Finalize **study design**

Keep the following details in mind when finalizing your study design:

- **Identify appropriate time to ask users targeted questions.** Some questions reveal the study's main objective, altering participant behavior. For example, if your goal is to assess whether participants use sources for verification, asking participants questions about sources after each task will make them overly engage with sources, leading to incorrect assessment of overreliance mitigations.
- **Watch out for unintended consequences of overreliance mitigations.** Overreliance mitigations can backfire (e.g., the very presence of sources and explanations can exacerbate overreliance).

Note:

Don't overly incentivize participants to look for mistakes in GenAI outputs. Incentives such as extra money for correct answers can change participant behavior. Strive to put users in situations closer to their natural state for effective assessments.

Conclusion

In this report, we summarized lessons learned from our work on fostering appropriate reliance on AI. We derived three UX goals for fostering appropriate reliance on AI from the barriers to appropriate reliance observed in multiple studies. These three UX goals inform the Overreliance Risk Identification and

Mitigation Framework, which guides AI builders through the same process we used to tackle overreliance on AI in various products.

Overreliance on AI is a complex phenomenon, which we explained in our previous syntheses of research literature. Existing research makes clear that overreliance mitigations can backfire. Therefore, it is essential to test their effectiveness with user research. In this report, we also share tips for user researchers to identify and assess overreliance on AI and the effectiveness of mitigations.

Fostering appropriate reliance on AI is an open area of research. There is much more we need to learn as a community.

One limitation of the approach presented in this report is that it works within the dominant UI paradigm of GenAI application—chatbots. There are promising research directions that interrogate this paradigm: for example, positioning LLMs as provocateurs that foster critical thinking, or agents that facilitate sensemaking, rather than assistants serving ready-made answers.

References

UK Department for Science, Innovation and Technology & AI Safety Institute. 2025. The International Scientific Report on the Safety of Advanced AI. DSIT paper series number 2025/001. https://www.gov.uk/government/publications/international-ai-safety-report-2025

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 822–835. https://doi.org/10.1145/3630106.3658941

Samir Passi & Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review*. Microsoft Technical Report MSR-TR-2022-12. Microsoft Corporation. https://aka.ms/overreliance_review

Samir Passi, Shipi Dhanorkar, & Mihaela Vorvoreanu. 2024. *Appropriate Reliance on Generative AI: Research Synthesis*. Microsoft Technical Report MSR-TR-2024-7. Microsoft Corporation. https://aka.ms/genai_reliance

Microsoft. 2025. Overreliance Risk Identification and Mitigation Framework. https://aka.ms/overreliance-framework

Advait Sarkar. 2024. AI Should Challenge, Not Obey. Commun. ACM 67, 10 (October 2024), 18–21. https://doi.org/10.1145/3649404

S. Y. -T. Lee and K. -L. Ma, "HINTs: Sensemaking on large collections of documents with Hypergraph visualization and INTelligent agents," in IEEE Transactions on Visualization and Computer Graphics, doi: 10.1109/TVCG.2024.3459961