Project Title:Designing Controlled Exposure for Resilient Human–AI Organizations

Project ID: S26_7
Team Members: Tianyi Ma, Yifan Tao, Joe Zou, Haomiao Shi

1. Project Understanding & Background Research

Problem Area

In many contemporary AI systems, seamless automation has become a dominant design goal. Interfaces are designed to reduce visible complexity and deliver outputs efficiently. While this approach improves usability and speed, it often conceals important aspects of system behavior, such as model uncertainty, internal reasoning processes, and data handling mechanisms.

As AI tools become embedded in everyday academic and professional workflows, this invisibility introduces a significant interaction challenge. When users cannot see how confident a system is or how information is processed, they may rely on outputs without fully understanding their limitations. Over time, this can reduce critical evaluation and weaken human oversight.

The issue is not simply about whether AI systems are accurate. Even advanced models can produce hallucinations, incomplete reasoning, or context-specific errors. The deeper concern lies in how these limitations are communicated to users, and whether the interface supports informed judgment or passive acceptance.

Why This Problem Matters

Human–AI interaction increasingly influences decision-making in research, education, and technical environments. When uncertainty is not clearly communicated, users may overestimate the reliability of AI-generated content. Small inaccuracies can go unnoticed and accumulate into larger problems, particularly in high-stakes contexts where decisions are made quickly.

Privacy and governance present additional concerns. Many AI systems process sensitive information in ways that are not visible to users. Data masking, filtering, or transformation often occurs in the background without explicit feedback. As a result, users may have limited awareness of how information is handled or whether compliance requirements are met.

When transparency decreases and human engagement becomes more passive, organizations may gradually lose the ability to question, monitor, and intervene in automated processes. This can increase vulnerability to trust failures, privacy breaches, and systemic errors.

Who Is Affected

This problem affects a range of stakeholders across academic and organizational settings.

Primary users include students, researchers, engineers, and other knowledge workers who rely on AI systems for summarization, analysis, drafting, and decision support. These individuals often interact with AI tools under time constraints and may incorporate outputs directly into academic or professional work.

Secondary stakeholders include instructors, supervisors, and organizational leaders who depend on the accuracy and responsible use of AI-assisted content. At a broader level, institutions and companies are also affected when misinformation, compliance violations, or governance breakdowns result in reputational or legal consequences.

Situations in Which the Problem Occurs

These issues commonly arise when AI tools are integrated into routine workflows. Examples include summarizing technical research documents, analyzing datasets that contain sensitive information, generating reports, and supporting decisions under time pressure.

In such situations, seamless interfaces may obscure uncertainty, system limitations, and data governance processes. As a result, errors or privacy risks may remain unnoticed until they create significant downstream consequences.

Research Expectations:

Automation Bias and Over-Reliance

Research in human factors has long documented the phenomenon of automation bias, in which individuals over-rely on automated systems and reduce their own vigilance (Parasuraman & Riley, 1997)[1]. When systems operate smoothly and appear reliable, users are more likely to accept outputs without verification. Studies show that increased automation can gradually shift humans from active decision-makers to passive monitors.

These findings suggest that seamless system behavior is not neutral. It changes how responsibility and attention are distributed between humans and machines.

Algorithmic Opacity and Transparency

Scholars have argued that transparency in AI systems is often misunderstood as simply providing explanations for outputs. Ananny and Crawford (2018)[2] note that algorithmic systems remain structurally opaque even when explanations are offered. Visibility of results does not necessarily translate into accountability.

This literature highlights that interaction-level opacity can limit users' ability to assess system reliability or governance processes.

Trust Calibration in Human–AI Collaboration

Recent work on human–AI teaming emphasizes the importance of trust calibration rather than blind trust. Bansal et al. (2021)[3] demonstrate that performance improves when humans understand when to rely on AI and when to question it. Both over-trust and under-trust can degrade collaborative outcomes.

This research suggests that uncertainty communication plays a critical role in supporting responsible AI use.

AI Brittleness and Distribution Shift

AI safety research identifies brittleness and distribution shift as key risks in deployed systems (Amodei et al., 2016)[4]. Models that perform well in training conditions may fail unpredictably in real-world contexts. These failures are often subtle and may not be immediately apparent to users.

This highlights the need to consider how interaction design influences the detectability of AI errors.

Organizational Governance and Risk Frameworks

Industry frameworks such as the NIST AI Risk Management Framework and Microsoft's Responsible AI guidelines[5] emphasize governance, monitoring, and compliance. However, these frameworks largely focus on policy and institutional controls rather than everyday user interaction.

This suggests a gap between governance theory and interaction practice, particularly in routine AI-assisted workflows.

Synthesis

Across human factors research, algorithmic accountability studies, and AI safety literature, a consistent concern emerges: AI systems can unintentionally diminish human oversight when uncertainty, reasoning processes, and data handling remain opaque. Automation bias research shows that seamless systems may reduce vigilance. Transparency scholarship highlights structural limits of explanation-based approaches. Trust calibration studies emphasize the delicate balance between reliance and skepticism in human–AI collaboration. Meanwhile, governance frameworks primarily operate at the policy level rather than at the level of everyday interaction.

Together, these bodies of work suggest that the risks associated with AI are not solely technical failures, but interactional and organizational dynamics shaped by how systems present themselves to users. While existing literature identifies automation bias, opacity, and brittleness as critical concerns, less attention has been given to how routine interface design decisions influence users' ability to detect uncertainty, question outputs, and intervene when necessary.

As AI systems become embedded in academic and professional workflows, understanding how human oversight erodes through interaction design becomes an important research problem. This project situates itself within that problem space by examining the relationship between interface transparency, trust calibration, and organizational resilience in AI-supported environments.

---

2. Market & Competitor Analysis (No Feature Pitching）

To better understand the current landscape of AI-supported systems and governance tools, this section examines several existing products that address aspects of AI assistance, monitoring, or accountability. The goal is to analyze how these systems approach transparency, oversight, and user interaction.

2.1 ChatGPT and Large Language Model Interfaces

Commercial AI assistants such as ChatGPT and Claude are designed to provide conversational support for tasks including summarization, drafting, coding assistance, and research support. These systems prioritize simplicity and seamless interaction. Users can input prompts and receive immediate responses without being exposed to internal reasoning processes or uncertainty metrics.

These tools are primarily designed for general users, including students, professionals, and developers. Their strength lies in accessibility and speed of interaction.

From a user-experience perspective, however, internal model confidence and data processing steps are not clearly visible. While disclaimers may acknowledge potential inaccuracies, uncertainty is not dynamically communicated in a structured way. Additionally, privacy handling mechanisms remain largely abstracted from everyday interaction. Users must trust that data is processed appropriately without receiving granular feedback.

This reflects a broader trend toward frictionless AI interaction, where ease of use may come at the cost of deeper transparency.

2.2 Microsoft Copilot and Enterprise AI Assistants

Microsoft Copilot integrates AI assistance directly into enterprise productivity tools such as Word, Excel, and Teams. It aims to enhance workflow efficiency by embedding AI into existing professional environments. These systems are designed for organizational users and knowledge workers operating within corporate infrastructures.

Copilot emphasizes productivity and integration with secure enterprise systems. However, its interface largely presents AI-generated outputs as suggestions without detailed visibility into reasoning processes or confidence indicators. While enterprise-level governance controls may exist in backend administrative dashboards, everyday users interact primarily with polished outputs rather than system-level signals.

From a UX standpoint, governance and oversight mechanisms appear to be separated from the moment of interaction. Users may benefit from AI augmentation but may not have direct visibility into how risk monitoring or model behavior is managed in real time.

2.3 IBM Watson OpenScale

IBM Watson OpenScale is designed specifically for monitoring AI models in production environments. It provides tools for bias detection, model drift monitoring, and performance auditing. The target audience includes data scientists, AI engineers, and compliance officers responsible for managing deployed models.

Unlike conversational AI tools, OpenScale focuses explicitly on governance and model evaluation. It provides dashboards for tracking fairness metrics and performance changes over time.

However, from a user-experience perspective, these governance tools are typically accessed by specialized technical roles rather than everyday end users. The interaction occurs at a monitoring layer rather than within routine AI-assisted workflows. As a result, governance visibility may not extend directly to non-technical users who interact with AI outputs on a daily basis.

2.4 Summary of Observed Gaps

Across these systems, a pattern emerges. Conversational AI tools prioritize seamless usability, often minimizing visible system complexity. Enterprise governance tools focus on monitoring and compliance but are typically separated from everyday interaction.

This suggests that transparency and oversight mechanisms are often either abstracted away from users or concentrated in specialized administrative dashboards. The integration of governance signals directly into routine AI interaction appears limited within current systems.

---

3. Stakeholder Identification

AI-supported systems embedded in academic and organizational workflows affect multiple groups with different roles, expectations, and constraints. The following stakeholder analysis outlines primary users, secondary users, and indirect stakeholders who may be involved in or impacted by reduced transparency and oversight in AI-assisted environments.

3.1 Primary Users

1. Students and Graduate Researchers

Role:
Students and researchers use AI systems for summarization, drafting, data analysis, and research assistance.

Goals:
They aim to increase productivity, understand complex materials more efficiently, and meet academic deadlines.

Concerns and Constraints:
They may operate under time pressure and limited technical visibility into how AI outputs are generated. If uncertainty signals or limitations are unclear, they may unknowingly rely on inaccurate or incomplete outputs. At the same time, they must adhere to academic integrity policies and data privacy expectations.

---

2. Engineers and Knowledge Workers

Role:
Professionals use AI systems for drafting reports, reviewing technical documents, and supporting decision-making.

Goals:
They seek efficiency, accuracy, and workflow integration within existing systems.

Concerns and Constraints:
Errors or hallucinations may have downstream consequences in professional contexts. Workers may not have access to detailed governance signals and may rely on AI outputs without fully understanding associated risks. Organizational expectations for speed can reduce opportunities for critical verification.

3.2 Secondary Users

1. Instructors and Academic Supervisors

Role:
They oversee student work and evaluate the responsible use of AI tools in academic contexts.

Goals:
They aim to ensure academic integrity, maintain quality standards, and prevent misuse of AI systems.

Concerns and Constraints:
They may have limited visibility into how AI tools are used by students. Over-reliance on AI outputs can create challenges in assessment and attribution of work. They must balance encouraging technological literacy with preventing uncritical dependence.

---

2. Team Leads and Managers

Role:
They supervise teams that incorporate AI tools into professional workflows.

Goals:
They aim to maintain productivity, reduce risk, and ensure compliance with internal policies.

Concerns and Constraints:
Managers may rely on aggregated outputs without understanding how AI uncertainty or privacy mechanisms function in daily interactions. They are accountable for outcomes even if they are not directly involved in AI interactions.

3.3 Indirect Stakeholders

1. Institutions and Organizations

Role:
Universities and companies provide infrastructure, policies, and governance frameworks for AI usage.

Goals:
They aim to protect institutional reputation, ensure compliance with legal standards, and promote responsible innovation.

Concerns and Constraints:
Failures in AI oversight can result in reputational damage, regulatory penalties, or loss of trust. Institutions must balance innovation with risk management.

---

2. Compliance Officers and Data Protection Authorities

Role:
They establish and monitor policies related to data privacy, AI governance, and risk mitigation.

Goals:
They seek transparency, auditability, and adherence to legal frameworks such as data protection regulations.

Concerns and Constraints:
They often operate at a policy level rather than at the level of daily interaction. A

disconnect between policy and actual user behavior may create hidden compliance gaps.

---

3. Broader Public and End Beneficiaries

Role:
The public may indirectly rely on outputs generated with AI assistance, especially in research, reporting, or professional decision-making contexts.

Goals:
They expect accuracy, fairness, and responsible data handling.

Concerns and Constraints:
They are affected by downstream consequences of misinformation, privacy breaches, or systemic AI failures but have limited visibility into how AI systems are used internally.

4. Design Thinking Milestones

This project will follow the five stages of the Design Thinking model to systematically explore the problem space and inform future design decisions. At this stage, the team is defining planned activities and anticipated outputs rather than executing them.

4.1 Empathize

Planned Activities

The team plans to conduct semi-structured interviews with graduate students, researchers, and knowledge workers who regularly use AI tools for summarization, drafting, or analysis. We also plan to conduct contextual inquiry to understand how AI systems are integrated into academic and technical workflows.

Expected Artifacts

Interview transcripts, observation notes, and affinity diagrams organizing themes related to trust, uncertainty, workflow integration, and governance awareness.

Key Questions

- How do users currently evaluate the reliability of AI-generated outputs?
- At what moments do they verify or question AI responses?
- Where do they experience over-reliance or hesitation?
- How aware are they of data-handling and privacy implications?

4.2 Define

Planned Activities

Insights gathered during the Empathize phase will be synthesized to identify recurring patterns, breakdowns, and tensions in human–AI interaction. The team will articulate structured problem statements reflecting observed challenges.

Expected Artifacts

User personas, interaction journey maps, and clearly defined problem statements grounded in research findings.

Key Questions

- Where does oversight weaken in AI-assisted workflows?
- What interaction moments contribute to trust miscalibration?
- How do governance concerns manifest in everyday use?

4.3 Ideate

Planned Activities

The team will conduct structured brainstorming sessions using techniques such as “How Might We” prompts and concept clustering to explore multiple directions for addressing identified challenges.

Expected Artifacts

Concept maps, idea clusters, and prioritized themes based on feasibility and alignment with user needs.

Key Questions

- What interaction strategies could support more calibrated trust?
- How might transparency be introduced without overwhelming users?
- What feedback mechanisms could support reflective decision-making?

4.4 Prototype

Planned Activities

The team plans to create low-fidelity representations (e.g., sketches, wireframes) to explore interaction concepts identified during the Ideation phase. These artifacts will help clarify potential workflows and user decision points.

Expected Artifacts

Annotated sketches and structured wireframes illustrating key interaction pathways.

Key Questions

- How might changes in visibility affect user interpretation?
- Does additional information introduce cognitive burden?
- Where do users hesitate or require support?

4.5 Test

Planned Activities

The team plans to conduct usability evaluation sessions with participants from the target user group. Participants will complete structured tasks while thinking aloud to reveal reasoning processes and potential breakdowns.

Expected Artifacts

Planned data collection will include qualitative observation notes, usability metrics, and participant reflections to inform future iteration.

Key Questions

- Do users demonstrate improved critical evaluation of AI outputs?
- Are uncertainty or privacy signals noticed and interpreted correctly?
- Does the interaction design support timely intervention?

---

5. Technical Foundation and Implementation Planning

This project involves an interactive system that supports AI-assisted workflows while making system behavior observable to users. Although the team is not implementing the system in Phase 1, this section outlines the expected technical scope and the conceptual technology choices needed to support later design and evaluation work.

5.1 Front-end Considerations

The front-end will need to support interactive oversight tasks in real time. Key considerations include how users will view AI activity as it happens, how the interface will present uncertainty and data-handling information without overwhelming users, and how the system will capture user actions during oversight tasks. The interface should also support consistent interaction flows for monitoring and intervention, while remaining responsive across typical desktop environments used in academic and technical settings.

At a conceptual level, a modern component-based web framework such as React is suitable because it supports modular UI development, state-driven rendering, and rapid

iteration of interaction flows. This is important for later phases where the team will refine prototypes based on user feedback and usability testing.

5.2 Back-end Considerations

The back-end must support controlled execution of AI-assisted actions and expose system events in a way that can be observed by the front-end. High-level requirements include receiving user requests, coordinating AI-related operations, applying governance checks, and emitting structured logs suitable for interface display and later analysis.

A lightweight Python service is an appropriate conceptual choice because it supports rapid prototyping, strong ecosystem support for AI-related tooling, and straightforward integration with web APIs. The back-end should also support a streaming mechanism so the interface can display system activity as it occurs, rather than relying only on static responses.

5.3 Data and AI Components

The project's AI component is expected to involve generating or summarizing text based on technical documents, including scenarios where errors or uncertainty may occur. The data component includes managing representative documents for study tasks and ensuring that sensitive fields can be handled in a controlled and auditable way.

From a planning perspective, the team will need to define how AI outputs are represented in a structured format, including metadata that can support user oversight. This includes fields such as confidence or uncertainty indicators, provenance or context references when available, and markers related to data sensitivity. The project will also require a curated dataset for evaluation tasks, including deliberately challenging cases that test users' ability to notice inaccuracies or privacy risks.

5.4 Conceptual Technology Choices

At a conceptual level, the project scope suggests the following technology categories:

Front-end: A React-based web interface to support modular interaction design and rapid iteration of user-facing components.

Back-end: A Python-based service to orchestrate AI-related operations, apply policy checks, and emit observable system events.

Communication: A real-time streaming approach, such as server-sent events, to support continuous updates between the back-end and the interface during oversight tasks.

Deployment: A container-based deployment model is appropriate for reproducibility and for enabling consistent setups across team members and study environments. AWS or Google Cloud may be a good option for our application.

These choices are presented at a high level to demonstrate readiness and feasibility without committing to implementation details at this phase.

5.5 Rationale and Alignment with Later Design Work

This technical foundation supports later design phases in three ways. First, it enables iterative UI development, which is essential for prototyping and testing interaction patterns related to oversight and transparency. Second, a structured back-end with event streaming supports the collection and presentation of interaction-relevant system activity, which will be needed for usability testing and research evaluation. Third, defining data and metadata requirements early ensures that study tasks can be designed around realistic workflows while remaining ethically and practically manageable in a classroom setting.

Overall, the technical plan is intended to ensure that later design work can be grounded in realistic system behavior and that the evaluation can measure user understanding and oversight actions in a controlled way.

Reference List

[1] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

[2] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645

[3] Bansal, G., Wu, T., Zhou, J., Fok, R., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. https://doi.org/10.1145/3411764.3445717

[4] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*. https://arxiv.org/abs/1606.06565

[5] National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework