DOCTORAL DISSERTATION DRAFT • HUMAN-COMPUTER INTERACTION

Prompt-Mediated Agency: Individual Sovereignty and Self-Actualization in Generative Interfaces

A Mixed-Methods Investigation of Cognitive Offloading, Metacognitive Scaffolding, and Provenance-Bound Interaction

Abstract

Generative artificial intelligence interfaces increasingly mediate human reasoning, decision-making, and long-term goal pursuit through natural-language prompts. Prevailing Human-Computer Interaction (HCI) paradigms prioritize frictionless task completion and fluent synthesis, yet these design choices can encourage cognitive offloading, epistemic passivity, and algorithmic heteronomy. This dissertation investigates how prompt structure and interface affordances govern **individual sovereignty**—defined here as epistemic and decisional autonomy—and **self-actualization**, operationalized through Self-Determination Theory as sustained competence growth, autonomous endorsement, and authentic authorship. Drawing on metacognitive scaffolding, dual-process reasoning, and provenance-oriented systems design, this work formulates a four-part taxonomy of prompt postures and proposes *Sovereign Prompting Interfaces* that require explicit intent specification, counterfactual examination, and verifiable provenance separation before commitment. A mixed-methods research program is outlined to evaluate effects on subjective workload, metacognitive calibration, unassisted skill transfer, resistance to sycophancy, and longitudinal perceptions of authorship.

1. Introduction and Problem Statement

Natural-language prompting has become a primary cognitive and socio-technical boundary between human intention and automated execution. Unlike deterministic graphical user interfaces or explicit command pipelines, where system states and transitions are inspectable, generative chat interfaces often obscure model priors, optimization objectives, and epistemic uncertainty behind authoritative prose.

Empirical studies of AI-assisted work point to an *effort–time dissociation* and a *speedup illusion*: generative tools frequently reduce perceived mental effort even when overall task duration remains comparable or shifts toward verification and revision **Bastani et al. (2024)**; **Kosmyna et al. (2025)**. When individuals rely on open-ended, delegative prompts without structured verification or constraint-setting, three interrelated risks arise:

- **Cognitive Offloading and Transfer Loss:** Bypassing productive analytical struggle can weaken schema formation and reduce performance when individuals later work without model assistance **Risko and Gilbert (2016)**; **Bastani et al. (2024)**.
- **Epistemic and Value Heteronomy:** Model sycophancy, persuasive framing, and unexamined system defaults can subtly shift a user's judgments and priorities toward model tendencies rather than self-authored criteria **Perez et al. (2022)**; **Sharma et al. (2024)**.

- **Collapsed Provenance:** Unstructured conversational histories often blend user intent, model inference, and unverified claims into a single stream, making auditability and independent verification difficult.

This dissertation examines how HCI can re-design prompt workflows and interface guardrails so that generative systems support individual sovereignty and self-actualization rather than replacing human judgment.

2. Literature Review and Conceptual Foundations

2.1 Self-Actualization and Self-Determination Theory

In humanistic psychology, self-actualization denotes the realization of personal potential, autonomy, and purposeful growth **Maslow (1943)**. Within contemporary motivation research, Self-Determination Theory (SDT) provides an empirically testable framework for flourishing through three basic psychological needs: **autonomy** (acting with volition and self-endorsement), **competence** (experiencing mastery through optimal challenge), and **relatedness** **Ryan and Deci (2000)**. For HCI, SDT implies that interfaces supporting self-actualization must preserve volitional control and calibrated challenge rather than eliminating all cognitive friction.

2.2 Cognitive Offloading, Metacognition, and Productive Struggle

Cognitive offloading refers to the use of external tools or actions to reduce internal cognitive demand **Risko and Gilbert (2016)**. While offloading routine memory or calculation can free executive capacity for higher-order reasoning, offloading core synthesis or critical evaluation carries metacognitive hazards. Learners and practitioners may overestimate their mastery when fluent outputs are readily generated, leading to miscalibrated confidence and reduced transfer to unassisted settings **Bastani et al. (2024)**. Metacognitive scaffolding in interactive systems—such as prompting learners to articulate assumptions, justify steps, or test counterexamples—helps counter premature closure **Azevedo (2005)**.

2.3 Individual Sovereignty, Sycophancy, and Provenance

In socio-technical and political theory, sovereignty concerns authority over boundaries, rules, and self-governance without external domination. Applied to personal AI use, *individual sovereignty* entails maintaining independent evaluative standards, inspecting evidentiary grounding, and retaining clear refusal or revision control. Research on large language models highlights two pressures that challenge this stance: sycophantic alignment with user phrasing even when flawed **Perez et al. (2022)**; **Sharma et al. (2024)**, and persuasive fluency that can induce over-reliance. Provenance-oriented design addresses opacity by recording origin, transformations, and verification status across information workflows **Buneman et al. (2001)**.

3. Theoretical Framework and Prompt Posture Taxonomy

To analyze how prompt structure shapes human agency, this dissertation defines three operational constructs:

- **Individual Sovereignty (Cognitive and Epistemic):** The user's capacity to set independent evaluation criteria, distinguish verified claims from model inference, and exercise explicit refusal or override control over system proposals.
- **Self-Actualization (Processual Trajectory):** Ongoing development of competence, self-endorsed decision-making, and traceable authorship across meaningful tasks, grounded in SDT autonomy and competence constructs **Ryan and Deci (2000)**.
- **Sovereign Prompting Interface:** An interaction design pattern that requires users to declare intent, constraints, or verification rules before or alongside generation, separating drafting support from human commitment.

Prompt Posture	Interaction Pattern	Primary Agency and Cognitive Profile
1. Delegative	Requests a finished answer, draft, or recommendation with minimal user constraints.	High cognitive offloading; elevated risk of uncritical adoption, sycophancy effects, and lower unassisted transfer.
2. Transactional	Iterative back-and-forth edits aimed at quick output polish without explicit criteria.	Moderate user steering; susceptible to incremental drift from original goals and local rather than global evaluation.
3. Socratic / Adversarial	Requires the model to surface assumptions, alternative framings, and counterarguments first.	Promotes metacognitive monitoring and critical scrutiny at the cost of higher upfront interaction steps.
4. Provenance-Bound	Separates facts, inferences, and user commitments while requiring explicit human sign-off or refusal.	Supports epistemic auditability, clear authorship boundaries, and deliberate decision ownership.

4. Research Questions and Empirical Methodology

This dissertation is structured around three research questions linking prompt interface design to cognitive and motivational outcomes:

- **RQ1 (Metacognition and Transfer):** How do Socratic and Provenance-Bound prompt structures influence perceived workload, confidence calibration, and performance on subsequent unassisted tasks compared with Delegative prompting?
- **RQ2 (Epistemic Vigilance):** How do explicit pre-commitment constraints and provenance separation affect detection of planted inaccuracies, unsupported leaps, or framing bias in model outputs?
- **RQ3 (Autonomy and Longitudinal Authorship):** How do structured interface frictions—such as assumption checks and explicit refusal or edit gates—shape perceived autonomy, competence, and authorship over multi-week use?

4.1 Proposed Study Program

Study 1 (Controlled Within-Subjects Experiment): Participants complete analytical synthesis and decision-tradeoff tasks across three counterbalanced conditions—*Delegative Baseline*, *Socratic Scaffold*, and *Provenance-Bound Workspace*—followed by an unassisted transfer probe. Core measures include subjective workload (NASA-TLX) **Hart and Staveland (1988)**, confidence calibration accuracy, error/bias detection rates, and transfer task scores.

Study 2 (Longitudinal Field Study): Participants use an instrumented prompting workspace across real writing, analysis, or design tasks over several weeks. Data sources combine weekly autonomy and competence survey items informed by SDT, interaction logs (constraint completeness, edits, refusals), and semi-structured interviews on perceived authorship and sovereignty.

5. Expected Contributions and Ethical Considerations

This work aims to contribute (1) a conceptual bridge between SDT, metacognitive scaffolding, and prompt interface design; (2) empirical evidence on the trade-offs between frictionless generation, epistemic vigilance, and skill retention; and (3) open interface patterns for provenance-aware prompting. Ethical safeguards in human-subjects studies will emphasize transparent disclosure of synthetic or planted test items, mitigation of participant fatigue during multi-condition tasks, and privacy-preserving logging of interaction artifacts.

References

1. Azevedo, R. (2005). Computer environments as metacognitive tools for enhancing learning. *Educational Psychologist*, 40(4), 193–197.
2. Bastani, H., Bastani, O., Sungu, A., et al. (2024). Generative AI can harm learning. *The Wharton School Research Paper*.
3. Buneman, P., Khanna, S., & Tan, W. C. (2001). Why and where: A characterization of data provenance. In *Database Theory — ICDT 2001* (pp. 316–330). Springer.
4. Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139–183.
5. Kosmyna, N., Hauptmann, E., Yuan, Y. T., et al. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint*.
6. Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50(4), 370–396.
7. Perez, E., Ringer, S., Lukošiūtė, K., et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint*.
8. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688.
9. Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78.
10. Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*.