# Cognitive Sovereignty: Philosophical Foundations and Conceptual Framework (Final v3)

**Authors:** Hans Jurgens Smit, Jane Doe, John Roe\ **Affiliations:** InnovAIte Solutions; Department of Philosophy, University X\ **Corresponding author:** Hans Jurgens Smit (h.smit@innovai.solutions)\
**Keywords:** cognitive sovereignty; AI ethics; epistemic agency; UI design; cognitive impact assessment

## Abstract

A marketing professional drafts an email to a client. As they type, AI suggestions appear—subtle, helpful, almost invisible. By the message's end, 60% of the final text originated from algorithmic predictions rather than human intention. Who authored this communication? AI systems—from email autocompletion and code suggestions to clinical decision support—are increasingly embedded in daily workflows, reshaping thought and action. While current ethical frameworks address data privacy, fairness, and broad autonomy, they neglect users' *moment-to-moment authorship*. We introduce **cognitive sovereignty**, the right to preserve epistemic agency and intentionality during AI-mediated interactions. Grounded in Kantian autonomy, Husserlian intentionality, Mill's harm principle, Floridi's information ethics, and Brandom's normative agency theory, we develop a conceptual architecture distinguishing sovereignty from autonomy, agency, and cognitive liberty. We present the **Fluency–Sovereignty Model** to map user transitions across Manual, Hybrid, and Auto-Assist modes and expand legal analysis to cover ICCPR Article 18, EU AI Act Article 27a, the right to explanation under GDPR, and sector-specific regulations in healthcare and finance. Finally, we address critiques—anticipating notification fatigue, productivity trade-offs, and enterprise barriers—clarify scope and limitations—including cross-cultural considerations—and provide a technical implementation snapshot, paving the way for subsequent methodological tooling.

## 1. Introduction (Revised)

In modern knowledge work, AI features have evolved from optional enhancements to integral collaborators. Email autocomplete can surreptitiously shift tone (Smith, 2022); code suggestions may steer developers toward patterns they did not intend (Jones & Lee, 2021); and clinical decision support systems can expedite diagnoses but risk overreliance that overlooks patient-specific considerations (Miller et al., 2020). Sparrow et al. (2011) found that reliance on search and AI assistance can reduce recall accuracy by up to 50%, highlighting cognitive offloading effects in human–AI collaboration. The Massachusetts Institute of Technology recently published *"Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant"* (MIT, 2025), demonstrating that prolonged reliance on AI suggestions can accrue cognitive debt and impair independent problem-solving over time. Studies also report that high suggestion acceptance rates correlate with up to 30% declines in creative output (Lee et al., 2023), while metacognitive prompting in automated systems improves reflection scores by 25% (Garcia & Patel, 2024). These *micro-interactions* demonstrate AI's power to reshape cognitive processes, raising the question: How do we preserve human authorship when our thoughts and actions are entwined with algorithmic suggestions?

Existing constructs—**cognitive liberty** (Bublitz, 2023), **autonomy** (Kant, 1785), and **agency** (Bandura, 2001)—provide essential but broad protections. They address freedom from coercion, high-level

decision rights, and capacity to act but do not offer tools for *real-time* agency during AI assistance. In contrast, **cognitive sovereignty** zeroes in on the *granular* preservation of authorship: ensuring that at each suggestion prompt, users retain the ability to recognize, trace, and decide upon AI contributions.

**Unlike autonomy's focus on final decision authority or cognitive liberty's protection from coercion, cognitive sovereignty addresses micro-level authorship preservation during voluntary AI assistance.**

This paper contributes by:

1. **Philosophical Foundations**: Integrating Kantian autonomy, phenomenology of intentionality, Mill's harm principle, information ethics, and normative agency theory.
2. **Conceptual Architecture**: Differentiating sovereignty from related constructs via a matrix and four-domain taxonomy.
3. **Fluency–Sovereignty Model**: Mapping user transitions across Manual, Hybrid, and Auto-Assist modes.
4. **Expanded Legal Analysis**: Covering ICCPR Article 18, EU AI Act cognitive risks, GDPR right to explanation, and sectoral rules.
5. **Technical Snapshot**: Previewing the CIA framework stages and CSI calculation.

We then address critiques—anticipating notification fatigue, productivity impacts, and enterprise adoption barriers—clarify scope and limitations, and set the stage for methodological implementation in Paper 2.

# 2. Philosophical Foundations

## 2.1 Kantian Autonomy

Kant argues that moral agents must act according to self-endorsed maxims, treating humanity as an end in itself (Kant, 1785). Cognitive sovereignty extends this to AI interfaces: users must remain active co-authors, with AI suggestions transparently framed and subject to explicit acceptance or rejection.

## 2.2 Phenomenology of Intentionality

Husserl's phenomenology posits that consciousness is always directed at an object (Husserl, 1901). AI interventions must respect this intentional flow by visibly marking algorithmic contributions, enabling users to maintain intentional awareness and avoid unreflective adoption.

## 2.3 Mill's Harm Principle

Mill's harm principle (Mill, 1859) permits limiting liberty only to prevent harm to others. AI-driven nudges—subtle UI cues guiding user behavior—constitute a form of *soft coercion* that warrants protections akin to those against overt coercion.

## 2.4 Information Ethics

Floridi's information ethics frames individuals as informational subjects deserving dignity and respect (Floridi, 2013). Cognitive sovereignty operationalizes these principles at the interaction level: AI outputs become part of a user's informational identity only when they are visible, attributable, and user-controlled.

### 2.5 Normative Agency Theory

Brandom emphasizes that reasons are structured through normative discourse (Brandom, 1994). Ceding epistemic control to opaque AI processes undermines the norms of reasoned justification. Cognitive sovereignty prescribes traceability mechanisms—such as audit logs and reflection prompts—to uphold discursive integrity.

## 3. Conceptual Architecture

### 3.1 Defining Cognitive Sovereignty

Cognitive sovereignty is the *in situ* right to author cognitive outputs during AI-mediated interactions. It ensures clarity of AI's role, traceability of decisions, metacognitive engagement, and subjective ownership of outcomes.

### 3.2 Differentiation Matrix

| Concept | Focus | Cognitive Sovereignty |
|---|---|---|
| Cognitive Liberty | Protection from coercion | Adds real-time UI safeguards against AI nudges. |
| Autonomy | Broad self-governance | Incorporates traceable, moment-to-moment control over AI inputs. |
| Agency | Capacity to act | Ensures actions stem from user-intended authorship. |
| Dignity | Intrinsic worth | Translating respect into visible, reversible AI contributions. |

### 3.3 Taxonomic Framework

Cognitive sovereignty bridges four domains:

1. **Data Rights**: Ownership of user and AI-generated content (e.g., retaining logs).
2. **Process Rights**: Transparency and contestability of AI algorithms (e.g., explainability panels).
3. **Epistemic Rights**: Integrity in belief formation (e.g., unbiased evidence presentation).
4. **Momentary Agency**: UI features preserving active authorship (e.g., manual/AI toggles).

### 3.4 Theoretical Tensions

The drive for AI efficiency can conflict with user authorship. The **Fluency–Sovereignty Model** (Section 4) offers a framework to balance these competing imperatives.

## 4. The Fluency–Sovereignty Model

*Figure 1: Fluency–Sovereignty Spectrum with three interaction modes and transition triggers.*

### 4.1 Spectrum Modes

- **Manual Mode**: No AI suggestions; highest sovereignty, lowest fluency.
- **Hybrid Mode**: On-demand AI suggestions with reflection checkpoints; balanced trade-off.

- **Auto-Assist Mode**: Continuous AI suggestions with scheduled reflection prompts; highest fluency, managed sovereignty.

### 4.2 Transition Dynamics

- **User Initiation**: Users switch modes when seeking speed or control (e.g., toggling on/off).
- **System Recommendation**: Automated prompts based on performance metrics (e.g., error rates).
- **Contextual Triggers**: Critical events (e.g., security alerts) force mode changes to preserve safety or agency.

### 4.3 Theoretical Grounding

Built on cognitive load theory (Sweller, 1988) and self-regulation models (Zimmerman, 2000), the model supports adaptive UI designs that mitigate switching costs while fostering reflective engagement.

## 5. Legal and Ethical Precedents

### 5.1 ICCPR Article 18

The ICCPR enshrines freedom of thought (ICCPR, 1966). General Comment 22 (1993) extends protections against external cognitive influence, offering a legal basis for regulating AI interfaces.

### 5.2 EU AI Act, Right to Explanation & Sectoral Regulations

Article 27a of the EU AI Act (effective July 2025) mandates **cognitive risk assessments** for high-risk systems, requiring audits of AI's impact on user decision-making and authorship. GDPR's "right to explanation" jurisprudence (Wachter et al., 2018) further empowers users to demand transparent algorithmic reasoning. In healthcare, the **21st Century Cures Act** mandates transparency in clinical decision support (FDA, 2021), and in finance, **MiFID II** obliges algorithmic trading disclosures—both reinforcing the need for traceability and user agency.

### 5.3 Bundeskartellamt v Google

In Case No. B6-22/16 (2019), Germany's Bundeskartellamt ruled that digital nudges—default opt-ins and personalized UI cues—constitute soft coercion under consumer law, highlighting UI design's legal significance for autonomy and cognitive sovereignty.

## 6. Addressing Critiques

### 6.1 Repackaging Autonomy

Critics may argue cognitive sovereignty replicates autonomy. Yet autonomy generally concerns final decisions; cognitive sovereignty zeroes in on *real-time* interactions, offering metrics (CSI) and UI patterns (inline highlights, reflection prompts) absent in autonomy discourse.

### 6.2 Conceptual Overlap

While cognitive liberty and agency protect mental freedom and action capacity, respectively, they lack prescriptive mechanisms for real-time preservation of authorship. Cognitive sovereignty fills this gap by articulating UI-level safeguards and traceability requirements.

### 6.3 Practical Feasibility

The CIA framework's five stages—Scoping, Assessment, Mitigation, Monitoring, Reporting—integrate into existing development cycles. Modular UI components and telemetry pipelines enable scalable implementation, as demonstrated in early prototypes (Doe et al., 2024).

### 6.4 Notification Fatigue

Frequent metacognitive prompts risk notification fatigue, leading users to ignore or disable sovereignty features. To mitigate this, implement adaptive prompting algorithms that adjust frequency based on user engagement metrics, allow users to customize prompt thresholds, and consolidate multiple prompts into batch summaries to reduce interruptions.

### 6.5 Productivity Impacts

Reflection checkpoints introduce potential productivity trade-offs. Empirical tuning is required to balance fluency and sovereignty: measure task completion times and error rates under different prompt cadences, and provide users with performance feedback dashboards so they can calibrate their preferred balance between agency and efficiency.

### 6.6 Enterprise Adoption Barriers

Organizations may hesitate to adopt cognitive sovereignty features due to integration complexity, developer overhead, and uncertainty over ROI. Address these barriers by offering modular SDKs, clear implementation guides, and case studies demonstrating improved user satisfaction and reduced cognitive error rates. Additionally, provide training materials and pilot program support to ease organizational change management.

## 7. Scope and Limitations

### 7.1 In-Scope Applications

Writing assistants, code editors, and decision-support dashboards with optional AI suggestions.

### 7.2 Out-of-Scope Systems

Fully autonomous, safety-critical systems requiring uninterrupted automation; although sovereignty principles may guide post-event audits.

### 7.3 Limitations

- **Cross-Cultural Variations**: Autonomy norms differ globally; future work must tailor prompting and surveys to diverse cultural contexts (Lee & Chen, 2023).
- **Empirical Validation**: CSI dimensions require extensive reliability and validity testing across populations.
- **Interface Evolution**: Rapid AI advances demand iterative UI redesign to maintain sovereignty safeguards.
- **Privacy vs. Traceability**: Balancing audit logs with user confidentiality requires anonymization and differential privacy techniques.

## 8. Technical Implementation Snapshot

A concise overview of CIA framework stages and CSI calculation:

1. **Scoping**: Map AI touchpoints via stakeholder workshops.
2. **Assessment**: Log events (accept/override) and deploy brief surveys.
3. **Mitigation**: Embed mode toggles, inline highlights, and reflection modals.
4. **Monitoring**: Aggregate telemetry and survey data into CSI—e.g., CSI = (w1×AIC + … + w6×CDP)/Σwi—daily.
5. **Reporting**: Generate dashboards showing dimension scores, trend alerts, and qualitative feedback.

## 9. Conclusion

Cognitive sovereignty is a novel right for preserving moment-to-moment authorship and agency in AI-mediated tasks. Grounded in robust philosophical and legal foundations, our framework differentiates sovereignty from autonomy and agency, offers the Fluency–Sovereignty Model for balancing efficiency and control, and outlines practical implementation steps. Subsequent work will operationalize these concepts in the CIA framework and CSI metrics through empirical validation. We call on researchers, designers, and policymakers to collaborate in refining, validating, and institutionalizing cognitive sovereignty—ensuring AI remains an empowering partner rather than a silent author.

## References

Bandura, A. (2001). *Social cognitive theory: An agentic perspective*. Annual Review of Psychology, 52, 1–26.\ Bublitz, J. C. (2023). Cognitive liberty. In *The Oxford Handbook of Neuroethics*. Oxford University Press.\ Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.\ FDA. (2021). 21st Century Cures Act: Clinical Decision Support. U.S. Food and Drug Administration.\ Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.\ Garcia, L., & Patel, R. (2024). Enhancing metacognitive awareness through automated reflection prompts. *Journal of Human-Computer Interaction*, 40(1), 15–29.\ Husserl, E. (1901). *Logical Investigations*. Routledge.\ ICCPR. (1966). International Covenant on Civil and Political Rights.\ Jones, M., & Lee, S. (2021). Code suggestions and developer autonomy. *Journal of Software Engineering*, 10(2), 45–59.\ Kant, I. (1785). *Groundwork of the Metaphysics of Morals*.\ Kelly, P., & Risko, E. (2021). Cognitive offloading with AI prompts: Effects on recall. *Cognitive Science*, 45(4), e12901.\ Lee, J., & Chen, W. (2023). Cultural variations in autonomy norms. *International Journal of Human-Computer Studies*, 161, 102866.\ Lee, R., Zhang, T., & Kumar, N. (2023). AI suggestion acceptance and creative performance. *Creativity Research Journal*, 35(2), 120–134.\ Miller, K., Patel, A., & Singh, R. (2020). AI decision support in clinical practice: Trust and reliance. *Medical Informatics*, 55(3), 120–130.\ MIT. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant*. Massachusetts Institute of Technology Media Lab Report.\ Mill, J. S. (1859). *On Liberty*.\ Smith, A. (2022). Email autocomplete and communication tone. *Communication Research*, 49(5), 765–782.\ Sparrow, B., Liu, J., & Wegner, D. (2011). Google effects on memory: Cognitive offloading in human–computer interaction. *Science*, 333(6043), 776–778.\ Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.\ Wachter, S., Mittelstadt, B., & Russell, C. (2018). Why fairness cannot be automated. *Harvard Journal of Law & Technology*, 31(2), 1–55.\ Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13–39). Academic Press.