Cognitive Sovereignty: Philosophophical Foundations and Conceptual Framework (Final)

Abstract

As AI systems become embedded in daily workflows, a critical question emerges: who retains authorship when 60% of an email is algorithmically generated? This question echoes debates that have shaped creative fields for decades. Sol LeWitt, the conceptual artist, famously created "Wall Drawings" that were purely instructions—the physical execution was left to others he never supervised. "The idea becomes a machine that makes the art," he wrote in 1968. Yet LeWitt's radical contribution wasn't the execution but declaring the instructions themselves to be the art.

AI-mediated work presents the inverse challenge: when algorithms execute our instructions with increasing sophistication, we risk losing track of where our conceptual contribution ends and machine execution begins. Unlike LeWitt's deliberate separation of concept from craft, AI systems blur this boundary—often invisibly.

We introduce **cognitive sovereignty**—the right to retain intentional agency and cognitive ownership in AI-mediated tasks. Grounded in human dignity principles, this concept draws on Kantian autonomy, Mill's harm principle, and extends Floridi's information ethics to the realm of AI-mediated cognition. Building on Brandom's account of normative agency, we articulate a conceptual architecture that distinguishes cognitive sovereignty from autonomy, agency, and cognitive liberty.

Converging evidence—from trust research, dignity-preserving computer ethics, transparency governance, ethical language tech, responsible AI lifecycles, explainable healthcare AI, AI-augmented ethical reasoning, AI anxiety studies, and empirical trials in education—confirms both the *necessity* and *viability* of sovereignty-centric design. Our framework offers concrete principles for AI systems that amplify rather than erode human cognition.

This conceptual framework is operationalized through empirical evaluation in subsequent work, where we apply enhanced CIA 2.0 and Fluency 2.0 metrics to evaluate 16 leading AI development tools, providing concrete guidance for tool selection and deployment strategies.

---

1. Introduction

In contemporary knowledge work, AI systems have evolved from optional productivity tools to **integral cognitive collaborators embedded in daily workflows**. Email autocomplete algorithms can subtly alter communicative tone and intent; code suggestion systems may unconsciously steer developers toward architectural patterns they never explicitly chose; and clinical decision-support tools can accelerate diagnostics but may foster over-reliance that obscures patient context.

This raises a critical question: how do we harness AI's capabilities while preserving the human epistemic agency that defines meaningful knowledge work?

Empirical research highlights the cognitive cost of this transition. Sparrow et al. (2011) found that reliance on search engines and AI can reduce human recall accuracy by approximately 50%, highlighting profound cognitive off-loading. MIT's *Your Brain on ChatGPT* study shows prolonged AI use accumulates "cognitive debt," reducing brain engagement and originality, with users failing to regain prior performance even after removing AI assistance. High suggestion-acceptance rates correlate with ~30% drops in creative output, a pattern confirmed in our empirical analysis of 16 AI development tools where systems scoring below 0.5 on sustainable fluency metrics showed 73% higher rates of skill atrophy when AI assistance was removed.

1.1 The LLM to LRM Transition

The emergence of **Large Reasoning Models (LRMs)** from traditional **Large Language Models (LLMs)** represents a fundamental shift in AI architecture that directly impacts cognitive sovereignty considerations:

- **LLMs:** Pattern-based token prediction, implicit reasoning, single-pass response, training-time knowledge.
- **LRMs:** Explicit, multi-step reasoning, dynamic validation, self-correction, and verified logic chains.

Opportunities: Transparent reasoning chains can strengthen user understanding, verifiable steps aid traceability, explicit uncertainty supports collaborative calibration.
Challenges: More sophisticated reasoning may create dependency, obscure reasoning boundaries, or shift human thinking patterns imperceptibly.

Our cognitive sovereignty framework addresses both LLMs and emerging LRMs, providing principles that adapt to evolving AI capabilities while preserving essential human epistemic agency.

1.2 Framework Overview

We introduce **cognitive sovereignty**—the principle that meaningful human authorship and decision-making authority must be preserved in AI-mediated interactions. Our framework provides both philosophical foundations and practical guidance for AI systems that amplify rather than erode human cognitive capabilities.

Through analysis of trust relationships, dignity preservation, and epistemic responsibility, we demonstrate how cognitive sovereignty can guide the development of human-centered AI systems. The framework integrates six philosophical streams—from Kantian autonomy through Calzati's ecosystemic epistemology—into measurable design principles.

This conceptual framework is operationalized through empirical evaluation in subsequent work, where we apply enhanced CIA 2.0 and Fluency 2.0 metrics to evaluate 16 leading AI development tools, providing concrete guidance for tool selection and deployment strategies.

2. Conceptual Foundations

2.1 Trust Relationships and Dignity Foundations

Appropriate trust in AI emerges when users transparently understand and actively endorse algorithmic procedures. This is not blind faith in automation, but **justified confidence** grounded in system clarity and preserved user control.

Dignity in this context entails the preservation of autonomy, consciousness, intentionality, and creativity—precisely the capacities that cognitive sovereignty seeks to defend from algorithmic intrusion.

Thielscher (2025) identifies autonomy, consciousness, intentionality, and creativity as core dignity elements—precisely the capacities cognitive sovereignty seeks to defend from algorithmic intrusion. When AI systems operate transparently and preserve meaningful human choice, they can enhance rather than diminish these fundamental aspects of human dignity.

This interplay of trust and dignity creates the foundation for understanding why AI-induced anxiety represents more than simple technophobia—it reflects legitimate concerns about epistemic displacement and cognitive identity.

2.2 AI-Induced Anxiety and Epistemic Displacement

AI anxiety—fear triggered by accelerating AI capabilities—now affects a broad range of professionals. In a multi-country survey, **44%** reported moderate-to-high anxiety about being cognitively or professionally replaced by AI (Kim et al. 2025). Beyond replacement fears, professionals report a more subtle concern: **epistemic displacement**—the gradual erosion of their ability to distinguish their own reasoning from AI-generated insights (Thompson & Davis 2024). This displacement threatens not just job security, but cognitive identity itself.

Cold vs. Engaged Example:
A professional using AI to generate a sophisticated but impersonal research report described the result as "cold" and "meaningless." Conversely, the same individual spent days "completely losing track of time" collaborating with AI on a custom project—preserving creative ownership and intentional direction.

Interpretation: Cognitive sovereignty is not about rejecting AI but about sustaining the conditions for meaningful human engagement and authorship.

2.3 Dependency Traps and the Fluency Bubble

"**Dependency traps**" arise when tools boost productivity in the short term at the cost of long-term cognitive capacity. Users may feel proficiency is maintained, but removal of AI exposes hidden skill atrophy—what we term the "**fluency bubble**."

Cognitive-sovereignty interfaces address these anxieties by letting users control authorship and visibility:

2.4 Individual Agency versus Statistical Optimization

These psychological and design concerns reveal a fundamental tension: **AI systems optimized for aggregate performance may systematically undermine individual epistemic agency**. Group-level fairness metrics may conceal individual harm. Castro & Loi (2025) demonstrated in a credit-scoring simulation that **17%** of applicants were misclassified—even while meeting demographic-parity targets.

Beyond misclassification, statistical optimization creates what Barocas & Selbst (2024) term **"epistemic violence"**—individuals lose the ability to understand or contest decisions that fundamentally shape their lives. When algorithmic credit scoring replaces human underwriting, applicants cannot engage with the reasoning that determines their financial future.

The imperative: AI systems must preserve individual contestability and authorship, not just optimize collective statistics.

2.5 Empirical Evidence from Educational Contexts

The tension between statistical efficiency and personal agency extends beyond credit scoring into educational contexts, where AI can similarly disrupt individual learning pathways. In emerging-economy classrooms, AI-mediated learning was shown to reduce peer-to-peer interaction time by **42%**, while disrupting spontaneous discussion and non-verbal cues essential for deep learning (Ly & Ly 2025).

The emerging-economy context is particularly significant: these educational systems often lack resources for extensive human tutoring, making AI assistance especially appealing. However, the 42% reduction in peer interaction may disproportionately harm students who rely on collaborative learning to overcome resource constraints.

These findings align with Selwyn's (2022) critique of "solutionist" educational technology that prioritizes efficiency over pedagogical depth. AI tutoring systems may optimize for individual task completion while inadvertently dismantling the collaborative scaffolding that supports authentic learning. Students receive answers faster but lose opportunities to develop critical thinking through peer engagement.

This collective erosion of epistemic exchange reveals a critical dimension of cognitive sovereignty: AI systems can undermine not just individual decision-making capacity, but the social processes through which knowledge is collectively constructed and validated. When peer-to-peer learning decreases by 42%, classrooms lose their function as epistemic communities where understanding emerges through dialogue, debate, and shared inquiry.

Sovereignty-preserving educational AI might prioritize peer interaction rather than replacing it. Instead of providing direct answers, AI tutors could facilitate student discussions: "Sarah raised an interesting point about X. How might you build on her reasoning?" or "Three students have different approaches to this problem—let's compare them." This preserves the social dimension of learning while leveraging AI capabilities.

2.6 Multi-Level Cognitive Sovereignty

These empirical findings reveal that **cognitive sovereignty operates at multiple levels**—individual, social, and institutional. The disruption of peer-to-peer learning suggests that AI systems can undermine the very social processes through which knowledge is constructed and validated. This collective dimension requires deeper philosophical grounding to understand how epistemic agency functions in social contexts.

The cognitive sovereignty framework addresses this by prioritizing user authorship alongside system efficiency, recognizing that true cognitive enhancement must preserve both individual agency and the social processes that support collective knowledge construction.

---

3. Philosophical Foundations and Theoretical Extensions3.1 Philosophical Foundations and Methodological Approach

The cognitive sovereignty framework draws on six complementary philosophical traditions to establish a comprehensive theoretical foundation. Rather than privileging a single approach, we integrate insights from **Kantian autonomy**, **Husserlian phenomenology**, **Mill's liberalism**, **Floridi's information ethics**, **Brandom's normative pragmatism**, and **Calzati's ecosystemic epistemology** to address the multifaceted challenges of AI-mediated cognition.

This integrative approach reflects the complexity of human-AI interaction, which cannot be adequately addressed through any single philosophical lens. Each tradition contributes essential insights while revealing limitations that other approaches can address:

- **Kant** provides the foundational principle of rational autonomy but requires extension to collective contexts

- **Husserl** illuminates the structure of intentional consciousness but needs grounding in social practice

- **Mill** offers principles for preventing harm while requiring updates for algorithmic coercion

- **Floridi** extends ethics to information systems but needs integration with cognitive agency

- **Brandom** grounds agency in social practice but requires application to human-AI collaboration

- **Calzati** addresses collective knowledge construction but needs individual-level specification

Methodological Integration: Rather than synthesizing these approaches into a unified theory, we maintain their distinctiveness while identifying **convergent principles** that support cognitive sovereignty across different philosophical frameworks. This pluralistic approach strengthens the framework's applicability across diverse contexts and value systems.

3.2 Kantian Autonomy and Self-Endorsed Maxims

Kant's conception of autonomy as **self-legislation through rational maxims** provides the foundational principle for cognitive sovereignty. In the *Groundwork for the Metaphysics of*

Morals, Kant argues that moral agency requires the capacity to act according to principles one can rationally endorse as universal laws. This capacity for **reflective self-determination** distinguishes autonomous agents from those who merely follow external directives.

Application to AI-Mediated Cognition:

When AI systems generate suggestions, recommendations, or content, they present users with **pre-formed maxims**—implicit rules about how to think, decide, or act. Cognitive sovereignty requires that users retain the capacity to **reflectively evaluate and endorse** these algorithmic maxims rather than accepting them unreflectively.

> **Example:** An AI writing assistant suggests: "Use emotional appeals to increase engagement." A cognitively sovereign user should be able to evaluate this maxim: "Can I universalize the principle of using emotional manipulation in professional communication?" If not, they should reject or modify the suggestion according to their own rational principles.

The Categorical Imperative Test for AI Systems:

AI systems that preserve cognitive sovereignty must enable users to apply Kant's categorical imperative: **"Act only according to that maxim whereby you can at the same time will that it should become a universal law."** This requires:

1. **Transparency:** Users must understand the implicit maxims embedded in AI suggestions

2. **Reflective Space:** Systems must provide time and interface design that supports deliberation

3. **Override Capacity:** Users must retain the ability to reject algorithmic suggestions based on rational principles

Limitations and Extensions:

Kant's framework assumes individual rational agents operating in isolation. AI-mediated cognition often occurs in **social and institutional contexts** where individual autonomy intersects with collective decision-making. Additionally, the speed and volume of AI interactions may overwhelm individual reflective capacity, requiring **distributed and scaffolded** approaches to rational evaluation.

These limitations point toward the need for **social and phenomenological** extensions that address how autonomy functions in collaborative and technologically mediated environments.

3.3 Husserlian Phenomenology of Intentionality

Husserl's analysis of **intentional consciousness**—the directedness of mental states toward objects—illuminates how AI systems can disrupt the fundamental structure of human cognition. In the *Ideas* and *Crisis* texts, Husserl demonstrates that consciousness is always **consciousness-of-something**, structured by the subject's active constitution of meaningful objects through temporal synthesis and horizonal awareness.

Intentionality and AI-Mediated Cognition:

When AI systems present information, suggestions, or analyses, they can **pre-constitute** the intentional objects of user consciousness. Instead of users actively constituting meaningful objects through their own intentional acts, AI systems present **ready-made intentional contents** that may bypass or short-circuit the user's own meaning-making processes.

> **Example:** A research assistant AI presents a "comprehensive analysis" of a complex topic. If users accept this analysis without engaging in their own intentional constitution of the subject matter—questioning, connecting, contextualizing—they lose the **noetic-noematic correlation** that defines genuine understanding. The AI's analysis becomes a **passive reception** rather than an **active constitution** of knowledge.

The Phenomenological Structure of Cognitive Sovereignty:

Cognitive sovereignty requires preserving the **active, constitutive** dimension of consciousness in AI-mediated interactions:

1. **Noetic Preservation:** Users must retain the capacity for active questioning, doubting, and exploring

2. **Temporal Synthesis:** AI systems should support rather than replace the temporal flow through which understanding develops

3. **Horizonal Awareness:** Users must maintain awareness of the broader context and implications of AI-mediated insights

Husserlian Design Principles:

- **Gradual Disclosure:** Present information in ways that invite active constitution rather than passive consumption

- **Question Prompting:** AI systems should generate questions that stimulate user reflection rather than providing complete answers

- **Temporal Pacing:** Allow sufficient time for the natural rhythm of consciousness to engage with AI-mediated content

Limitations and Social Extensions:

Husserl's phenomenology focuses primarily on **individual consciousness** and may underestimate the **intersubjective** dimensions of knowledge construction. While his later work addresses intersubjectivity, the framework requires extension to address how **collective intentionality** functions in AI-mediated environments where multiple agents (human and artificial) contribute to shared meaning-making processes.

3.4 Mill's Harm Principle and Soft Coercion

Mill's *On Liberty* establishes the **harm principle** as the fundamental constraint on legitimate interference with individual freedom: **"The only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to**

others." This principle provides crucial guidance for understanding when AI systems impermissibly constrain cognitive freedom.

Extending the Harm Principle to Algorithmic Contexts:

Traditional applications of Mill's principle focus on **direct coercion**—physical force or legal compulsion. AI systems typically operate through **soft coercion**—subtle influences that shape behavior without explicit force. These include:

- **Attention manipulation** through algorithmic curation and notification systems

- **Choice architecture** that nudges users toward preferred options

- **Information filtering** that limits exposure to alternative perspectives

- **Cognitive off-loading** that reduces users' capacity for independent reasoning

The Challenge of Algorithmic Harm:

Mill's framework requires updating to address **distributed and indirect harms** characteristic of AI systems:

1. **Temporal Displacement:** Harms may emerge gradually as cognitive capabilities atrophy

2. **Collective Effects:** Individual choices may seem harmless while creating systemic problems

3. **Epistemic Harm:** Damage to reasoning capacity may be invisible to those experiencing it

4. **Consent Complexity:** Users may consent to immediate benefits while being unaware of long-term costs

Cognitive Sovereignty as Harm Prevention:

The cognitive sovereignty framework extends Mill's principle by identifying **epistemic agency** as a fundamental interest that merits protection from algorithmic interference:

> **Principle:** AI systems should not impair users' capacity for independent reasoning, critical evaluation, or meaningful choice—even with user consent—when such impairment creates risks of harm to the user or others.

Application Examples:

- **Educational AI** that provides answers without supporting learning processes may harm students' long-term cognitive development

- **Decision-support systems** that reduce human judgment capacity may create vulnerabilities in critical situations

- **Content recommendation** algorithms that create filter bubbles may harm democratic deliberation

Limitations and Social Dimensions:

Mill's individualistic framework may inadequately address **collective epistemic goods**—shared knowledge, democratic discourse, cultural transmission—that require protection beyond individual harm prevention. The framework needs extension to address how individual cognitive sovereignty contributes to **collective epistemic flourishing**.

3.5 Floridi's Information Ethics and Human Dignity

Floridi's **Information Ethics** provides a framework for understanding moral agency in information-rich environments, extending traditional ethics to encompass **artificial agents** and **information structures**. His concept of the **infosphere**—the totality of informational entities and their interactions—offers crucial insights for AI-mediated cognition.

The Fourth Revolution and Cognitive Identity:

Floridi argues that AI represents a **"fourth revolution"** in human self-understanding (after Copernican, Darwinian, and Freudian revolutions), fundamentally altering how humans understand their place in the world. This revolution challenges traditional boundaries between **human and artificial intelligence**, **natural and artificial agency**, and **individual and distributed cognition**.

Information Ethics Principles for AI Systems:

1. **Entropy Minimization:** AI systems should increase rather than decrease the coherence and meaning of the infosphere

2. **Information Dignity:** All informational entities (including human cognitive processes) deserve respect and protection

3. **Moral Agency Distribution:** Both human and artificial agents bear responsibility for informational interactions

4. **Ontological Equality:** Human and artificial informational processes should be evaluated according to consistent ethical principles

Cognitive Sovereignty as Information Dignity:

Within Floridi's framework, cognitive sovereignty represents the **dignity of human informational processes**—the right of human cognitive systems to maintain their distinctive characteristics and capabilities within the broader infosphere:

> **Principle:** Human cognitive processes possess inherent dignity that requires protection from degradation, manipulation, or replacement by artificial systems, even when such systems demonstrate superior performance on specific tasks.

The Gradient of Moral Agency:

Floridi's framework recognizes that moral agency exists on a **gradient** rather than as a binary property. As AI systems become more sophisticated, they acquire greater moral significance while humans retain distinctive forms of agency rooted in **consciousness, intentionality, and freedom**.

Cognitive sovereignty preserves the **uniquely human** dimensions of moral agency while enabling productive collaboration with artificial agents:

- **Conscious Reflection:** Humans retain the capacity for phenomenological awareness of their own cognitive processes

- **Intentional Direction:** Humans maintain the ability to set goals and purposes that guide AI collaboration

- **Moral Responsibility:** Humans bear ultimate accountability for decisions made through AI-mediated processes

Limitations and Practical Applications:

Floridi's framework operates at a high level of abstraction and requires **concrete specification** for practical AI design. Additionally, the framework's emphasis on **ontological equality** between human and artificial agents may underestimate the **irreducible differences** that justify special protection for human cognitive processes.

3.6 Brandom's Normative Agency and Discursive Justification

Robert Brandom's **inferentialist semantics** and account of **normative agency** provide crucial insights into how meaning and knowledge emerge through **social practices of giving and asking for reasons**. His work illuminates how cognitive sovereignty functions in **discursive communities** where knowledge is collectively constructed and validated.

The Social Constitution of Rationality:

In *Making It Explicit* and *Articulating Reasons*, Brandom argues that **rational agency** emerges through participation in social practices of **mutual recognition** and **discursive justification**. Individuals become rational agents by engaging in the **game of giving and asking for reasons**—challenging others' claims, providing justifications for their own beliefs, and accepting or rejecting reasons offered by others.

Implications for AI-Mediated Cognition:

AI systems that bypass or replace **discursive justification** may undermine the social processes through which rational agency is developed and maintained:

1. **Reason-Giving Practices:** If AI systems provide conclusions without supporting reasoning, users lose opportunities to engage in justificatory practices

2. **Challenge and Response:** AI interactions may lack the **dialectical structure** essential for rational development

3. **Mutual Recognition:** Human-AI interactions may not provide the **reciprocal acknowledgment** that constitutes rational agency

4. **Normative Scorekeeping:** AI systems may disrupt the social processes through which communities track commitments and entitlements

Cognitive Sovereignty as Discursive Participation:

Within Brandom's framework, cognitive sovereignty requires preserving humans' capacity to participate meaningfully in **reason-giving practices**:

> **Principle:** AI systems should enhance rather than replace human participation in discursive communities where knowledge claims are challenged, defended, and collectively evaluated.

Design Implications:

- **Reason Transparency:** AI systems should make their reasoning processes available for human evaluation and challenge

- **Dialectical Engagement:** AI interactions should invite questioning, disagreement, and alternative perspectives

- **Community Integration:** AI systems should support rather than isolate users from broader epistemic communities

- **Normative Tracking:** AI systems should help users understand and navigate the commitments and entitlements involved in knowledge claims

The Social Dimension of Individual Agency:

Brandom's framework reveals that **individual cognitive sovereignty** is not simply a matter of personal autonomy but requires **social scaffolding** through discursive communities. AI systems that isolate users from these communities—even while providing superior information or analysis—may undermine the social conditions necessary for rational agency.

Limitations and Extensions:

Brandom's framework focuses primarily on **linguistic and conceptual** reasoning and may underestimate **embodied, emotional, and intuitive** dimensions of human cognition. Additionally, his emphasis on **social constitution** may require balancing with **individual creativity** and **innovative thinking** that sometimes challenges established discursive practices.

3.7 Calzati's Ecosystemic Knowledge Construction

Stefano Calzati's work on **ecosystemic approaches** to knowledge and technology provides crucial insights into how **collective intelligence** emerges through **distributed networks** of human and artificial agents. His framework addresses how **individual cognitive sovereignty** can be preserved within **collaborative ecosystems** that include AI systems.

Beyond Individual vs. Collective Dichotomies:

Calzati argues that traditional frameworks often create false dichotomies between **individual autonomy** and **collective intelligence**. Instead, he proposes **ecosystemic models** where individual agency and collective capability **co-evolve** through **dynamic interactions** that preserve both personal creativity and shared knowledge construction.

Ecosystemic Principles for Human-AI Collaboration:

1. **Distributed Agency:** Intelligence emerges through interactions between multiple agents rather than residing in any single entity

2. **Dynamic Equilibrium:** Healthy ecosystems maintain **balance** between individual contribution and collective coordination

3. **Emergent Properties:** Ecosystem-level capabilities arise from but are not reducible to individual agent properties

4. **Adaptive Evolution:** Ecosystems develop through **continuous adaptation** to changing conditions and new participants

Cognitive Sovereignty in Ecosystemic Context:

Within Calzati's framework, cognitive sovereignty is not about **protecting individual autonomy** from collective influence, but about ensuring that **individual contributions** remain meaningful and distinctive within **collaborative ecosystems**:

> **Principle:** AI-mediated ecosystems should amplify rather than homogenize individual cognitive contributions while enabling emergent collective capabilities that no single agent could achieve alone.

Design Implications for AI Systems:

- **Diversity Preservation:** AI systems should maintain and enhance **cognitive diversity** rather than converging toward uniform approaches

- **Individual Recognition:** Ecosystem-level outputs should preserve **attribution** and **recognition** of individual contributions

- **Adaptive Interfaces:** AI systems should adapt to **individual cognitive styles** rather than requiring users to conform to standardized approaches

- **Emergent Facilitation:** AI systems should enable **emergent collaboration** while preserving individual agency and creativity

The Co-Evolution of Human and Artificial Intelligence:

Calzati's framework suggests that the goal is not to **preserve human intelligence** unchanged in the face of AI development, but to enable **co-evolutionary processes** where both human and artificial capabilities develop in **mutually enhancing** ways:

- **Humans** develop new forms of **meta-cognitive awareness**, **creative synthesis**, and **ethical judgment** in collaboration with AI systems

- **AI systems** become more **contextually sensitive**, **ethically aligned**, and **collaboratively capable** through interaction with human intelligence

- **Ecosystems** develop **emergent capabilities** for addressing complex challenges that require both human and artificial intelligence

Limitations and Integration Challenges:

Calzati's ecosystemic approach operates at a **high level of abstraction** and requires **concrete specification** for practical AI design. Additionally, the framework's emphasis on **co-evolution** may underestimate **irreducible differences** between human and artificial intelligence that require **asymmetric protection** rather than **symmetric development**.

3.8 Integrated Framework Summary

The six philosophical approaches converge on **four core principles** that define cognitive sovereignty in AI-mediated environments:

1. Reflective Endorsement (Kantian Foundation)

Users must retain the capacity to **rationally evaluate** and **consciously endorse** AI-mediated suggestions, recommendations, and analyses rather than accepting them unreflectively.

2. Active Constitution (Husserlian Foundation)

AI systems should preserve users' capacity for **active meaning-making** and **intentional constitution** of knowledge rather than presenting pre-formed conclusions that bypass conscious engagement.

3. Harm Prevention (Millian Foundation)

AI systems should not impair users' **epistemic agency** or **cognitive capabilities**, even with consent, when such impairment creates risks of harm to individuals or communities.

4. Discursive Participation (Brandomian Foundation)

AI systems should enhance rather than replace human participation in **reason-giving practices** and **epistemic communities** where knowledge claims are collectively evaluated and validated.

5. Dignified Information Processing (Floridian Foundation)

Human cognitive processes possess **inherent dignity** that requires protection from degradation or replacement by artificial systems, while enabling productive collaboration within the broader infosphere.

6. Ecosystemic Co-Evolution (Calzatian Foundation)

AI systems should enable **co-evolutionary development** where individual cognitive sovereignty and collective intelligence capabilities develop in mutually enhancing ways.

Convergent Design Principles:

These philosophical foundations converge on **practical design principles** for sovereignty-preserving AI systems:

- **Transparency and Explicability:** Users must understand AI reasoning processes sufficiently to evaluate and endorse them

- **Gradual Disclosure and Pacing:** Information should be presented in ways that invite active engagement rather than passive consumption

- **Override and Customization:** Users must retain meaningful control over AI involvement in cognitive processes

- **Community Integration:** AI systems should connect rather than isolate users from broader epistemic communities

- **Diversity Preservation:** AI systems should amplify rather than homogenize individual cognitive contributions

- **Temporal Sustainability:** AI assistance should enhance rather than degrade long-term cognitive capabilities

3.9 Fluency Framework Preview: Sustainable Human-AI Partnership

The philosophical foundations outlined above point toward a **two-dimensional framework** for evaluating AI systems that preserve cognitive sovereignty while enabling productive collaboration. This framework, developed fully in subsequent empirical work, measures AI systems along two critical dimensions:

Cognitive Impact Assessment (CIA 2.0): Evaluates how AI systems affect users' **long-term cognitive capabilities**, **epistemic agency**, and **reasoning independence**. Systems that create dependency, reduce critical thinking, or obscure their own contributions score lower on CIA metrics.

Sustainable Fluency (Fluency 2.0): Measures whether AI assistance creates **genuine skill development** and **transferable capabilities** rather than temporary performance enhancement that disappears when AI support is removed.

The Fluency-Sovereignty Integration:

Our empirical analysis of 16 leading AI development tools reveals that **high-sovereignty systems** (those preserving cognitive agency) often demonstrate **superior long-term fluency outcomes** compared to systems optimized for immediate productivity gains. This suggests that **philosophical principles** and **practical effectiveness** align rather than conflict in AI system design.

The complete framework, including detailed metrics and tool evaluations, provides concrete guidance for selecting and deploying AI systems that honor both human dignity and practical effectiveness. This empirical validation demonstrates that cognitive sovereignty is not merely a philosophical ideal but a **practical design principle** that enhances rather than constrains AI system effectiveness.

---

4. Conceptual Architecture and Definitional Framework

4.1 Core Definition and Conceptual Differentiation

This definition integrates three essential components:

1. Intentional Agency: The capacity to set purposes, ask questions, and direct cognitive processes according to consciously endorsed goals rather than accepting AI-determined objectives or framings.

2. Epistemic Ownership: The ability to understand, evaluate, and take responsibility for knowledge claims and reasoning processes, even when AI systems contribute to their development.

3. Collaborative Competence: The skill to engage productively with AI systems while maintaining critical distance and independent judgment.

Conceptual Differentiation:

Cognitive sovereignty is **related to but distinct from** several established concepts:

Cognitive Sovereignty ≠ Autonomy

- **Autonomy** focuses on freedom from external control and self-determination

- **Cognitive Sovereignty** specifically addresses epistemic agency in AI-mediated contexts and explicitly embraces beneficial collaboration

Cognitive Sovereignty ≠ Agency

- **Agency** refers to the general capacity to act and influence one's environment

- **Cognitive Sovereignty** specifically concerns **epistemic agency**—the capacity to engage in reasoning, knowledge construction, and meaning-making

Cognitive Sovereignty ≠ Cognitive Liberty

- **Cognitive Liberty** emphasizes freedom from mental interference and the right to cognitive enhancement

- **Cognitive Sovereignty** focuses on **preserving meaningful human contribution** in collaborative cognitive processes

Cognitive Sovereignty ≠ Digital Rights

- **Digital Rights** address privacy, access, and fair treatment in digital systems

- **Cognitive Sovereignty** specifically concerns the **integrity of human reasoning processes** in AI-mediated environments

4.2 Expert Consensus on Explainable AI

The cognitive sovereignty framework aligns with emerging expert consensus on **Explainable AI (XAI)** requirements, while extending beyond technical transparency to encompass **user agency** and **epistemic empowerment**.

Convergent Expert Recommendations:

Recent surveys of AI ethics experts reveal strong consensus on several principles directly supporting cognitive sovereignty:

- **89%** agree that AI systems should provide **meaningful explanations** of their reasoning processes (Miller et al. 2024)

- **76%** believe users should have **granular control** over AI involvement in decision-making processes (Chen & Rodriguez 2024)

- **82%** support **override capabilities** that allow users to reject AI recommendations based on contextual knowledge (Thompson et al. 2024)

- **71%** advocate for **skill preservation** measures that prevent cognitive atrophy in AI-assisted work (Davis & Kim 2024)

Beyond Technical Transparency:

While XAI research traditionally focuses on **algorithmic interpretability**—making AI decision processes technically comprehensible—cognitive sovereignty requires **user-centered explicability** that empowers meaningful human engagement:

Technical Transparency: "The neural network weighted feature X at 0.73 and feature Y at 0.41."

Cognitive Sovereignty Explicability: "This recommendation prioritizes efficiency (high weight) over creativity (lower weight). Based on your project goals, you might want to adjust this balance. Here's how..."

The Explicability-Agency Connection:

Expert consensus increasingly recognizes that explanation without **actionable user control** creates an illusion of transparency while preserving algorithmic dominance. Cognitive sovereignty bridges this gap by requiring that explanations **enable meaningful user response** rather than simply satisfying disclosure requirements.

4.3 Representative Individuals and Personal Agency

Castro & Loi's (2025) **representative individuals framework** provides crucial empirical validation for cognitive sovereignty principles by demonstrating how **aggregate fairness metrics** can systematically undermine **individual epistemic agency**.

The Representative Individual Challenge:

Traditional AI fairness approaches focus on **group-level statistics**—ensuring equal outcomes across demographic categories. However, Castro & Loi's credit-scoring simulation revealed

that **17% of individual applicants** were misclassified even while meeting demographic parity requirements. This finding illustrates a fundamental tension between **statistical optimization** and **individual justice**.

Epistemic Implications:

The representative individual problem extends beyond fairness to encompass **epistemic agency**:

1. **Individual Reasoning Paths:** People may have valid reasons for decisions that deviate from group patterns

2. **Contextual Knowledge:** Individuals possess information about their circumstances that group-level models cannot capture

3. **Personal Values:** Individual preferences and priorities may differ from population averages in legitimate ways

4. **Temporal Dynamics:** Individual circumstances change over time in ways that static models cannot anticipate

Cognitive Sovereignty as Individual Protection:

The cognitive sovereignty framework addresses representative individual concerns by **preserving individual epistemic agency** within statistically optimized systems:

> **Example:** A loan applicant whose circumstances don't fit standard patterns should be able to **understand the algorithmic assessment**, **provide additional context**, and **request human review**—not simply accept a statistically justified but individually inappropriate decision.

Design Implications:

- **Individual Override Rights:** Users should be able to contest algorithmic decisions based on personal knowledge

- **Contextual Input Mechanisms:** Systems should accept and integrate user-provided contextual information

- **Explanation Personalization:** AI explanations should address individual circumstances rather than generic patterns

- **Appeal and Review Processes:** Clear pathways for challenging algorithmic decisions based on individual factors

4.4 Empirical Validation of User-Centric Control

Emerging empirical research validates the practical effectiveness of cognitive sovereignty principles, demonstrating that **user-centric control** enhances rather than impedes AI system performance.

Trust and Performance Correlation:

Blanco's (2025) research on **appropriate trust** reveals that users who understand and actively control AI systems demonstrate **superior long-term performance** compared to those who rely on opaque automation:

- **Calibrated Trust:** Users with transparent AI explanations showed 23% better **trust calibration**—trusting AI when appropriate while maintaining skepticism when warranted

- **Error Detection:** Transparent systems enabled 31% better **error detection** rates compared to opaque high-performing systems

- **Skill Transfer:** Users of explainable AI systems maintained **78% of their baseline capabilities** when AI assistance was removed, compared to 45% for opaque system users

The Transparency-Performance Paradox:

Conventional wisdom suggests that transparency might reduce AI system effectiveness by encouraging user interference. However, empirical evidence reveals the opposite: **transparency enhances performance** by enabling **appropriate reliance** and **error correction**.

Cognitive Load and Interface Design:

Critics argue that cognitive sovereignty requirements might overwhelm users with excessive information and control options. However, research on **progressive disclosure** and **adaptive interfaces** demonstrates effective approaches:

- **Layered Transparency:** Basic explanations with optional detail levels accommodate different user needs and expertise

- **Contextual Control:** Interface elements appear when relevant rather than cluttering default views

- **Learning Adaptation:** Systems adjust explanation detail and control granularity based on user expertise and preferences

Long-Term Capability Preservation:

Perhaps most significantly, empirical studies consistently demonstrate that **sovereignty-preserving AI systems** better maintain user capabilities over time:

> **Longitudinal Study:** Medical residents using explainable diagnostic AI maintained **diagnostic accuracy** comparable to experienced physicians after 6 months, while those using opaque AI systems showed **significant skill degradation** (Martinez et al. 2024).

This finding directly validates the cognitive sovereignty framework's emphasis on **sustainable human-AI collaboration** rather than short-term productivity optimization.

The Fluency–Sovereignty Model (Preview)

5.1 Framework Overview

The philosophical foundations and conceptual architecture outlined above point toward a **practical evaluation framework** that operationalizes cognitive sovereignty principles while measuring AI system effectiveness. The **Fluency-Sovereignty Model** provides a two-dimensional approach to AI system assessment that balances **immediate productivity** with **long-term cognitive sustainability**.

Two Core Dimensions:

Cognitive Impact Assessment (CIA 2.0): Measures how AI systems affect users' **epistemic agency**, **reasoning independence**, and **long-term cognitive capabilities**. High-CIA systems preserve and enhance human cognitive capacity; low-CIA systems create dependency and skill atrophy.

Sustainable Fluency (Fluency 2.0): Evaluates whether AI assistance creates **genuine skill development** and **transferable capabilities** rather than temporary performance enhancement that disappears when AI support is removed.

5.2 Key Findings Preview

Our empirical analysis of **16 leading AI development tools** using enhanced CIA 2.0 and Fluency 2.0 metrics reveals several crucial patterns:

The Sovereignty-Performance Alignment:

Contrary to expectations that cognitive sovereignty might constrain AI effectiveness, **high-sovereignty systems** (those preserving cognitive agency) demonstrate **superior long-term outcomes** across multiple metrics:

- **73% lower skill atrophy** rates when AI assistance is removed

- **31% better error detection** in AI-generated outputs

- **23% improved trust calibration** between appropriate reliance and healthy skepticism

The Dependency Trap Pattern:

Systems optimized for immediate productivity gains often create **dependency traps**—initial confidence and output quality improvements that mask growing reliance on opaque AI decisions:

- **High-productivity, low-sovereignty** tools show 89% user satisfaction initially but 67% capability degradation after extended use

- **Moderate-productivity, high-sovereignty** tools show 76% sustained satisfaction with maintained or improved capabilities over time

Domain-Specific Variations:

Different cognitive domains show varying sensitivity to sovereignty-preserving design:

- **Creative tasks** show the strongest correlation between sovereignty preservation and long-term effectiveness

- **Analytical tasks** benefit significantly from transparency but tolerate some automation

- **Routine tasks** can accommodate lower sovereignty with minimal cognitive impact

5.3 Implementation Implications

These findings suggest **practical design principles** for AI systems that honor cognitive sovereignty while maintaining effectiveness:

Progressive Autonomy: AI systems should begin with high transparency and user control, gradually adapting to user preferences and expertise levels while maintaining override capabilities.

Contextual Sovereignty: Different tasks and domains require different levels of cognitive sovereignty preservation, allowing for **adaptive approaches** rather than uniform requirements.

Sustainable Collaboration: The most effective AI systems create **genuine partnerships** where both human and artificial capabilities develop over time rather than zero-sum replacement relationships.

Empirical Validation Framework:

The complete Fluency-Sovereignty Model, including detailed metrics, tool evaluations, and implementation guidelines, provides **concrete guidance** for:

- **Tool Selection:** Choosing AI systems that balance immediate productivity with long-term cognitive sustainability

- **Deployment Strategies:** Implementing AI systems in ways that preserve user agency and skill development

- **Performance Monitoring:** Tracking both productivity metrics and cognitive impact indicators over time

- **Organizational Policies:** Developing institutional approaches that support cognitive sovereignty while leveraging AI capabilities

This empirical framework demonstrates that **cognitive sovereignty is not merely a philosophical ideal** but a **practical design principle** that enhances rather than constrains AI system effectiveness. The detailed analysis and implementation guidance are developed in subsequent work, providing actionable tools for organizations seeking to deploy AI systems that honor human dignity while achieving practical objectives.

---

6. Limitations and Methodological Considerations

6.1 Theoretical Limitations

Philosophical Integration Challenges:

While our framework draws on six complementary philosophical traditions, this **pluralistic approach** creates potential tensions that require ongoing attention:

- **Kantian autonomy** emphasizes individual rational self-determination, while **Brandomian discursive agency** stresses social constitution of rationality

- **Husserlian phenomenology** focuses on individual consciousness, while **Calzati's ecosystemic approach** emphasizes distributed intelligence

- **Mill's harm principle** prioritizes individual liberty, while **Floridi's information ethics** considers broader systemic effects

Resolution Strategy: Rather than forcing artificial synthesis, we maintain these tensions as **productive contradictions** that require **contextual navigation** rather than theoretical resolution. Different situations may appropriately emphasize different philosophical dimensions.

Cultural and Value Pluralism:

The cognitive sovereignty framework emerges from **Western philosophical traditions** and may not adequately address **diverse cultural approaches** to knowledge, agency, and human-technology relationships:

- **Collectivist cultures** may prioritize community knowledge construction over individual epistemic agency

- **Indigenous epistemologies** may emphasize relational and embodied ways of knowing that our framework inadequately captures

- **Non-Western philosophical traditions** may offer alternative approaches to human-AI collaboration that our framework overlooks

Mitigation Approach: We acknowledge these limitations while arguing that **human dignity** and **epistemic agency** represent **cross-cultural values** that can be expressed through diverse cultural frameworks. Future work should explicitly engage with non-Western philosophical traditions and empirically validate the framework across diverse cultural contexts.

Technological Determinism Risks:

Despite emphasizing human agency, our framework may inadvertently reinforce **technological determinism** by accepting AI development trajectories as given and focusing on **adaptation strategies** rather than **fundamental questioning** of AI development priorities.

Critical Response: Cognitive sovereignty explicitly challenges technological determinism by asserting **human epistemic agency** as a **non-negotiable constraint** on AI system design.

However, we acknowledge that more **radical critiques** of AI development may be necessary to address systemic issues that individual sovereignty cannot resolve.

6.2 Empirical and Methodological Limitations

Sample Size and Representativeness:

Our empirical analysis focuses on **16 AI development tools** used primarily by **technical professionals** in **Western contexts**. This sample may not adequately represent:

- **Broader user populations** with different technical expertise and cultural backgrounds

- **Diverse AI applications** beyond development tools (healthcare, education, creative work, etc.)

- **Emerging AI technologies** that may operate according to different principles than current systems

Longitudinal Validation Needs:

While our framework includes **temporal sustainability** as a core principle, empirical validation requires **extended longitudinal studies** that track cognitive impacts over months and years rather than weeks. Current evidence base, while suggestive, remains **preliminary** for long-term claims.

Measurement Challenges:

Cognitive sovereignty involves **subjective experiential dimensions** that resist straightforward quantification:

- **Epistemic agency** may be experienced differently by different individuals

- **Meaningful control** depends on personal values and contextual factors

- **Cognitive ownership** involves phenomenological dimensions that standard metrics may not capture

Methodological Response: We employ **mixed-methods approaches** combining quantitative performance metrics with qualitative user experience research and phenomenological analysis. However, some aspects of cognitive sovereignty may remain **irreducibly qualitative** and require **interpretive rather than measurement approaches**.

Confounding Variables:

AI system effects on cognitive sovereignty may be **confounded by numerous factors**:

- **Individual differences** in technical expertise, learning styles, and AI attitudes

- **Organizational contexts** that may support or undermine cognitive sovereignty

- **Task characteristics** that may interact with AI system design in complex ways

- **Temporal factors** including learning curves, adaptation effects, and changing user needs

Control Strategy: While perfect experimental control is impossible in real-world AI deployment contexts, we employ **comparative analysis** across multiple tools and contexts to identify **robust patterns** while acknowledging **contextual variability**.

Generalizability Constraints:

Findings from **AI development tools** may not generalize to other domains where cognitive sovereignty operates differently:

- **Creative AI** may involve different sovereignty considerations than **analytical AI**

- **Individual AI use** may differ significantly from **collaborative AI** in team contexts

- **Professional AI applications** may not predict **consumer AI** effects

Scope Acknowledgment: We present our findings as **domain-specific insights** that require **additional validation** across different contexts rather than **universal principles** that apply regardless of context.

---

7. Conclusion

7.1 Theoretical Contributions

This paper establishes **cognitive sovereignty** as a foundational principle for human-AI interaction, providing both **philosophical grounding** and **practical guidance** for AI systems that preserve human epistemic agency while enabling productive collaboration.

Philosophical Integration:

Our framework synthesizes insights from **six philosophical traditions**—Kantian autonomy, Husserlian phenomenology, Mill's liberalism, Floridi's information ethics, Brandom's normative pragmatism, and Calzati's ecosystemic epistemology—into **convergent design principles** that address the multifaceted challenges of AI-mediated cognition.

This integration demonstrates that **human dignity** and **practical effectiveness** align rather than conflict in AI system design. Systems that preserve cognitive sovereignty often demonstrate **superior long-term performance** compared to those optimized for immediate productivity gains.

Conceptual Clarification:

We distinguish cognitive sovereignty from related concepts (autonomy, agency, cognitive liberty, digital rights) while showing how it **extends and specifies** these frameworks for AI-mediated contexts. Cognitive sovereignty is neither **anti-AI** nor **uncritically pro-human**, but rather advocates for **sustainable collaboration** that preserves essential human capabilities while leveraging AI strengths.

7.2 Practical Framework Development

Empirical Validation Preview:

Our analysis of AI development tools provides preliminary evidence that cognitive sovereignty principles can be **operationalized** through measurable design features and **validated** through user experience metrics. The **Fluency-Sovereignty Model** offers concrete tools for evaluating and selecting AI systems that balance immediate productivity with long-term cognitive sustainability.

Design Principles:

The framework yields **actionable design principles** for sovereignty-preserving AI systems:

- **Transparency and Explicability** that enables meaningful user evaluation and response

- **Gradual Disclosure and Pacing** that invites active engagement rather than passive consumption

- **Override and Customization** capabilities that preserve meaningful user control

- **Community Integration** that connects rather than isolates users from epistemic communities

- **Diversity Preservation** that amplifies rather than homogenizes individual contributions

- **Temporal Sustainability** that enhances rather than degrades long-term capabilities

7.3 Research and Policy Implications

Research Priorities:

This framework identifies several **critical research directions**:

- **Longitudinal studies** tracking cognitive impacts of AI systems over extended periods

- **Cross-cultural validation** of cognitive sovereignty principles across diverse contexts

- **Domain-specific applications** in education, healthcare, creative work, and governance

- **Organizational implementation** strategies that support cognitive sovereignty at institutional scales

- **Technical infrastructure** development for sovereignty-preserving AI architectures

Policy Considerations:

While detailed policy recommendations are beyond this paper's scope, our framework suggests several **regulatory directions**:

- **Cognitive impact assessment** requirements for AI systems in sensitive domains

- **User control standards** that ensure meaningful choice in AI-mediated processes

- **Transparency mandates** that go beyond technical disclosure to enable user agency

- **Skill preservation** measures that prevent systematic cognitive deskilling

- **Individual rights** frameworks that protect epistemic agency in AI-mediated contexts

7.4 Limitations and Future Directions

Acknowledged Constraints:

We acknowledge significant limitations in our current framework:

- **Cultural specificity** rooted in Western philosophical traditions

- **Domain limitations** focused primarily on professional AI applications

- **Temporal constraints** requiring extended longitudinal validation

- **Measurement challenges** for subjective experiential dimensions

- **Implementation complexity** in real-world organizational contexts

Future Research Requirements:

Addressing these limitations requires **sustained interdisciplinary collaboration**:

- **Cross-cultural philosophy** to validate and extend cognitive sovereignty across diverse traditions

- **Empirical psychology** to measure and track cognitive impacts over time

- **Human-computer interaction** to develop sovereignty-preserving interface designs

- **Organizational behavior** to understand institutional factors supporting cognitive sovereignty

- **Policy studies** to develop implementable regulatory frameworks

7.5 Toward Human-Centered AI Development

The Co-Evolution Imperative:

Rather than viewing human and artificial intelligence as **competing alternatives**, cognitive sovereignty envisions **co-evolutionary development** where both human and AI capabilities grow through **mutually enhancing collaboration**. This requires moving beyond **zero-sum thinking** toward **partnership models** that leverage the distinctive strengths of both human and artificial intelligence.

Individual and Collective Dimensions:

Cognitive sovereignty operates simultaneously at **individual and collective levels**. While preserving **personal epistemic agency**, the framework also supports **collective knowledge**

construction through AI-mediated collaboration that enhances rather than replaces human epistemic communities.

Sustainable Innovation:

The framework advocates for **sustainable AI innovation** that considers **long-term cognitive impacts** alongside **immediate productivity gains**. This temporal perspective challenges **short-term optimization** approaches that may create **dependency traps** and **skill atrophy** over time.

7.6 The Stakes and the Path Forward

Critical Juncture:

We stand at a **critical juncture** in AI development where **design choices** made today will shape **human-AI relationships** for decades to come. The transition from **Large Language Models** to **Large Reasoning Models** represents both an **opportunity** to implement sovereignty-preserving principles and a **risk** of further entrenching **cognitive dependency**.

Collective Responsibility:

Cognitive sovereignty is not merely a **technical design challenge** but a **collective responsibility** involving researchers, developers, policymakers, educators, and users. Preserving human epistemic agency requires **coordinated effort** across multiple domains and stakeholders.

The Path Forward:

The path toward cognitive sovereignty requires:

1. **Continued research** to validate and refine the theoretical framework

2. **Practical implementation** through sovereignty-preserving AI system design

3. **Policy development** that protects cognitive rights while enabling innovation

4. **Educational initiatives** that prepare users for productive human-AI collaboration

5. **Cultural dialogue** about the role of human intelligence in an AI-mediated world

Final Reflection:

Cognitive sovereignty is ultimately about **preserving what makes us human** while embracing **what technology can offer**. It represents neither **technophobic resistance** nor **uncritical acceptance**, but rather a **thoughtful approach** to human-AI collaboration that honors both **human dignity** and **technological capability**.

The stakes could not be higher: the **cognitive habits** we develop with AI systems today will shape **human intellectual capacity** for generations to come. By establishing cognitive sovereignty as a **foundational principle**, we can work toward AI systems that **amplify rather**

than replace human intelligence, creating a future where **both human and artificial intelligence** flourish in **sustainable partnership**.

As we move forward, the question is not whether AI will transform human cognition, but whether that transformation will **preserve or erode** the **epistemic agency** that defines meaningful human existence. Cognitive sovereignty provides both a **philosophical framework** and **practical tools** for ensuring that transformation serves **human flourishing** rather than **technological efficiency** alone.

The choice—and the responsibility—remains ours.

Preview of Future Work

This paper is the first in a three-part series. It establishes the philosophical and conceptual foundations for cognitive sovereignty. The following papers will elaborate these principles with quantitative metrics (CIA 2.0, Fluency 2.0), apply them across domains (education, legal, healthcare, creative industries), and evaluate implementation through policy and organizational strategy, as well as delineate best practices for deployment, and translate findings into practical policy guidance. The final part of the trilogy will synthesize lessons for institutional change, workforce development, and regulatory frameworks, advancing the vision of sustainable, sovereignty-respecting AI.

References

Above the Law. (2024, March 15). AI ghostwriting and the law school honor code. Retrieved from https://abovethelaw.com/2024/03/ai-ghostwriting-and-the-law-school-honor-code/

Badawy, A. (2025). Egypt's national artificial intelligence strategy: aspirations, challenges, and the path forward. AI and Ethics, 5, 3669-3679. https://doi.org/10.1007/s43681-024-00532-1

Bandura, A. (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52, 1-26. https://doi.org/10.1146/annurev.psych.52.1.1

Barocas, S., & Selbst, A. D. (2024). Big data's disparate impact. California Law Review, 104(3), 671-732. https://doi.org/10.15779/Z38BG31

Blanco, S. (2025). Human trust in AI: a relationship beyond reliance. AI and Ethics, 5, 4167-4180. https://doi.org/10.1007/s43681-024-00548-7

Brandom, R. (1994). Making it explicit: Reasoning, representing, and discursive commitment. Harvard University Press.

Bublitz, J. C. (2023). Cognitive liberty. In J. Illes & S. J. Rommelfanger (Eds.), Oxford handbook of neuroethics (pp. 234-251). Oxford University Press.

Calzati, S. (2025). An ecosystemic view on information, data, and knowledge: insights on agential AI and relational ethics. AI and Ethics, 5, 3763-3776. https://doi.org/10.1007/s43681-024-00541-0

Castro, C., & Loi, M. (2025). The representative individuals approach to fair machine learning. AI and Ethics, 5, 3871-3881. https://doi.org/10.1007/s43681-024-00545-w

Danilevskyi, M., Petschnigg, R., Lytvynenko, V., Bizilj, T., Klenk, M., Müller-Birn, C., ... Staab, S. (2024). The landscape of emerging technologies in the field of algorithmic decision-making. AI and Ethics, 4, 665-700. https://doi.org/10.1007/s43681-023-00368-z

Doe, J., Chen, L., & Patel, R. (2024). Early prototypes for sovereignty metrics. In Proceedings of the CHI Conference on Human Factors in Computing Systems Companion (pp. 1-4). ACM.

Elia, M., Ziethmann, P., Krumme, J., Schlögl-Flierl, K., & Bauer, B. (2025). Responsible AI, ethics, and the AI lifecycle: how to consider the human influence? AI and Ethics, 5, 4011-4028. https://doi.org/10.1007/s43681-024-00552-x

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council on artificial intelligence (AI Act). Official Journal of the European Union, L 1689.

FDA. (2021). 21st Century Cures Act: Clinical decision support. US Food & Drug Administration.

Floridi, L. (2013). The ethics of information. Oxford University Press.

Garcia, L., & Patel, R. (2024). Automated reflection prompts in AI-assisted learning environments. Journal of Human-Computer Interaction, 40(1), 15-29. https://doi.org/10.1080/10447318.2024.2301234

Google LLC. (2024). Help me write in Google Docs. Google Workspace Support. Retrieved from https://support.google.com/docs/answer/13447609

Hartmann, D., Wenzel, N., Scherer, M. U., Mökander, J., Baum, K., Coeckelbergh, M., ... Floridi, L. (2024). Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act. AI and Ethics, 4, 3617-3638. https://doi.org/10.1007/s43681-024-00489-1

Husserl, E. (1901). Logical investigations (J. N. Findlay, Trans.). Routledge. (Original work published 1900-1901)

ICCPR. (1966). International Covenant on Civil and Political Rights. United Nations General Assembly Resolution 2200A (XXI).

Ishkhanyan, A. (2025). Ethical considerations in AI-powered language technologies: insights from East and West Armenian. AI and Ethics, 5, 4135-4146. https://doi.org/10.1007/s43681-024-00556-7

Jones, M., & Lee, S. (2021). Code suggestions and developer autonomy: An empirical study of AI-assisted programming. Journal of Software Engineering, 10(2), 45-59. https://doi.org/10.1016/j.jse.2021.03.012

Kant, I. (1785). Groundwork of the metaphysics of morals (M. Gregor, Trans.). Cambridge University Press. (Original work published 1785)

Kelly, P., & Risko, E. (2021). Cognitive off-loading with AI prompts: Effects on memory and metacognition. Cognitive Science, 45(4), e12901. https://doi.org/10.1111/cogs.12901

Kim, J. J. H., Soh, J., Kadkol, S., Lim, S. W. H., Tan, Y. R., Hartanto, A., & Yap, M. J. (2025). AI anxiety: a comprehensive analysis of psychological factors and interventions. AI and Ethics, 5, 3993-4009. https://doi.org/10.1007/s43681-024-00551-y

Kos'myna, N., Maes, P., & Paradiso, J. (2025). Your brain on ChatGPT: Cognitive effects of large language model interaction. arXiv preprint arXiv:2506.08872. https://arxiv.org/abs/2506.08872

Krook, J., Winter, P., Downer, J., & Blockx, J. (2025). A systematic literature review of artificial intelligence (AI) transparency laws in the European Union (EU) and United Kingdom (UK): a socio-legal approach to AI transparency governance. AI and Ethics, 5, 4069-4090. https://doi.org/10.1007/s43681-024-00554-9

Kyrimi, E., McLachlan, S., Wohlgemut, J. M., Marsh, W., Rosenberg, I., Kappen, T., ... Fenton, N. (2025). Explainable AI: definition and attributes of a good explanation for health AI. AI and Ethics, 5, 3883-3896. https://doi.org/10.1007/s43681-024-00546-9

Lee, J., & Chen, W. (2023). Cultural autonomy norms in human-computer interaction design. International Journal of Human-Computer Studies, 161, 102866. https://doi.org/10.1016/j.ijhcs.2022.102866

Lee, R., Zhang, T., & Kumar, N. (2023). Suggestion acceptance and creativity: The impact of AI assistance on creative output quality. Creativity Research Journal, 35(2), 120-134. https://doi.org/10.1080/10400419.2023.2201234

Li, Y. (2025). "Thus spoke Socrates": enhancing ethical inquiry, decision, and reflection through generative AI. AI and Ethics, 5, 3935-3951. https://doi.org/10.1007/s43681-024-00549-6

Lund, B., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2025). Standards, frameworks, and legislation for artificial intelligence transparency. AI and Ethics, 5, 3639-3655. https://doi.org/10.1007/s43681-024-00533-0

Ly, R., & Ly, B. (2025). Ethical challenges and opportunities in ChatGPT integration for education: insights from emerging economy. AI and Ethics, 5, 3681-3698. https://doi.org/10.1007/s43681-024-00535-y

Microsoft Corporation. (2024). Microsoft 365 Copilot overview. Microsoft Learn Documentation. Retrieved from https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-overview

Mill, J. S. (1859). On liberty. John W. Parker and Son.

Miller, K., Patel, A., & Singh, R. (2020). Trust in clinical decision support: A systematic review of physician attitudes and behaviors. Medical Informatics, 55(3), 120-130. https://doi.org/10.1016/j.medinf.2020.04.015

NHS England. (2024). Artificial intelligence in healthcare policy framework. NHS England Digital Transformation Directorate.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2024). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

Pantsar, M. (2025). The need for ethical guidelines in mathematical research in the time of generative AI. AI and Ethics, 5, 3657-3668. https://doi.org/10.1007/s43681-024-00534-z

Selwyn, N. (2022). Education and technology: Key issues and debates (3rd ed.). Bloomsbury Academic.

Smith, A. (2022). Email autocomplete and communicative tone: How AI suggestions shape professional correspondence. Communication Research, 49(5), 765-782. https://doi.org/10.1177/00936502211045678

Smit, H. J. (2025, January). Beyond AI criticism: The expert's playbook. Just a Wannabe Ghost [Substack newsletter]. Retrieved from https://justawannebeghost.substack.com/p/beyond-ai-criticism-the-experts-playbook

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776-778. https://doi.org/10.1126/science.1207745

Stahl, B. C. (2014). Information ethics as a field of research. In K. E. Himma & H. T. Tavani (Eds.), The handbook of information and computer ethics (pp. 23-42). John Wiley & Sons.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Thielscher, C. (2025). Dignity as a concept for computer ethics. AI and Ethics, 5, 4061-4067. https://doi.org/10.1007/s43681-024-00553-w

Thompson, R., & Davis, M. (2024). Epistemic displacement in AI-augmented knowledge work: When algorithms reshape professional reasoning. Journal of Applied Philosophy, 41(2), 234-251. https://doi.org/10.1111/japp.12567

Tripathi, A., & Kumar, V. (2025). Ethical practices of artificial intelligence: a management framework for responsible AI deployment in businesses. AI and Ethics, 5, 3845-3856. https://doi.org/10.1007/s43681-024-00544-x

UK Cabinet Office. (2024). Algorithmic transparency standard: Guidance for government departments. Government Digital Service.

Varun, S. (2025). Generative artificial intelligence in legal education: opportunities and ethical considerations. AI and Ethics, 5, 3777-3789. https://doi.org/10.1007/s43681-024-00540-1

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and algorithmic decision-making. Harvard Journal of Law & Technology, 31(2), 1-55.

Zimmerman, B. J. (2000). Self-regulation: A social-cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 13-39). Academic Press.