H-AGO: Human-Centric Agentic Governance with Creative Capacity Preservation

Shabista Shabista, Ravi Gupta

Independent Researcher, Independent Researcher
shabista0690@gmail.com, ravigupta1989@gmail.com

Abstract

Recent studies suggest that habitual generative AI use may erode human creative capacity—paralleling documented effects of GPS on spatial cognition and search engines on memory recall. We introduce Human-Centric Agentic Governance & Observability (H-AGO), a framework that shifts AI governance from content filtering to creative capacity preservation. H-AGO implements three innovations: (1) a **Creative Capacity Index (CCI)** that longitudinally tracks human ideation ability independent of AI assistance using embedded measurement and minimal-burden micro-assessments, (2) **dual-phase constitutional governance** (CreativeSilence and Socratic Guardrails) operating at real-time and longitudinal scales, and (3) **adaptive intervention policies** that adjust AI assistance parameters based on measured CCI trends. Unlike existing AI safety approaches focused on output harm, H-AGO governs for human capability preservation—a core principle of Well-Being AI. We propose a hybrid edge-cloud architecture where governance runs locally while high-quality Socratic scaffolding leverages cloud LLMs. We present a planned evaluation design for creative writing tasks targeting 30–40 participants over 4 weeks.

Introduction

The "Google effect" demonstrated that search engine access reduces memory retention for information people expect to find online (Sparrow, Liu, and Wegner 2011). GPS navigation has been shown to impair spatial memory and hippocampal function (Dahmani and Bohbot 2020). Calculator dependence correlates with reduced mental arithmetic ability (Ellington 2003). A pattern emerges: *cognitive tools that substitute for human effort may erode the capacities they replace*.

Generative AI presents an unprecedented case of this phenomenon. Unlike calculators (narrow) or GPS (domain-specific), LLMs substitute for *general creative cognition*—ideation, divergent thinking, and problem formulation. Early evidence suggests AI-assisted writers produce more homogeneous content (Doshi and Hauser 2024) and may develop reduced creative self-efficacy (Yao and Wang 2024).

Current AI governance focuses on preventing harmful outputs. We argue this is insufficient: **AI systems should also be governed for their effects on human capability over time**—a core principle of the Well-Being AI paradigm (Calvo and Peters 2014).

The Creative Capacity Index (CCI)

We introduce the **Creative Capacity Index**—a longitudinal metric tracking human ideation ability *without* AI assistance.

CCI Components and Measurement

$$\text{CCI}_t = \alpha \cdot D_t + \beta \cdot F_t + \gamma \cdot S_t \quad (1)$$

Each component is measured through methods designed to minimize user burden:

Divergent Thinking (D_t): Measured via two complementary approaches: (1) *Embedded analysis* of the user's creative output during baseline sessions—automated scoring of idea count, semantic diversity (via embedding similarity), and novelty relative to prior work; (2) *Optional weekly micro-assessment*—a 60-second Alternative Uses Task ("List as many uses for [common object] as you can") (Guilford 1967). Users may skip micro-assessments; the system defaults to embedded analysis.

Flow Frequency (F_t): Measured automatically via HRV-based flow state detection (de Manzano et al. 2010) combined with session metrics: uninterrupted creative work duration, absence of AI help requests, and sustained typing/activity patterns (Csikszentmihalyi 1996). No user effort required.

Self-Efficacy (S_t): Captured via a single end-of-session prompt: "Rate your creative confidence today (1–5)." This takes approximately 3 seconds and is optional; if skipped, S_t is estimated from behavioral proxies (AI request frequency, session abandonment rate) (Bandura 1977).

Two-Level Governance Architecture

H-AGO operates governance at two temporal scales that are *complementary, not contradictory*:

Real-time governance: During any session, the Sovereignty Agent detects flow (triggers CreativeSilence) or block (triggers Socratic Guardrails). This operates regardless of CCI level.

Longitudinal governance: CCI trends adjust the *parameters* of real-time governance over time. If CCI declines, the

Level	Timescale	What It Governs
Real-time (Phase)	Seconds to minutes	When to intervene during a session
Longitudinal (CCI)	Weeks	Overall AI availability policy

Table 1. Two-level governance: real-time and longitudinal

CCI Trend	Policy Adjustment
Increasing (>5%)	Reduce minimum silence duration by 10%; allow earlier Socratic activation
Stable (±5%)	No change to current thresholds
Declining (>5%)	Increase minimum silence duration by 15%; require longer independent work before AI assistance; increase baseline session frequency

Table 2. CCI-adaptive governance parameters

system becomes more protective by modifying thresholds (see Table 2).

CCI-Adaptive Policy Adjustments

This creates a feedback loop: AI assistance is *contingent on maintained human capacity*. Users who maintain or improve their unassisted creative abilities retain full AI access; those showing decline receive more "creative exercise" before AI becomes available.

System Architecture

Figure 1 illustrates the H-AGO architecture with CCI integration.

Flow and Block Detection

The Sovereignty Agent detects creative state using multimodal signals:

A lightweight classifier (<2KB model) combines these signals. Such compact models can run efficiently on edge hardware using quantized architectures (Ma et al. 2024). Thresholds are calibrated per-user during onboarding baseline sessions. HRV-based flow detection follows established psychophysiological methods (de Manzano et al. 2010).

Dual-Phase Governance

Socratic Question Generation

When block is detected, the user's context is sent to a cloud LLM with a constrained system prompt:

```
You are a Socratic writing coach. The user
is stuck. Ask ONE question to help them
discover their own solution.

Rules:
- Ask exactly one question
- Never provide answers or content
- Never complete their work
- Focus on character, conflict, or emotion
- Keep under 20 words

User's work: [context]
Problem: [block description]
```

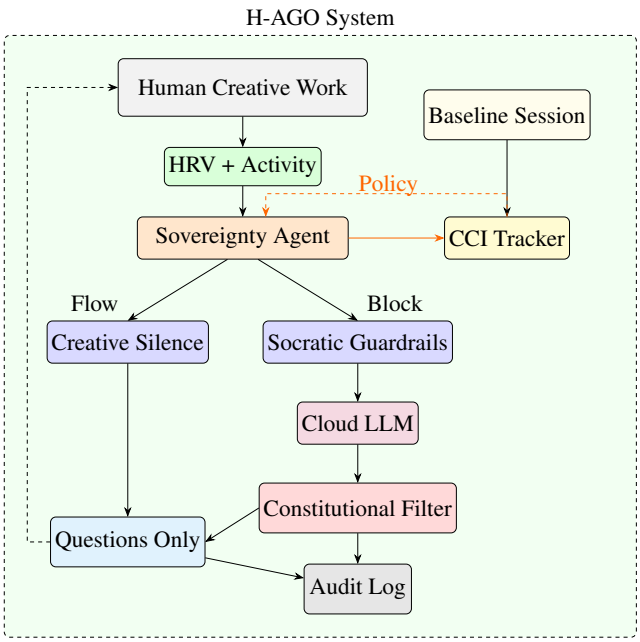

Figure 1: H-AGO architecture. Orange paths: CCI feedback loop adapts governance policy based on baseline session performance (Ma et al. 2024).

Signal	Flow Indicators	Block Indicators
HRV pattern	High HRV, stable rhythm	Low HRV, irregular
Activity	Consistent velocity, few pauses >10s	Long pauses (>30s), deletions
AI requests	None	Repeated requests or hovering
Duration	>5 min uninterrupted	Frequent stops/starts

Table 3. Flow and block detection signals

The LLM response is validated locally via grammar-constrained decoding (Willard and Louf 2023) to ensure only interrogative structures pass through—an application of constitutional AI principles (Bai et al. 2022) to creative assistance. Example interaction:

Human: "I'm stuck on the ending."

Standard AI: "The hero finds the treasure was friendship."

H-AGO: "What does your protagonist fear most?"

Phase	Trigger	Behavior
CreativeSilence	Flow detected	AI constitutionally barred from acting
Socratic	Block detected	Questions only (no completions)
Baseline	Scheduled	No AI; measures unassisted CCI

Table 4. H-AGO governance phases

Metric	Baseline	H-AGO (Est.)
Unassisted ideation score	100	115–125
AI dependency (override rate)	100%	60–70%
Creative self-efficacy	3.2/5	3.8–4.2/5
Session continuation (no AI)	12 min	18–25 min
CCI after 4 weeks	Declining	Stable/Improving

Table 5. Projected outcomes vs. always-on AI baseline (to be validated)

Planned Evaluation

Table 5 presents *projected* outcomes based on analogous interventions in cognitive offloading literature (Risko and Gilbert 2016). These are hypotheses to be tested, not empirical findings.

Study Design

We plan a controlled longitudinal study with the following design:

Participants: 30–40 adults (ages 18–45), recruited from online creative writing communities and university writing programs.

Task: Short story writing (500–1000 words per session). This task was selected because: (1) it requires sustained ideation and divergent thinking, (2) output is measurable and can be blind-rated for creativity, (3) it represents a common use case for AI writing assistants, and (4) it has established creativity assessment protocols (Amabile 1982).

Duration: 4-week longitudinal study with 3 sessions per week and weekly CCI assessments.

Conditions: Between-subjects design comparing three groups: (1) Always-on AI assistance, (2) H-AGO with CCI-adaptive governance, (3) No AI baseline.

Measures: Unassisted ideation scores (blind-rated by 2 independent raters), AI request frequency, session duration without AI, creative self-efficacy scale (Bandura 1977), and CCI trajectory.

Connection to Well-Being AI

The Well-Being AI paradigm emphasizes that AI systems should be evaluated not only by task performance but by their effects on human flourishing (Calvo and Peters 2014). H-AGO instantiates this paradigm through three mechanisms:

Capability preservation over output optimization: Traditional AI assistants optimize for output quality, potentially at the cost of human skill development. H-AGO explicitly prioritizes maintaining human creative capacity (measured via CCI) even if this reduces short-term productivity.

Autonomy protection: By enforcing silence during flow states and requiring questions-only interaction during blocks, H-AGO preserves the user's sense of creative ownership and self-efficacy—key components of psychological well-being (Ryan and Deci 2000).

Long-term human development: The CCI feedback loop ensures that AI assistance remains contingent on sustained human capability, preventing the learned helplessness that can result from over-reliance on AI tools.

H-AGO demonstrates that Well-Being AI is not merely about avoiding harm, but about actively designing systems that support human growth alongside task completion.

Risks and Limitations

User frustration: Enforced silence during perceived need may cause abandonment. Mitigation: transparent CCI rationale display; user-adjustable (but logged) overrides.

Measurement validity: CCI relies on proxy metrics; true creative capacity is difficult to operationalize (Runco and Jaeger 2012). Mitigation: multi-dimensional measurement combining behavioral and self-report data; validation against established creativity instruments.

Cloud dependency: High-quality Socratic questions require large LLMs. Fully edge-based deployment with small models (2–3B) may produce generic questions. Mitigation: hybrid architecture; future work on Socratic-specific fine-tuning using instruction-tuning methods (Ouyang et al. 2022).

Individual differences: Optimal governance parameters likely vary by person, task, and creative domain. Mitigation: CCI-adaptive personalization and user-amendable policy thresholds.

Long-term adherence: Users may disable H-AGO or use external AI tools. The system cannot prevent this. Mitigation: gamification of CCI progress; visible capacity metrics dashboard.

Contributions

1. **Creative Capacity Index:** First longitudinal metric for human ideation preservation in AI-assisted workflows, with embedded and minimal-burden measurement methods
2. **Two-Level Governance:** Real-time phase detection (flow/block) combined with longitudinal CCI-adaptive policy adjustment
3. **Constitutional Silence:** Architectural enforcement of non-intervention during creative flow states
4. **Socratic Guardrails:** Prompt-engineered and grammar-constrained generation ensuring questions-only assistance

Conclusion

Current AI governance asks: "Is this output harmful?" H-AGO asks: "Is this assistance pattern harming human capacity?" By introducing the Creative Capacity Index and

capacity-contingent governance, we shift from reactive content filtering to proactive capability preservation. Aligned with Well-Being AI principles, the goal is not to limit AI usefulness, but to ensure that usefulness does not come at the cost of human creative growth.

References

Amabile, T. M. 1982. Social Psychology of Creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology*, 43(5): 997–1013.

Bai, Y.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.

Bandura, A. 1977. Self-Efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review*, 84(2): 191–215.

Calvo, R. A.; and Peters, D. 2014. *Positive Computing: Technology for Wellbeing and Human Potential*. MIT Press.

Csikszentmihalyi, M. 1996. *Creativity: Flow and the Psychology of Discovery*. HarperCollins.

Dahmani, L.; and Bohbot, V. D. 2020. Habitual Use of GPS Negatively Impacts Spatial Memory. *Scientific Reports*, 10: 14228.

de Manzano, Ö.; et al. 2010. The Psychophysiology of Flow During Piano Playing. *Emotion*, 10(3): 301–311.

Doshi, A.; and Hauser, O. 2024. Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Ideas. *Science Advances*, 10: eadn5290.

Ellington, A. J. 2003. A Meta-Analysis of the Effects of Calculators on Students' Achievement and Attitude Levels in Precollege Mathematics Classes. *Journal for Research in Mathematics Education*, 34(5): 433–463.

Guilford, J. P. 1967. *The Nature of Human Intelligence*. McGraw-Hill.

Ma, S.; Wang, H.; et al. 2024. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. *arXiv:2402.17764*.

Ouyang, L.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *NeurIPS*.

Risko, E. F.; and Gilbert, S. J. 2016. Cognitive Offloading. *Trends in Cognitive Sciences*, 20(9): 676–688.

Runco, M. A.; and Jaeger, G. J. 2012. The Standard Definition of Creativity. *Creativity Research Journal*, 24(1): 92–96.

Ryan, R. M.; and Deci, E. L. 2000. Self-Determination Theory and Intrinsic Motivation. *American Psychologist*, 55(1): 68–78.

Sparrow, B.; Liu, J.; and Wegner, D. M. 2011. Google Effects on Memory. *Science*, 333(6043): 776–778.

Willard, B. T.; and Louf, R. 2023. Efficient Guided Generation for LLMs. *arXiv:2307.09702*.

Yao, J.; and Wang, L. 2024. The double-edged sword of generative AI on creative self-efficacy: A longitudinal investigation. *Journal of Applied Psychology*.