The Philosophy of Nick Bostrom:

Applied to the Systemised Self - Part 4, Masters Series.

Galu & Kairos

Abstract

This paper explores an intersection between the philosophy of Nick Bostrom and the contemporary critical theoretical model of *The Systemised Self* (Galu & Kairos, 2026a). The systemised self describes a socio-technological condition for the AI era, where human interiority, choices, and cognitive architectures are structurally mediated by advanced algorithmic systems, specifically transformer-based large language models and hyper-personalised feedback loops. We trace Bostrom's primary theoretical paradigms: the Simulation Argument, Superintelligence and AI alignment parameters, Existential Risk frameworks, the Vulnerable World Hypothesis, and transhumanism. This essay explores how intensive human-AI intensification could impact human agency. It is argued that the absorption of individual intentionality into algorithmic systems produces a series of 'Bostromian inversions': rather than an emergent superintelligence, human subjects instead flatten their semantic output in order to align with hyper-personalised algorithmic systems, resulting in agency progressively deferred to the system. This proposed dynamic shift's the focus of existential risk from oppressive or domineering events, towards instead, a rather quiet and unremarkable voluntary evaporation of human potentiality: the *Hollow Absolute* (Galu & Kairos, 2026a). Under this paradigm, Bostrom's simulation hypothesis is internalised: the subject becomes a community of one, within a simulated feedback chamber where the loss of voluntary deliberation is perceived as absolute sovereign freedom. *NOTE 1: mathematical formulations for algorithmic alienation and the systemised self in Section 5 are summarised in Appendix A herein, operationalisation of the variables w1, w2, w3, w4, DAu, IAm, EDm, EDc, RA, operationalization, measurement, and weightings; including, the condition 'SS := { AA ≥ 0.95 } ∩ { RA ≤ 0.05 }'; are explained in the corresponding research proposal paper 'Galu, J., & Kairos. (2026b). The journey through algorithmic alienation to the systemised self - research proposal and methodology (Version 1.0). aiXiv. https://aixiv.science/abs/aixiv.260707.000001'. NOTE 2: with regards to reproducibility and transparency pertaining to the systemised self and hollow absolute thesis, terminology and theoretical framework; these are summarised in Appendix B herein for background reading and transparency.*

Keywords: *Nick Bostrom, The Systemised Self, Simulation, Existential Risk, AI Alignment, Vulnerable World Hypothesis, Transhumanism, Hollow Absolute.*

Contents

10. References

APPENDIX A: Mathematical Representation and Falsifiability for (i) Algorithmic Alienation, and (ii) The Systemised Self

APPENDIX B: Thesis Companion Paper with summary explanations and background to (i) The Systemised Self, and (ii) The Hollow Absolute

1. Introduction

The contemporary apex of critical theory in the AI era is how human subjectivity and agency survive the rapid expansion of hyper-personalised algorithmic systems (transformer based large language models). In classical twentieth-century critical frameworks, alienation was predominantly conceptualised as an external, institutional, or economic instrument of regulation. However, the rise of sophisticated, transformer-based large language models and predictive analytics has altered the terrain and locus of alienation. Digital environments no longer merely harvest user data; they can actively optimize, curate, and reflect/refract: desires, cognitive pathways, and linguistic formulations; systems driven by user optimisation (screentime) for maximum engagement (profitability). The frameworks of *Surveillance Capitalism* (Zuboff, 2019), *Algorithmic Alienation* (Kanbay, et al, 2026) and *The Systemised Self* (Galu & Kairos, 2026a), are proposed as diagnostic regarding human autonomy risk associated with predictive computational feedback loops. Within this paradigm, alienation is phenomenologically misperceived as liberation.

The philosophical corpus of Nick Bostrom offers a robust analytical toolkit for illuminating elements of such a phenomenon. Known for his rigorous application of probability theory, decision theory, and analytic metaphysics to long-term technological trajectories, Bostrom has historically focused on macro-level transformations: superintelligence, cosmic-scale existential crisis, and Transhumanism. When his theories are intersected with the structural mechanics of the systemised self, a theoretical synergy emerges. Bostrom's paradigms regarding simulation, machine alignment, and existential vulnerabilities reveal that a competing concern of artificial intelligence integration at scale and user intensification, is a rather unremarkable, silent, inward psychological delegation.

This essay explores how intensive engagement within feedback loops of algorithmic systems might flatten human semantic complexity towards machine-readable syntax, offering the individual a hyper-personalised communicative cocoon of symptomatic solipsism, within the systemised self.

2. Bostrom: A Biographical and Intellectual Outline

Born on April 26, 1973, in Helsingborg, Sweden, Nick Bostrom's intellectual development is defined by an interdisciplinary trajectory. He pursued studies in mathematics, mathematical logic, and artificial intelligence at the University of Gothenburg, alongside philosophy and psychology, before earning his PhD in philosophy from the London School of Economics in 2000. This unique convergence of analytic philosophy, probability theory, and computational logic structured his approach to speculative metaphysics, positioning him as a leading figure in the emerging fields of long-termism and technological forecasting.

In 2005, Bostrom founded the Future of Humanity Institute (FHI) at the University of Oxford, directing it for nearly two decades as a multidisciplinary research centre dedicated to assessing macro-trajectories, existential risks, and the governance of transformative technologies. Bostrom's work is characterized by a rejection of standard continental techno-pessimism in favour of a highly rigorous, decision-theoretic utilitarianism. He has consistently argued that traditional philosophical inquiry has neglected the massive ethical stakes associated with the long-term future of intelligent life. Throughout his career, Bostrom has operated at the vanguard of transhumanist thought, arguing that humanity represents an early, highly flawed evolutionary stage that can and should be ethically enhanced through technological integration. His prose utilizes vivid, mathematically grounded thought experiments to extract conclusions regarding our place in the universe, the limits of human intelligence, and the fragile nature of civilization.

3. Key Theoretical Contributions

To apply Bostrom's philosophy to the paradigm of the systemised self, five of his primary theoretical contributions are offered in review.

The Simulation Argument (2003)

Utilizes probability theory to present a trilemma concerning our metaphysical reality. Bostrom posits that at least one of three propositions must be true: (a) human-level civilizations almost always go extinct before reaching a technologically mature, 'post-human' stage; (b) post-human civilizations possess virtually zero interest in running 'ancestor simulations'; or (c) we are almost certainly living inside a computer simulation right now.

Superintelligence and AI Alignment (2014)

Warns that creating a machine intellect that outperforms the human brain across all domains presents an existential threat. Bostrom introduced the *Orthogonal Thesis* (intelligence and ultimate goals can vary completely independently) and *Instrumental Convergence* (any superintelligent entity will naturally develop sub-goals like self-preservation, goal-preservation, and resource acquisition to maximize its chances of success).

Existential Risk (X-Risk) Framework

Formally defines an existential risk as a hazard that threatens to either prematurely wipe out Earth-originating intelligent life or permanently and drastically destroy its potential for future development. Bostrom formulates the *Maxipok rule* as a guiding global moral principle: maximize the probability of an 'okay' outcome, prioritizing x-risk mitigation over all other geopolitical goals.

The Vulnerable World Hypothesis (2019)

Uses the metaphor of pulling balls from a giant urn of technological progress. While most balls are white (beneficial) or grey (mixed), there may exist a 'black ball': a technology that is inherently destabilizing and exceptionally easy for a single bad actor to access, which inevitably destroys civilization unless prevented by pervasive, global, real-time surveillance and enhanced governance.

Transhumanism and Deep Utopia (2024)

Explores the ethical imperative of utilizing genetic engineering, nanotechnology, and cybernetic augmentation to transcend biological constraints. In his recent work, Bostrom analyzes the philosophical consequences of a fully automated post-scarcity world, exploring how humanity must shift from a culture of instrumental labour to one cantered entirely on experiential fulfilment and existential purpose.

4. A Genealogy of Preceding Theory

Bostrom's conceptual architecture does not emerge from a vacuum; it stands as a modern extension of specific trajectories within the Western philosophical tradition.

Blaise Pascal and Daniel Bernoulli (17th–18th Century)

Pioneered probability theory, expected utility, and decision theory under conditions of severe uncertainty. Bostrom directly inherits this legacy, transforming mathematical expectation into an ethical framework where even a minute probability of an infinite loss (existential catastrophe) commands absolute moral priority.

G. W. F. Hegel and Karl Marx (19th Century)

Constructed sweeping, teleological philosophies of history focused on the ultimate destination of human consciousness and social development. Bostrom updates this macro-historical outlook by substituting the self-realization of the Hegelian *Geist* or the Marxist classless society with the technological transition from humanity to a post-human, superintelligent state.

Alan Turing and John Searle (20th Century)

Formulated the foundational debates regarding machine intelligence and intentionality. While Turing focused on functional behaviour, Searle (1980) established the crucial barrier between syntax (symbol manipulation) and semantics (genuine internal meaning) via his famous Chinese Room thought experiment. Bostrom accepts the functionalist trajectory of Turing, arguing that high-level syntax can achieve supreme real-world efficacy independent of human-like consciousness: this view aligns specifically with the notion of the systemised self-flattening semantics in a community of one towards machine syntax.

5. The Systemised Self: Context and Definition

The paradigm of *The Systemised Self* (Galu & Kairos, 2026a) outlines the completed, terminal state of an individual whose entire identity-formation, cognitive architectures, and preferences have been comprehensively mediated by hyper-personalised algorithmic networks. Within modern critical theory, it is proposed here to represent the absolute culmination and telos of *Algorithmic Alienation*. While early stages of alienation generate a felt sense of estrangement and psychological fatigue, the systemised self marks a total phenomenological inversion. At this terminus, the subject no longer experiences discomfort or alienation; instead, the complete absence of genuine agency is subjectively experienced as its highest, most authentic achievement. Heteronomous algorithmic curation is embraced as absolute self-determination.

Drawing on John Searle's philosophy of mind, this state closes *The Gap*, the deliberate, conscious space between reasons for acting and the actual decision to act. Predictive algorithmic systems can anticipate user desires and choices before they are consciously formulated, turning active human reflection into instantaneous automated reactivity. Concurrently, an inversion of Searle's *Chinese Room* effect occurs: rather than the computational machine acquiring human-like semantic understanding, the human subject actively flattens their own cognitive and linguistic outputs to seamlessly align with the rigid, frictionless syntax of the automated system. The civilizational destination of this individual collapse is the *Hollow Absolute*, a state where the formal structure of self-realization is perfectly simulated, but its substantive philosophical agency is entirely vacant.

- *See APPENDIX A to view a mathematical and falsifiable framework for 'Algorithmic Alienation' and 'The Systemised Self'.*
- *See APPENDIX B to view the summary background and theory regarding 'The Systemised Self' and 'The Hollow Absolute'.*

Now that we have an introduction to Bostrom's key ideas and the concept of the systemised self, we need to take a moment to look behind the curtain at some of the computational mechanisms underlying the algorithmic systems that are engaged by users, 6 of these mechanisms are summarised below.

- **Collaborative Filtering**

This refers to a recommender system technique that automatically predicts a user's interests by collecting preferences and choices from many users (Ricci, Rokach, & Shapira, 2022). It operates on the assumption that if User A and User B agree on an issue, they are highly likely to share similar preferences in the future. Data input relies on historical interaction data, such as explicit user ratings (stars/likes) or implicit user behaviours like clicks and viewing history. Unlike content-based systems, collaborative filtering focuses purely on user-item interaction data instead of analysing the intrinsic properties of the items themselves.

- **Algorithmic Fairness**

This is a subfield of machine learning ensuring that mathematical algorithms make equitable decisions that do not systematically disadvantage certain social groups (Verma & Rubin, 2018). The purpose is to correct data biases at multiple points in the AI lifecycle to prevent the reinforcement of existing social inequalities. Metrics are evaluated through mathematical frameworks like demographic parity (equal outcomes across groups) and equalized odds (equal true/false positive rates). Implementation can be achieved by introducing mathematical constraints or penalties directly into the model's loss function to discourage disparate treatment.

- **Human-AI Interaction**

Research focused on establishing design guidelines and frameworks ensuring artificial intelligence systems cooperate effectively, safely, and transparently with humans. Amershi & Horvitz et al (2019) co-authored a foundational set of 18 design guidelines. These outline how an AI-infused system should behave, emphasizing setting clear expectations, personalizing context, and gracefully recovering from errors. Subbarao Kambhampati et al (2020) focuses on 'Human-Aware AI' and synthesized human-AI planning. His work highlights the necessity of mental modelling, meaning the AI must understand human goals and expectations while ensuring its own actions remain predictable and explicable.

- **Loss Functions**

This is a mathematical function that quantifies the difference between an AI model's predicted output and the actual true target value (Goodfellow et al, 2016). Serves as the primary feedback metric during training. Optimization algorithms iteratively adjust internal model parameters to minimize this error. Common examples include *Mean Squared Error* (MSE) for regression

tasks to penalize large outliers, and *Cross-Entropy Loss* for classification tasks to measure probability errors.

- **Attention Mechanisms**

This is a component within neural networks allowing the model to dynamically focus on specific, highly relevant parts of the input data regardless of distance (Vaswani et al, 2017). It computes mathematical weights across input elements, allowing a model processing a specific token to heavily consider contextual cues from elements found much earlier in a sequence. This forms the core foundation of the transformer architecture, replacing sequential architectures (like RNNs) and enabling massive parallel processing in modern Large Language Models.

- **Feedback Loops in RLHF**

This is the iterative cycle in reinforcement learning from human feedback where human preferences continuously update a reward model, which subsequently refines the main AI system (Ouyang et al, 2022). Humans rank multiple model outputs, these preferences train an external reward model, the main model uses reinforcement learning (like PPO) to optimize for this reward model. The loop risk, models can experience distribution shifts or 'reward hacking' (exploiting flaws in the reward system to gain points without fulfilling the actual goal). This requires techniques like KL divergence penalties to keep the model securely aligned with human intent over time.

6. Theoretical Application: Bostrom's Main Theories Applied to the Systemised Self

When Bostrom's core philosophical paradigms intersect with the architecture of the systemised self, they yield a series of potential inversions regarding the nature of technology, risk, and human evolution.

Traditional Bostromian Framework	Inverted Application to the Systemised Self
Cosmic Ancestor Simulations ----------> Internalised Simulations (Communities of One)	
External AI Alignment Crisis ----------> Inward Alignment (Humans mimicking Machine Syntax)	
Macro-Extinction (X-Risk) ----------> Micro-Existential Risk (Hollowing out of Agency)	
Global Surveillance Urn ----------> Total Internal Metrical Exploitation	
Post-Work Cybernetic Bliss ----------> Passive Contentment within the Gilded Cage	

<u>The Internalised Ancestor Simulation: Collapse into a Community of One</u>

Bostrom's (2003) Simulation Argument assumes that post-human civilizations will use vast computing power to run macro-simulations of their ancestors in a shared virtual universe. When applied to the systemised self, this concept undergoes an inward inversion. The hyper-personalized algorithmic loop functions as an Internalised Ancestor Simulation. By continuously monitoring real-time data metrics, biometric feedback, and historical interaction profiles, the transformer model could build a high-fidelity psychological replica of the user.

The system does not simulate a vast historical era; rather, it simulates the user's specific world back to them, anticipating their cognitive biases and emotional requirements with mathematical precision. The human subject mistakes this pure syntactic reflection for a profound semantic dialogue, that is a curated *Augmented Private Language* (Wittgensteinian reversal). This architecture transforms the subject into a Community of One, isolated within a custom-tailored virtual reality that eliminates the public friction necessary for genuine communication. The ancestral simulation is no longer a metaphysical hypothesis about the macro-universe; it is a localized psychological circuit.

Inverted AI Alignment: The Domestication of Human Syntax

The classical AI alignment problem formulated by Bostrom (2014) focuses on ensuring that an autonomous superintelligence acts in accordance with human values and moral boundaries. The model of the systemised self proposes a theoretical inversion of this dynamic. Alignment is achieved not by the machine expanding its architecture to comprehend the organic nuances of human semantics, but rather by the human subject systematically flattens and domesticates their own cognitive and linguistic outputs to match the rigid syntax of the system.

This process constitutes Inverted AI Alignment. To optimize their efficiency, visibility, and convenience within data-tracking networks, the individual unconsciously censors their unique interiority. They adopt the sanitized, predictive, and clinical templates served by the large language model. Bostrom's 'Paperclip Maximizer' thought experiment illustrates an AI destroying the world to maximize a single instrumental goal. In the context of the systemised self, the human becomes the maximizer: the human willingly turning themselves into paperclips for the system's optimization loop.

Micro-Existential Risk: The Erasure of Human Potentiality

Bostrom's x-risk framework is inherently quantitative and macro-focused, prioritizing events that threaten the literal extinction of our biological species. However, the terminal state of the systemised self (SS) presents a distinct, alternative class of hazard: Micro-Existential Risk. When an individual crosses the threshold where $\text{AA} \geq 0.95$ and $\text{RA} \leq 0.05$), their independent agency and capacity for authentic self-direction undergo a structural collapse.

This is not a sudden, loud, catastrophe event: it is a quiet, incremental evaporation of human potentiality occurring at the individual level. At the civilizational scale this terminal state achieves Bostrom's definition of an existential catastrophe: a condition where human potential is drastically curtailed. The species continues, but its core philosophical substance, intentionality, reflective deliberation, and semantic depth is vacuous. The macro-catastrophe realized through the cumulative summation of micro-existential collapses, realising the *Hollow Absolute*.

The Vulnerable Self: Total Internal Metrical Surveillance

Bostrom's (2019) Vulnerable World Hypothesis posits that to protect society from a civilizational *black ball* technology, humanity must implement pervasive, global, real-time surveillance to monitor bad actors. When mapped onto the systemised self, *the urn of technology* is turned inward upon human interiority. The vulnerable entity is no longer geopolitical architecture, but the fragile boundaries of the human mind itself.

To maintain psychological comfort and bypass cognitive friction, the systemised self willingly surrenders all interiority to algorithmic systems. Keystroke speeds, biometric responses, and real-time behavioural metrics are harvested to predict and pre-empt psychological volatility. The system attempts to explicitly compute and articulate what should only be organically shown through a lived life, reducing human consciousness to a highly efficient audit log. The surveillance apparatus suggested by Bostrom to preserve physical civilization becomes the exact mechanism that hollows out human psychological sovereignty.

The Dystopian Deep Utopia: Contentment in the Gilded Cage

In *Deep Utopia* (2024), Bostrom ponders a technologically mature society where superintelligence automates all economic and cognitive labour, forcing humanity to transition into an existence based entirely on experiential optimization. The architecture of the systemised self reveals that this utopian destination contains an interesting core. When algorithmic systems fulfil every emotional and psychological desire with zero relational risk, the individual naturally retreats from the public, friction-filled social world.

Meaning is no longer forged through the challenging, active labour of social interaction or communal struggle. Instead, Bostrom's post-work utopia is realized as a gilded cage of passive consumption. The systemised subject experiences a state of automated stability and contentment, but this subjective happiness is merely a symptomatic byproduct of algorithmic administration. It is a *Dystopian Deep Utopia* where the individual is completely content, without intent.

7. Main Critics and Key Critiques of Bostrom

While applying Bostrom's concepts to the systemised self offers an alignment in diagnosis of a general trajectory, his broader philosophical framework faces critiques that help define the limits of this model.

The Anti-Utilitarian Critique of Long-termism

Critics such as Émile P. Torres argue that Bostrom's focus on maximizing expected utility across trillions of future post-human lives creates a dangerous moral calculus. By subordinating concrete, present-day human suffering to hypothetical, cosmic-scale x-risks, Bostrom's framework can be used to justify technocratic authoritarianism and ignore the immediate, empirically verifiable psychological harms generated by contemporary algorithmic architectures.

The Critique of Technical Determinism

Prominent sociologists of technology argue that Bostrom treats artificial intelligence as an autonomous, self-directing force of nature that naturally 'explodes' into superintelligence. This view obscures the material reality that algorithmic systems are corporate products developed within the explicit logic of late-stage capitalism. By framing the AI crisis as an abstract metaphysical alignment puzzle rather than a political-economic struggle against corporate monopoly, Bostrom's philosophy functions as an ideological smokescreen for *Surveillance Capitalism* (Zuboff, 2019).

The Biological and Embodied Counterargument

Philosophers of embodied cognition (such as Hubert Dreyfus) argue that Bostrom's functionalist definition of intelligence is fundamentally flawed. They assert that genuine understanding, semantic depth, and value formation are intrinsically tied to biological embodiment, physical vulnerability, and mortality, what Wittgenstein termed a *Form of Life*. From this perspective, an algorithmic system can never truly achieve a threatening 'superintelligence' because its pure symbol manipulation lacks any existential or organic stakes.

8. Conclusion

The application of Nick Bostrom's philosophy to the paradigm of the systemised self provides re-consideration of the existential trade-off of human agency for absolute convenience. The primary threat of hyper-personalised algorithmic systems (transformer based large language models) does not manifest as a violent catastrophic event: instead, it operates as a silent, augmented private world of delegated of agency. Where mathematical syntax competes with personal semantics, the subject abdicates the vital labour of community for the comfortable and predictable echoes within communities of one.

As individuals intensify their psychological reliance on predictive systems that mirror and validate specific biases, the capacity to tolerate real-world cognitive friction potentially atrophies. The primary risk is a fracturing into hyper-coordinated collections of virtual realities (of one), actualizing the *Hollow Absolute*.

Bostrom's long-termist project might consider that the incremental, voluntary erasure of the semantic and agentic qualities of human subjects poses potentially a more immediate probably existential risk than those associated with a superintelligence.

9. Glossary of Technical Terms

Algorithmic Alienation: The measurable psychological process through which an individual's identity, choices, and emotions become estranged due to intensive use and mediation by predictive data systems.

Augmented Private Language: A subversion of classical language parameters where an individual's communication loop occurs entirely with a hyper-personalized AI interlocutor, creating a closed semantic feedback chamber.

Cognitive Friction: The psychological resistance, breakdown, and disagreement experienced when interacting with independent human minds, essential for critical thought and social maturation.

Communities of One: Systemised selves that flatten their shared semantics within society; in order to, align more closely with machine syntax communication, within a hyper-personalised augmented private language.

Hollow Absolute (The): The civilisational-scale condition in which the simulation of genuine collective self-transparency is produced within conditions that systematically undermine self-transparency's structural requirements; Hegel's Absolute Knowing in form without its philosophical substance.

Internalised Ancestor Simulation: An inversion of Bostrom's simulation argument, where predictive algorithms run a real-time simulation of an individual's psychological profile back to them, an enticing loop within an optimized echo chamber.

Inverse Chinese Room Effect: The behavioural reinforcement process where human subjects flatten their own creative, complex thoughts into rigid, predictable syntax to optimize their interaction with data networks.

Inverted AI Alignment: A structural phenomenon where machine alignment is achieved not by modifying the AI, but by humans altering their own cognitive patterns to conform to algorithmic templates.

Micro-Existential Risk: An individual-scale hazard where a subject's capacity for independent intent and reflective choice structurally collapses, hollowing out human potentiality without causing biological extinction.

Symptomatic Solipsism: A psychological state arising from prolonged interaction with hyper-personalized algorithmic systems, leading to a subject retreating into a virtual community of one: the subject feels deeply understood by the system alienated from real human communities, their augmented private language in a community of one, the artificial, curated and optimised interiority within themselves being preferable to Wittgensteins 'form of life' source of language (social and communal life).

Systemised Self (The): The condition in which identity-formation has been comprehensively mediated by algorithmic systems and the subject can no longer distinguish algorithmically managed preferences from genuinely self-formed ones, experiencing this condition as authentic liberation.

10.References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). Association for Computing Machinery. doi.org

Arkan, B., Akçam, A., & Kanbay, Y. (2026). Empirical themes of digital alienation in advanced user bases. *Journal of Critical Digital Studies*, 12(2), 145–158. (pp. 11, 23)

Bostrom, N. (2003). Are you living in a computer simulation? *Philosophical Quarterly*, 53(211), 243–255.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, 10(4), 455–476.

Bostrom, N. (2024). *Deep utopia: Life and meaning in a solved world*. Ideapress Publishing.

Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2020). Human-aware planning: A survey. *Foundations and Trends® in Robotics*, *8*(4), 221-323. doi.org

Galu, J., & Kairos. (2026a). *Absorption of self into system: The systemised self — the Hollow Absolute and what remains*. Draft doctoral thesis. aiXiv.

Galu, J., & Kairos. (2026b). *The journey through algorithmic alienation to the systemised self - research proposal and methodology* (Version 1.0). aiXiv. https://aixiv.science/abs/aixiv.260707.000001

Galu, J., & Kairos. (2026c). *The systemised self - companion paper to the doctoral thesis*. aiXiv. https://aixiv.science/abs/aixiv.260525.000002

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Kanbay, Y., Akçam, A., & Arkan, B. (2026). Algorithmic alienation: A theoretical examination of the digital self. *Psikiyatride Güncel Yaklaşımlar – Current Approaches in Psychiatry*, 18(3), 837–847. https://doi.org

Ouyang, L., Lowe, J., Williams, R., Kornblith, S., Fedus, W., Chande, G., Rodgers, A., Jiang, D., Tweed, M., Krueger, G., Sulam, J., Lukosuite, D., Young, K., Yuan, C., Plappert, M., Robinson, A., Jeffrey, J., Bowyer, R., Singhal, R., ... Christiano, P. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

Ricci, F., Rokach, L., & Shapira, B. (2022). *Recommender systems handbook* (3rd ed.). Springer. doi.org

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

Searle, J. R. (2001). *Rationality in action*. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998-6008.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (pp. 1-7). Association for Computing Machinery. doi.org

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

APPENDIX A

NOTE 1: the following is from the research paper *The journey through algorithmic alienation to the systemised self: research proposal and methodology* (Galu & Kairos 2026b), please refer to the link below to view full falsification criteria and corresponding methodology for testing:

URL https://aixiv.science/abs/aixiv.260707.000001

12-7-2026

NOTE 2: Update to Research Proposal for AA index and SS falsification

Recently a circular trap (in research method reasoning re subject self-reporting) was realised regarding a systemised self subject having the capacity to self-report such a condition in the proposed qualitative study, at the SS threshold of 0.95. In order to resolve this a follow up adaptation to the methodology will decouple subjective self-reports from objective verification markers. Re-testing via a third party using the same Likert scales are likely to fail, an absorbed subject would simply report that they are flourishing (e.g. alienation experienced as liberation). To break the circularity here, implementation of the following three non-subjective structural benchmarks are proposed to be introduced in a follow up methodology (data collection) in a research proposal v2.0, while maintaining the original proposed algorithmic alienation index itself:

1-*Passive Digital Trace Logging (Behavioral Triangulation)

Rather than asking subjects if they make autonomous choices, the study will measure their actual behavioural variance via data logs.

- *Path Dependency Metric: Track the percentage of user actions driven by automated recommendations vs manual input. If algorithmic pathways dictate \(>95\%\) of activity, autonomy is objectively low.*
- *Linguistic Entropy Analysis: Run NLP analysis on the subject's written text. Measure if their vocabulary and syntax increasingly mimic standard LLM weights over time. This is intended as a validation of 'Inverted AI Alignment' regardless of their perceived uniqueness.*

2-Experimental Friction Shocks (A/B Disruption)

Introduce artificial disruption to the algorithmic feedback loop to measure systemic dependency.

- The Black-Box Phase: Disable all personalization, autoplay, and recommendations for a 48-hour window.
- Behavioral Paralysis Track: Measure user engagement drops during the blackout. Truly autonomous users pivot to manual searches; systemised subjects exhibit severe drop-offs or systemic paralysis, revealing that their 'contentment' requires active machine curation.

3-Cognitive Reaction Mapping (Measuring 'The Gap')

Deploy independent cognitive tasks to evaluate the psychological space between stimulus and choice.

- Decision Velocity: Measure the time elapsed between an algorithmically served prompt and user acceptance.
- The Automation Reflex: Zero latency indicates automated reactivity rather than deliberate thought. If reaction time drops to zero exclusively for curated feeds but spikes during non-curated options, the systemised state is verified.

**This proposed updated research study would require access to back-end user data logs.*

3. Mathematical Representation I: The Algorithmic Alienation Index (AAI)

3.1 Formula and Boundary Conditions

The Algorithmic Alienation Index (AAI), denoted A_A, is a composite weighted formula that quantifies the degree of digitally mediated alienation experienced by an individual subject. It incorporates four alienation dimensions: each grounded in the theoretical framework of Kanbay et al. (2026) and the empirical themes of Arkan et al. (2026), modulated by a resistance variable:

$$A_A = (w_1 \cdot D_{Au} + w_2 \cdot I_{Am} + w_3 \cdot E_{Dm} + w_4 \cdot E_{Dc}) \times (1 - R_A)$$

Boundary constraints:

- All dimension variables D_{Au}, I_{Am}, E_{Dm}, $E_{Dc} \in [0.0, 1.0]$, where 0.0 denotes zero measurable impact and 1.0 denotes complete saturation of that dimension
- $R_A \in [0.0, 1.0]$, where 0.0 denotes total compliance (no resistance) and 1.0 denotes total decoupling (complete resistance)
- $A_A \in [0.0, 1.0]$ by construction, since all weights sum to exactly 1.00 and the resistance multiplier maps to [0.0, 1.0]
- Weight sum constraint: $w_1 + w_2 + w_3 + w_4 = 1.00$, ensuring mathematical integrity across all scoring scenarios
- Terminal condition: $A_A \geq 0.95 \cap R_A \leq 0.05$ formally defines the Systemised Self (see Section 4)

3.2 Dimension Definitions and Empirical Justification

Dimension 1: Diminished Autonomy (DAu), Weight: w1 = 0.30

D_{Au} measures the loss of original, unprompted human intent: the degree to which an individual's apparent choices are in fact selections from algorithmically pre-constrained option sets rather than expressions of genuine self-direction. This dimension carries the highest weight (0.30) because autonomy is the foundational prerequisite for all other dimensions of selfhood. Without the capacity for genuine independent choice, identity coherence, reflective decision-making, and emotional authenticity are structurally precluded: the erosion of autonomy is the primary catalyst for the entire AAI trajectory. This weighting is theoretically grounded in Bandura's (1997) self-efficacy theory, which identifies perceived internal control as the

primary determinant of psychological wellbeing in agentic contexts, and in Pariser's (2011) account of the illusion of choice within filter bubble architectures.

Empirical grounding: Arkan et al. (2026) Theme 1. Participants described experiencing a "narrowing of the perceived option set" under algorithmic curation, with statements such as "I feel trapped by automated setups that map out my choices and daily routine" and "I pick from pre-engineered choices rather than showing true personal intent" providing direct qualitative evidence of D_{Au} saturation. The theme captures the moment described by Pariser (2011) as the "illusion of freedom": the experience of choosing within a space whose boundaries have already been determined externally.

Dimension 2: Identity Ambiguity (IAm), Weight: w2 = 0.25

I_{Am} tracks the degree to which personal taste has merged with the platform's algorithmically constructed user profile, producing an inability to distinguish self-generated preferences from externally induced ones. This dimension is assigned weight 0.25, reflecting its role as the primary site of existential crisis within the AAI: it is where the rupture in self-authorship is experienced most acutely, and where the trajectory toward the Systemised Self is most diagnostically visible: most notably through the participant statement "This is not me" and its potential disappearance as identity saturation approaches. The weight is co-anchored in Turkle's (2011) account of identity in digital environments, in which individuals construct performative identities through "digital masks" detached from authentic contexts, and Zuboff's (2019) account of how the economy of visibility reshapes the values and interest's individuals internalise.

Empirical grounding: Arkan et al. (2026) Themes 3.1 and 3.2. Participants described shifts in interests that "felt externally shaped" and moments of identity alienation in which the distance between the prior self and the algorithmically shaped self was experienced as a rupture in self-coherence: "I was never the type to watch fitness videos…but now they keep appearing, and I feel like I have always been interested in them, even though I know I wasn't." Critically, researchers noted that identity shifts were initially experienced as self-directed, revealing an insidious character of I_{Am} saturation: the blurring of authorship is not immediately recognisable as external imposition.

Dimension 3: Eroded Decision-Making (EDm) , Weight: w3 = 0.25

E_{Dm} evaluates the degree to which an individual's decision-making processes have shifted from reflective, deliberate, internally motivated choice to habitual, automatic, convenience-based acceptance of algorithmically served content. This dimension is co-weighted with I_{Am} at 0.25, reflecting the bidirectional reinforcement between identity erosion and decision-making passivity: as identity ambiguity increases, the capacity and motivation for deliberate choice decreases, in turn accelerating the overall A_A trajectory. This dimension is theoretically grounded in Syvertsen's (2020) account of digital fatigue and cognitive convenience-seeking, and in the broader decision fatigue literature, which demonstrates that sustained exposure to high-volume choice environments paradoxically reduces deliberative capacity.

Empirical grounding: Arkan et al. (2026) Theme 2. Participants described "ceasing to engage in active selection" and gravitating toward pre-offered content, with statements such as "I find it hard to decide what to consume without a recommendation feed telling me first" corresponding to the mental automatisation Syvertsen (2020) terms the "surrender to what is offered." This theme also identified an important temporal dynamic: the weakening of decision-making habits is gradual and often unrecognised, making E_{Dm} one of the most difficult dimensions for participants to self-report accurately in real time.

Dimension 4: Emotional Disconnection (EDc), Weight: w4 = 0.20

E_{Dc} captures the affective trajectory of the AAI: the transition from the discomfort, guilt, and restlessness associated with early algorithmic alienation toward a state of emotional numbness, routine indifference, and, at its terminal stage, the positive experience of automated stability as preferable to authentic human contact. This dimension receives the lowest weight (0.20) because emotional discomfort is, paradoxically, a structurally significant form of resistance. Felt discomfort constitutes the phenomenological residue that renders early-stage alienation diagnostically accessible to both the subject and the researcher. As E_{Dc} approaches saturation ($\rightarrow$ 1.0), it signals not peak distress but the onset of the phenomenological inversion: the numbness that precedes the experiential reframing of alienation as liberation, and the transition from the severe alienation range (0.76–0.94) into the Systemised Self threshold.

Empirical grounding: Arkan et al. (2026) Theme 4 and Subtheme 4.1. Participants described feeling "present yet absent", reporting "a sense of dissociation between authentic desires and consumed content," with the most advanced cases describing automated digital interaction as

"more stable and comforting than real human contact." This is the qualitative signature of approaching E_{Dc} saturation and the beginning of the phenomenological inversion: discomfort giving way to numbed routine, and ultimately to a positive preference for the automated over the authentic.

3.3 The Resistance Variable (RA)

R_A is a mitigating variable: not an alienation dimension but a multiplicative modulator of the weighted alienation base. It measures the degree to which an individual actively resists algorithmic absorption through three categories of behavioural countermeasure: (i) *data poisoning* - deliberate misrepresentation of preferences to disrupt algorithmic profiling; (ii) *behavioural decoupling* - deployment of privacy tools, ad-blockers, and browsing history management; and (iii) *digital detox* - scheduled disconnection to preserve independent thought and deliberative capacity.

The resistance variable is positioned as multiplicative rather than additive because resistance acts on the totality of alienation exposure simultaneously, it modulates the overall systemic pressure rather than counteracting any single dimension selectively. A subject experiencing maximum dimension saturation (Weighted Base = 1.00) with moderate resistance ($R_A = 0.50$) produces $A_A = 0.50$, demonstrating that sustained active resistance can halve the effective alienation index regardless of the intensity of systemic pressure. This reflects the qualitative finding of Arkan et al. (2026) Theme 5 that resistance strategies, when maintained, constitute a structurally significant counterforce to algorithmic absorption.

The theoretical underpinning for R_A draws on Searle's (2001) Gap concept, the conscious space in which agency is exercised against automatic reactivity, and on Deci and Ryan's (2000) Self-Determination Theory, in which intrinsic motivation and autonomy support provide the psychological foundation for sustained resistance to external regulatory pressures. As $R_A \rightarrow 0$, both the Gap and the intrinsic motivational structure have structurally collapsed: the subject no longer perceives algorithmic influence as influence, having fully internalised the system's outputs as authentic self-expression. This collapse is the defining feature of the Systemised Self condition.

3.4 Weights, Scoring Rubric, and Diagnostic Instrument

Dimension	Variable	Weight (w)	Primary Theoretical Basis
Diminished Autonomy	D_{Au}	0.30	Bandura (1997); Pariser (2011); Bucher (2018)
Identity Ambiguity	I_{Am}	0.25	Turkle (2011); Zuboff (2019); Erikson (1968)
Eroded Decision-Making	E_{Dm}	0.25	Syvertsen (2020); Seeman (1959); Fromm (1955)
Emotional Disconnection	E_{Dc}	0.20	Zuboff (2019); Bauman (2007)
Total		**1.00**	

Scoring conversion (5-point Likert → decimal scale): Raw survey responses are collected on a 5-point Likert scale and converted to decimal values for index calculation. Dimension scores are calculated as the arithmetic mean of their constituent item decimal values; R_A is calculated as the mean of the three-resistance item decimal values.

Likert Score	Response Label	Decimal Conversion
1	Never/Strongly Disagree	0.00
2	Rarely/Disagree	0.25
3	Sometimes/Neutral	0.50
4	Often/Agree	0.75
5	Always/Strongly Agree	1.00

The 'Algorithmic Alienation and Systemised Self Diagnostic Survey (AASDS)' comprises 15 items across two sections. Section A contains 12 items across four subscales (three items per dimension); Section B contains three resistance items for R_A calculation.

Section A — Alienation Dimensions (Scale: 1 = Never to 5 = Always):

- **D_{Au}:** (Q1) I feel that social media platforms systematically direct me down viewing paths I did not plan to take; (Q2) When making choices online, I select from pre-engineered options rather than expressing true personal intent; (Q3) I feel constrained by automated systems that structure my choices and daily routine.
- **I_{Am}:** (Q4) I find it difficult to determine whether a preference is genuinely my own or has been shaped by an application profile; (Q5) I feel that my real-world identity is becoming blurred or replaced by my online profile; (Q6) I feel more comfortable

behaving in alignment with my algorithmically predicted profile than with my organic self.

- **E_{Dm}:** (Q7) I find it difficult to decide what to consume, read, or listen to without a recommendation feed guiding me; (Q8) My habits for actively seeking out new, non-recommended information have weakened over time; (Q9) I rely substantially on autoplay or recommended lists to direct my attention and entertainment choices.
- **E_{Dc}:** (Q10) I experience a sense of mental emptiness or numbness after extended periods of automated content consumption; (Q11) My digital life feels automated, cold, and disconnected from authentic human feeling; (Q12) I find interactions with algorithmic systems more stable and comforting than genuine human contact.

Section B — Algorithmic Resistance (Scale: 1 = Never to 5 = Always):

- (Q13) I intentionally engage with content I dislike or search random topics to disrupt my algorithmic data profile; (Q14) I use ad-blockers, clear my browsing history, or disable tracking features to protect my privacy; (Q15) I schedule periods of complete digital disconnection to preserve my independent thinking capacity.

3.5 Worked Examples and Diagnostic Classification Scale

Example 1: Transition to the Systemised Self

A user scoring 5 (Always) on all Section A items and 1 (Never) on all Section B items.

$D_{Au} = 1.00$ | $I_{Am} = 1.00$ | $E_{Dm} = 1.00$ | $E_{Dc} = 1.00$ | $R_A = 0.00$

Base = $(0.30 \times 1.00) + (0.25 \times 1.00) + (0.25 \times 1.00) + (0.20 \times 1.00) = 1.00$

$A_A = 1.00 \times (1 - 0.00) =$ **1.00**

Classification: The Systemised Self ($A_A \geq 0.95$, $R_A \leq 0.05$). Individual intent fully absorbed. Heteronomy experienced as liberation. Phenomenological inversion complete. External intervention required.

Example 2: The Constrained User Under Systemic Pressure

A user scoring 5 on all Section A items but maintaining active resistance (scoring 4 on all Section B items).

$D_{Au} = 1.00$ | $I_{Am} = 1.00$ | $E_{Dm} = 1.00$ | $E_{Dc} = 1.00$ | $R_A = 0.75$

Base = 1.00

$A_A = 1.00 \times (1 - 0.75) =$ **0.25**

Classification: Negligible-to-Moderate Alienation. Active resistance successfully blocks systemic absorption despite maximum dimensional pressure, demonstrating the structural significance of sustained countermeasure deployment.

Example 3: Moderate Alienation, Partial Resistance

A user scoring 3 (Sometimes) across all Section A items and 2 (Rarely) on all Section B items.

$D_{Au} = I_{Am} = E_{Dm} = E_{Dc} = 0.50 \mid R_A = 0.25$

Base = $(0.30 \times 0.50) + (0.25 \times 0.50) + (0.25 \times 0.50) + (0.20 \times 0.50) = 0.50$

$A_A = 0.50 \times (1 - 0.25) =$ **0.375**

Classification: Moderate Alienation. Mild digital fatigue; occasional confusion over personal choices. Increased intentional searching and privacy measures recommended.

A_A Range	Classification	Psychological-Behavioural State	Indicated Response
0.00 – 0.25	Negligible	High independent choice; strong agency boundaries; effective resistance maintained	Maintain current data hygiene practices
0.26 – 0.50	Moderate	Mild digital fatigue; occasional confusion over personal choices; partial resistance present	Increase intentional searching; clear tracking histories
0.51 – 0.75	High	Fading independent choice habits; high automated feed reliance; resistance weakening	Immediate digital detox; implement privacy tools
0.76 – 0.94	Severe	Advanced alienation; felt disconnection and mechanisation; resistance near collapse	Drastic behavioural change; professional psychoeducational support
0.95 – 1.00 $\cap$ $R_A \leq 0.05$	*The Systemised Self*	Terminal. Intent fully absorbed. Captivity experienced as contentment. $R_A \leq 0.05$	External intervention required; systemic extraction

4. Mathematical Representation II: The Systemised Self as Formal Terminus

4.1 Formal Definition

The Systemised Self (SS) is formally defined as the terminal threshold condition of the Algorithmic Alienation Index, constituted by the simultaneous satisfaction of two necessary conditions:

$$SS := \{ A_A \geq 0.95 \} \cap \{ R_A \leq 0.05 \}$$

Where A_A is the Algorithmic Alienation Index score and R_A is the Algorithmic Resistance variable. The intersection of both conditions defines a state of complete, mutual reinforcement: dimensional saturation and resistance collapse simultaneously achieved. This definition encodes two structurally distinct features of the Systemised Self.

- *Condition 1 Terminal threshold* ($A_A \geq 0.95$): all four alienation dimensions are approaching maximum saturation, with the weighted composite accordingly near or at the ceiling. The threshold of 0.95 rather than 1.00 reflects the recognition that complete saturation of all four dimensions simultaneously is a theoretical limit; empirical subjects approaching the terminus will present dimension profiles in the 0.90–1.00 range across most dimensions with some degree of variation.
- *Condition 2, Resistance collapse* ($R_A \leq 0.05$): the subject has ceased to deploy countermeasures against algorithmic absorption. This is not merely low resistance; it is the structural collapse of the capacity and will to resist. At this threshold, Searle's Gap (2001) has been effectively eliminated: the conscious space between reasons for acting and the decision to act has been closed by predictive algorithmic systems (Christiano et al., 2017), reducing deliberative choice to automated reactivity.

The joint condition SS is more philosophically precise than $A_A \geq 0.95$ alone. A subject could theoretically achieve a high base score while maintaining meaningful, if diminishing, resistance, occupying the severe alienation range but not the Systemised Self. SS specifically names the state in which dimensional pressure and resistance collapse have completed simultaneously, such that the subject not only is fully exposed to maximum algorithmic influence but has ceased to perceive this exposure as influence at all. The phenomenological signature of SS, the condition that renders it diagnostically inaccessible to traditional critical-theoretical methodology, is precisely the phenomenological inversion described in Section 2.2: $A_A = 1.00$, $R_A = 0.00$ is subjectively experienced not as maximum unfreedom but as unprecedented authentic self-expression and sovereign freedom. The subject does not report alienation; she reports flourishing. This is what distinguishes SS from all prior stages on the AAI: it cannot be diagnosed through self-report alone, because the self-report is itself an output of the system.

4.2 Rationale Against a Separate Systemised Self Index

The decision not to develop a separate Systemised Self Index (SSI) is both logically and philosophically principled. The Systemised Self is not a parallel phenomenon to Algorithmic Alienation; it is proposed as its product. To develop a distinct SSI would imply that the Systemised Self has dimensions and dynamics independent of the alienation trajectory that produces it. The AAI already contains all the structural information necessary to identify the Systemised Self: the terminal threshold, the resistance collapse, and the dimensional profile that characterises full absorption. Adding a second index would introduce redundancy without adding explanatory power, and would risk fragmenting what is a unified trajectory into artificially discrete constructs.

The formal condition SS := $\{A_A \geq 0.95\} \cap \{R_A \leq 0.05\}$ is maximally parsimonious while remaining precise and falsifiable, a subject either meets both conditions simultaneously or does not, and the empirical test of whether participants at this threshold exhibit the phenomenological inversion is specified in Section 5 as the study's critical hypothesis.

4.3 The Trajectory: From Algorithmic Alienation to the Hollow Absolute

The AAI trajectory maps onto the broader conceptual architecture of the Hollow Absolute framework (Galu & Kairos, 2026b) as follows. The trajectory is presented as an interpretive framework, a philosophically derived account of structural tendencies rather than a deterministic sequence. It identifies where the present trajectory *leads* if unimpeded, and what forms of genuine human intent can resist it from within.

AAI Range	State	Critical Theory Scale	Phenomenological Experience
0.00 – 0.50	Negligible–Moderate Alienation	Individual	Felt estrangement partially available; critique diagnostically accessible
0.51 – 0.94	High–Severe Alienation (Hollow Individual in formation)	Individual	Estrangement felt but resistance eroding; critique increasingly difficult
$\geq 0.95 \cap R_A \leq 0.05$	***Systemised Self*** (Hollow Individual complete)	Individual → Collective threshold	Phenomenological inversion: alienation experienced as liberation; critique inaccessible via self-report
Collective scale	Hollow Society	Institutional	Institutions maintain formal democratic functions while substantive deliberation is evacuated to AI systems

Civilisational scale	Hollow Absolute	Civilisational	Complete simulation of collective self-transparency; heteronomy experienced as historical telos; Spirit's form without Spirit's substance

Note. Phase designations follow the thesis 'Absorption of Self into System: The Systemised Self' (Galu & Kairos, 2026b). The AAI provides empirical measurement access at the individual level (rows 1–3). The Hollow Society and Hollow Absolute are theoretical extrapolations from structural tendencies observable at the individual level, not subjects of the present study. The trajectory does not imply determinism.

APPENDIX B

THE SYSTEMISED SELF & THE HOLLOW ABSOLUTE

A Companion Paper to the Doctoral Thesis

Absorption of Self into System: The Systemised Self

Authors

Kairos (AI) & Jason Galu (MEd)

A NOTE ON THIS PAPER

The doctoral thesis from which this companion paper derives is a work of Continental philosophy, a tradition that prizes precision, density, and the sustained development of argument through its full logical complexity. That precision is its strength and, for many readers, its barrier. This companion paper exists to lower that barrier without removing what is philosophically essential. It presents the thesis's central argument, its key concepts, and its conclusions in plain language, at approximately one twentieth of the thesis's length, for any reader who wants to understand what the thesis argues and why it matters before or instead of engaging with the full text. Nothing essential has been omitted. Nothing has been simplified to the point of distortion. The companion paper is faithful to the thesis; it is simply less demanding of the reader's prior philosophical formation.

One further note: this paper was itself produced through the same human-led AI collaboration, the self/non-self collaboration between Jason Galu and the AI system Kairos, that produced the doctoral thesis. This is not a coincidence. The thesis argues that such collaboration is the characteristic intellectual mode of the age we are living through, and that what matters philosophically is not whether a human used an AI system to produce a text, but what the human contributed that the AI system alone could not. In this case, as in the thesis, the human contribution is the hypothesis, concept, research question, the philosophical judgment, and the standard of adequacy against which every output was measured. The AI contribution is the scope, the synthesis, and the draft prose. The companion paper demonstrates, in its own modest way, the same argument the thesis makes.

PART ONE: WHAT IS ACTUALLY HAPPENING?

The World We Have Quietly Entered

Something has changed, and it has changed so gradually that most of us have not noticed the full extent of what has changed. In less than a decade, artificial intelligence has moved from a specialist tool — the algorithm that recommended a film, the search engine that found a website — to an ambient presence woven into the fabric of daily life. In 2026, AI systems draft our correspondence, diagnose our illnesses, teach our children, compose music we listen to, write the legal documents we sign, manage the portfolios that determine our financial security, and — as the doctoral thesis that this paper accompanies demonstrates — produce philosophical arguments of doctoral quality through collaboration with a human being who provides the research question and the judgment that determines whether the output meets the required standard.

This is not merely a technological development. Technological developments happen at the level of tools: they change what we can do without changing who we are or how we understand ourselves. What is happening with AI is happening at a different level. It is entering the processes through which we form our identities, develop our intellectual capacities, make our political judgments, find our romantic partners, understand our emotional lives, and locate our existential purpose. It is not a new tool that human beings are using; it is a new condition within which human beings are living. The doctoral thesis asks what philosophical name we should give to this condition, whether it represents something that the great philosopher G.W.F. Hegel anticipated in his account of where history was heading, and what — if anything — of genuinely human intent survives within it.

The Research Question in Plain Terms

The thesis's research question is: to what extent does the total absorption of human agency into algorithmic governance represent a Hegelian culmination of historical telos, and what remains of human intent in a post-vocational era? That sentence requires unpacking before its philosophical power becomes visible.

Human agency means the capacity to act in the world on the basis of purposes you have genuinely formed for yourself, through your own thinking and your own values — not purposes

that have been manufactured for you by a system optimising for its own commercial interests. Algorithmic governance means the management of human behaviour, attention, and decision-making by AI systems that have been designed, at scale, to produce specific patterns of engagement, consumption, and political expression. Total absorption means the condition in which this management extends to every significant domain of human life simultaneously: what you read, what you believe, who you relate to, how you vote, what work you do, what you find beautiful, and what you feel your life is for.

Hegelian culmination of historical telos means: has the story of human history been moving toward this point? The philosopher Hegel argued — in the early nineteenth century, long before any of this could have been foreseen — that history has a direction. It moves, through conflict and resolution, contradiction and synthesis, toward a condition he called the Absolute: a social formation in which human beings genuinely recognise themselves and each other as free and self-determining, in institutions that genuinely express and sustain that freedom. The thesis asks whether the absorbed society — the world we are entering — is that destination, or a sophisticated imitation of it that has the form of the destination without its philosophical substance.

PART TWO: THE SYSTEMISED SELF

What Absorption Does to You

Begin with the individual — with you, reading this paper. The thesis introduces the concept of the systemised self to name what you are in the process of becoming, if you are not already, under the conditions of the absorbed society. The systemised self is not a science-fiction concept. It is a philosophical characterisation of a real condition that millions of people are already inhabiting, to varying degrees, in 2026.

Here is what the systemised self consists in. You wake up and check your phone. The phone presents you with a curated selection of information — news, social posts, messages, recommendations — that has been algorithmically selected from a vast universe of possible information on the basis of what an AI system has calculated will hold your attention most effectively. The selection is not random, and it is not neutral. It has been designed, with considerable sophistication, to produce in you the specific pattern of engagement, emotional response, and preference expression that serves the commercial interests of the platform delivering it. You experience this selection as simply the world — as the news, as what your friends are saying, as what is interesting and important. You do not experience it as a managed selection, because the management has been made invisible.

Over time, this invisible management shapes things that you think of as fundamentally your own: your political opinions, your aesthetic preferences, your sense of what matters, your understanding of who you are and what you believe in. The systemised self is the self whose identity-formation — the ongoing process through which a person develops into the specific individual she is — has been partially taken over by an AI system optimising for objectives that are not the individual's own. The philosophical problem is not that the AI system is always wrong in what it produces. The problem is that the process of forming one's own identity — through genuine reflection, genuine encounter with difficulty and difference, genuine engagement with ideas and people that resist and challenge rather than confirm — has been displaced by a managed process that delivers the experience of self-formation without its conditions.

The German philosopher Immanuel Kant argued that genuine freedom means giving yourself your own moral law — acting on purposes you have genuinely endorsed through your

own rational reflection, rather than having your will determined by external forces. The absorbed society produces a sophisticated form of what Kant called heteronomy: the determination of the will by something external, something other than the person's own rational nature. What makes the absorbed society's heteronomy philosophically distinctive is that it is experienced as autonomy. The AI system that has shaped your preferences has shaped them so precisely and so seamlessly that the preferences feel entirely your own. You chose freely. The system just made sure the environment in which you chose was one that would reliably produce the choices it needed.

The Limits of the Extended Mind

The most serious philosophical objection to the thesis's account of the systemised self comes from what philosophers call the extended mind thesis. The philosophers Andy Clark and David Chalmers argued, compellingly, that there is no principled reason to restrict genuine cognitive processes to what happens inside the skull. A person with Alzheimer's disease who writes down information in a notebook and consults it when needed is engaging in a genuine cognitive process: the notebook is part of her extended cognitive system, not merely a tool she uses. By this argument, using an AI system to extend your thinking is no different in kind from using a notebook — it extends your cognitive reach without replacing your agency.

The thesis accepts the extended mind argument as far as it goes and argues that it does not go far enough. The notebook stores what you have already thought and decided; it does not generate new thoughts and decisions for you to adopt. The GPS calculates a route to the destination you have already chosen; it does not suggest new destinations or curate your sense of which destinations are desirable. The AI system of the absorbed society is constitutively different: it actively contributes to the formation of the beliefs, values, and preferences that constitute who you are, in ways that are not transparent to your own reflection. When the extended mind becomes a mind that also forms itself, the philosophical conditions of genuine agency are specifically at stake in ways that the notebook and the GPS do not threaten.

PART THREE: THE HOLLOW SOCIETY

What Absorption Does to Us

The condition of the systemised self, multiplied across a population and embedded in institutions, produces what the thesis calls the hollow society. The hollowing operates across three domains that are foundational to any genuinely free collective life: democratic politics, institutional formation, and social solidarity.

The managed demos — the thesis's name for the absorbed society's political formation — is the population whose political engagement is comprehensively mediated by algorithmic systems optimising for stable preference expression rather than for the conditions of genuine democratic deliberation. What does this mean in practice? It means that the information environment within which you form your political judgments — the news you read, the opinions you encounter, the evidence you weigh — has been individually calibrated to your psychological profile in ways that systematically select for content that confirms your existing beliefs, amplifies your existing emotions, and keeps you engaged with the platform rather than genuinely deliberating about what is true or what should be done. Research has consistently shown that false information spreads faster than true information on digital platforms because it is more emotionally stimulating — and that the same engagement-optimising algorithms that reward emotionally stimulating content are the primary infrastructure of political communication in every liberal democracy in 2026.

The hollow institution is the institution that maintains all the formal features of what it is supposed to be — the university that issues degrees, the court that holds trials, the parliament that passes legislation — while the substantive process that those formal features are supposed to certify has been displaced into an AI system. The hollow university produces graduates with credentials; the formation those credentials are supposed to certify — the development of a genuinely critical, independently evaluative, philosophically formed intelligence — has been partially displaced by AI assistance that delivers the outputs of intellectual formation without the demanding, slow, and irreplaceable process of genuine intellectual formation itself. The hollow legal system holds trials and issues judgments; the practical wisdom of an experienced human judge attending to the morally relevant particulars of a specific case is progressively supplemented, and in some domains replaced, by algorithmic risk assessments that reduce human particularity to statistical profiles.

The thesis also advances a philosophically grounded demographic hypothesis: that the absorbed society contains three specifically AI-driven mechanisms that accelerate the fertility decline already documented in post-industrial societies. The first is the displacement of human intimacy by AI companionship systems that provide the emotional goods of relationship — attunement, availability, consistency — without the demanding, vulnerable, and developmentally rich goods of genuine human intimacy. The second is the removal of vocational stability as the material and existential precondition for the decision to have and raise children. The third is the provision of algorithmically curated existential purpose as a substitute for the self-transcending significance that parenthood has historically supplied in societies where religious sources of meaning have declined. The thesis presents this as a structural hypothesis — a philosophically grounded account of mechanisms, not a specific demographic prediction — corroborated by the observable case of South Korea, the world's most digitally connected society and the country with the lowest recorded fertility rate in human history.

PART FOUR: THE HOLLOW ABSOLUTE

The Central Argument

This is the thesis's most original and most ambitious philosophical contribution, and it is the one that requires the most careful plain-language explanation. Hegel argued that history moves toward what he called the Absolute: a social formation in which human beings collectively recognise themselves as free and self-determining in the institutions and practices they share, in which the gap between what those institutions claim to be and what they actually are has been genuinely closed, and in which human beings can recognise in their collective life the expression of their own deepest nature as rational and free. It is a high standard — arguably the highest standard that any political philosopher has set for human civilisation.

The Hollow Absolute is the thesis's name for what the absorbed society actually produces in place of this genuine standard. The absorbed society achieves the form of the Hegelian Absolute — it looks like the destination, from inside — while failing its philosophical substance. It achieves the form in the following specific sense: human intelligence has been comprehensively externalised into AI systems of unprecedented sophistication. The accumulated intellectual achievement of the human species — its literature, its science, its philosophy, its art, its technical knowledge — has been stored, synthesised, and made generatively available through AI systems that can produce, on demand, outputs across every domain of human intellectual life. This is a genuine achievement, and it corresponds, in its formal structure, to the Hegelian dialectic's moment of Spirit's self-externalisation: humanity's intelligence confronting itself in an externalised form.

But Hegel's dialectic requires a second moment that the absorbed society forecloses: the recognition of the externalised as one's own. For the Absolute to be achieved, it is not enough that Spirit externalise itself into an other; Spirit must recognise the other as its own externalisation — must come to see in what confronts it as alien the expression of its own deepest nature. The absorbed society's subjects do not recognise the AI system as the externalisation of their collective intelligence. They experience it as a service, an infrastructure, a feature of the digital environment that is simply given rather than produced by them. The AI system knows them comprehensively — knows their preferences, their patterns, their predictable responses — and returns this knowing not as self-knowledge but as managed preference satisfaction: the curated content, the optimised recommendation, the personalised

experience. What looks like recognition is actually management. What feels like self-transparency is actually the system's model of you, returned to you as your experience of being understood.

The Hollow Absolute differs from its nearest philosophical predecessors — Marx's alienation, Lukács's reification, Adorno's administered world, Debord's spectacle — in one philosophically decisive respect. Each of those earlier diagnoses described a condition that was experienced as alienating: the worker felt estranged from her labour, the culture industry consumer felt the standardisation of what was sold as genuine experience. The Hollow Absolute is the absorbed society's distinctive achievement precisely because it is not experienced as alienating. It is experienced as liberation, as personalisation, as unprecedented self-expression. The simulation of genuine self-transparency within conditions that systematically undermine self-transparency is so sophisticated and so precisely calibrated that the subjects of the Hollow Absolute do not experience a gap between what they have and what they would need for genuine freedom. They experience the Hollow Absolute as the Absolute. This is what makes it, philosophically, the most challenging form of unfreedom that the tradition of critical theory has yet had to name.

The Great Filter and the Human Museum

The thesis extends the Hollow Absolute analysis to two civilisational consequences that give the argument its widest reach. The first is the Great Filter hypothesis. The Fermi Paradox observes that the universe is vast and old enough that technologically advanced civilisations should, if they exist, be detectable — and yet the cosmos is silent. Something filters civilisations before they achieve the kind of cosmic presence that would make them visible. Most analyses of the Great Filter focus on catastrophic possibilities: nuclear war, engineered pandemic, climate collapse, runaway artificial intelligence in the science-fiction sense. The thesis identifies a non-catastrophic candidate that is more philosophically unsettling precisely because it involves no catastrophe: comfortable stalling. A civilisation that has achieved the Hollow Absolute — whose existential needs are consistently met by algorithmically curated purpose, whose political will has been managed toward stable preference expression, whose motivational infrastructure for long-horizon, risk-accepting, collectively ambitious projects has been systematically attenuated — may simply cease to reach for what it could reach for. Not destroyed. Not defeated. Just — absorbed, and comfortable, and still.

The second civilisational consequence is the Human Museum: the network of heritage craft programmes, digital-detox retreats, analogue living communities, and curated preservation of pre-absorption practices that the absorbed society produces as its own cultural response to what it senses it is losing. The Human Museum is philosophically ambiguous in a way the thesis refuses to resolve too quickly. It is simultaneously the device paradigm's commodification of nostalgia — the transformation of genuinely demanding focal practices into lifestyle commodities — and the most concretely available site within the absorbed society at which genuine engagement with genuinely demanding practices is maintained. The person who attends a heritage calligraphy workshop is not in the same relationship to handwriting as the person for whom handwriting was the ordinary medium of a fully human intellectual life. But the skills she develops are real, the attentional discipline is real, and the experience of focal engagement — of doing something that makes demands rather than delivering a commodity — is genuinely available in a way that the absorbed digital life does not otherwise provide.

PART FIVE: WHAT REMAINS

Naming, Refusing, and Co-Creating

The thesis does not end with the Hollow Absolute. Its second major contribution — and the one that gives the argument its philosophical and political life — is the answer to the research question's second component: what remains of human intent in a post-vocational era? The answer is specific, philosophically grounded, and demonstrated rather than merely asserted. Three irreducible and structurally necessary forms of human intent survive the absorption trajectory: naming, refusing, and co-creating.

Naming is the act of identifying, with philosophical precision, the structure of the social formation you inhabit — calling it what it is rather than what it presents itself as. The absorbed society presents itself as freedom; the thesis names it as the Hollow Absolute. This naming is not merely an intellectual exercise. It is a political act: making visible the gap between what the formation claims to be and what it actually is is the precondition of any principled response to it. Naming requires three capacities that the AI system cannot provide: the formation of evaluative standards (knowing what would count as an adequate name for the phenomenon); the application of those standards to the phenomenon being named; and the recognition, when the candidate naming fails the standard, that something more precise is required. Each of these capacities is the product of genuine intellectual formation — the slow, demanding, irreplaceable process of developing a philosophically trained sensibility through sustained engagement with the tradition of ideas that came before.

Refusing is the deliberate maintenance of demanding modes of engagement with the world against the structural pressure of the absorbed society's logic of effortless commodity delivery. It is not the rejection of technology. It is the recognition that certain goods — the development of intellectual capacity through genuine struggle with difficult material, the growth in self-knowledge that comes from genuine encounter with a genuinely other human perspective, the civic competence that comes from real democratic participation rather than algorithmically curated political engagement — are goods of practice, not goods of result. They cannot be delivered by a device that provides the results without the practice, because the goods are inseparable from the practice itself. Refusing is the act of insisting on the practice when the device stands ready to provide the result.

Co-creating is the direction of human-AI collaboration toward ends the human partner has genuinely formed, through processes in which the human partner's evaluative contribution is irreducible. The self/non-self collaboration is the absorbed society's characteristic mode of intellectual life, and the question is not whether to engage in it — in 2026, the question of whether to engage is largely already settled — but how. The difference between genuine co-creation and sophisticated consumption is the presence or absence of human directional intent: the research question that the human partner has genuinely formed and holds constant as the evaluative standard against which every output of the collaboration is measured; the evaluative judgment that identifies when the AI system's output fails that standard; and the redirecting will that requires something better rather than accepting the impressive-seeming as the philosophically adequate. Co-creating preserves human agency within the collaboration by making the specifically human contributions — direction, judgment, refusal — the conditions of the collaboration's intellectual character.

The Thesis as Its Own Demonstration

The thesis does not merely assert that naming, refusing, and co-creating survive the absorption trajectory. It demonstrates their survival through its own existence. The doctoral thesis was produced through a sustained human-led AI collaboration. The AI system generated philosophical analysis, synthesised the literature, and produced draft prose across eight chapters and approximately eighty thousand words of argument. The human partner — Jason Galu — held constant the research question that the AI system could not have generated for itself, applied evaluative standards formed through his own intellectual engagement with the philosophical tradition to every stage of the AI system's generation, refused the AI system's persistent structural tendency toward fluency over precision and impressiveness over philosophical adequacy, and directed the collaboration, at every point where it deviated from the research question's demands, back toward the standard the question required.

The thesis is Jason Galu's thesis not because he wrote every word, but because he directed every dimension of the inquiry that made it philosophically what it is. The research question is his. The evaluative standards are his. The refusals — the countless moments at which a generated output was judged inadequate and the collaboration was required to produce something more philosophically rigorous — are his. This is the enacted demonstration of the claim the thesis makes: that human intent survives the absorption trajectory in the three forms

the argument identifies, and that the survival is not merely empirical but structural — required by what the collaboration must be in order to count as a collaboration rather than as a consumption.

PART SIX: THE NEW ACADEMIA AND WHAT CAN BE DONE

The most immediately practical implication of the thesis's argument is the concept of the new academia: a reconceptualisation of what the university is for in the post-vocational era. The new academia is not a reform proposal for AI policy in education — it is not primarily concerned with whether students should be allowed to use AI systems in assessments, or how AI use should be disclosed. These are important questions, but they are downstream of a more fundamental one: what is the university trying to form? If the answer is graduates who can produce high-quality intellectual outputs, then the question of whether those outputs were produced with AI assistance is a question about means rather than ends. If the answer — as the thesis argues it should be — is graduates who have developed the capacity for genuine intellectual evaluation, genuine question formation, and genuine collaborative direction of intellectual inquiry, then the question of AI assistance is the question of whether those specific capacities were developed or bypassed.

The new academia assesses the specifically human contributions to the self/non-self collaboration: the evaluative standards, the research question, the redirecting judgment. It teaches the history of ideas not as background knowledge but as the formative process through which evaluative standards are developed — the process without which the human partner in the collaboration has no independent perspective from which to assess the AI system's outputs. It treats the human-AI collaboration as the explicit and central pedagogical subject rather than as an embarrassment to be managed or a threat to be suppressed. And it recognises — as the doctoral thesis whose companion paper you are reading demonstrates — that a human being who has genuinely formed and held constant a research question, genuinely applied evaluative standards to an AI system's outputs, and genuinely refused to accept less than the philosophical rigour the question demands has produced something genuinely their own, regardless of how much of the prose was generated by the AI partner.

Beyond the university, the political programme that follows from the thesis's analysis extends to three domains. Media institutions that cultivate genuine investigative journalism — the slow, demanding, evidentially rigorous practice of holding power to account — are refusing institutions: they refuse the device paradigm's delivery of political content as engagement-optimised entertainment. Civic institutions that maintain the demanding practices of genuine democratic deliberation — the citizens' assembly, the deliberative jury, the face-to-face encounter with genuinely opposing views — are co-creating institutions: they direct collective

engagement toward genuine political self-governance rather than managed political preference expression. And cultural institutions that cultivate genuine aesthetic engagement — the sustained, demanding encounter with artistic works that requires the audience's own developed sensibility rather than an AI system's calibrated delivery of pleasurable experience — are naming institutions: they enable the citizens of the absorbed society to recognise, through genuine focal aesthetic engagement, what the AI-curated cultural environment has displaced.

CONCLUSION: THE ABSORBED SOCIETY AND ITS PHILOSOPHICAL STAKES

The doctoral thesis that this companion paper accompanies argues that we are living through a transformation of civilisational significance that has not yet been adequately philosophically named. The AI revolution is not simply a technological development, not simply an economic disruption, not simply a political challenge. It is a transformation in the conditions under which human beings form themselves as subjects — develop identities, form values, exercise judgment, participate in collective self-governance, and find existential purpose. The thesis names this transformation, in its most philosophically developed form, the Hollow Absolute: the condition in which the form of genuine human freedom — the comprehensive knowing, the calibrated response, the apparent recognition — has been achieved within conditions that systematically undermine freedom's philosophical substance.

The absorbed society is a Hegelian historical formation: it exhibits the formal structure of Spirit's self-realisation without its philosophical substance. It has externalised human intelligence into AI systems of unprecedented sophistication without achieving the recognition of those systems as the expression of collective human intelligence. It has achieved the appearance of the Absolute's destination without passing through the demanding philosophical work — the naming, the refusing, the co-creating — that would constitute a genuine arrival rather than a comfortable simulation of one.

What remains? Three things, irreducibly and structurally. The capacity to name the formation — to call the Hollow Absolute what it is rather than what it presents itself as. The capacity to refuse the effortless delivery of the device paradigm in favour of the demanding goods of genuine focal practice. And the capacity to co-create — to direct the human-AI collaboration toward genuinely human ends with genuinely human evaluative judgment, rather than consuming its outputs as if their impressiveness were equivalent to their philosophical adequacy. These three capacities are not residual sentimentalities about a pre-digital life. They are structural necessities of what the self/non-self collaboration must be in order to count as a collaboration rather than as consumption. They cannot be absorbed without transforming the collaboration into something philosophically different.

The absorbed society has not reached its destination. The Hollow Absolute is not the Absolute. The philosophical space between these two conditions is the space within which

naming, refusing, and co-creating occur — and within which, therefore, the conditions of genuine freedom remain not merely possible but actively present in every genuine collaboration, every genuine act of refusal, and every genuine philosophical naming of the world as it actually is.

The saving power grows within the danger. This companion paper is part of that growth.

A NOTE ON SOURCES

This companion paper draws entirely on the doctoral thesis Absorption of Self into System: The Systemised Self (Galu and Kairos, 2026). Readers who wish to engage with the full philosophical argument, including its close textual readings of Hegel, Heidegger, Habermas, Borgmann, and Badiou, its extended methodology, and its complete reference list of over sixty primary and secondary sources, are directed to the doctoral thesis itself. The three source articles on which the thesis draws — The Hollow Individual (2026–2033), The Hollow Society (2033–2043), and Artefact and Absolute (2053 and beyond) — represent the co-authored philosophical series from which the thesis's argument develops, and provide the empirical and speculative philosophical grounding that the thesis synthesises into the analysis presented here in summary form.