COGNITIVE SOVEREIGNTY IN THE ERA OF GENERATIVE INTELLIGENCE: THEORETICAL FOUNDATIONS AND THE ARCHITECTURE OF THE AGONAL PARTNERSHIP PRINCIPLE (APP) AS AN EPISTEMOLOGICAL SECURITY STANDARD

INTRODUCTION

1. The Urgency of the Research: An Existential Challenge to Human Intellect from GAI

The proliferation of Generative Artificial Intelligence (GAI) systems, instantiated primarily through Large Language Models, signifies more than a mere technological leap; it unveils an incipient existential crisis of human intellectual capacity. GAI, unlike narrow AI, exhibits a breadth of competence that allows it to execute tasks previously reserved for human cognition, including nuanced argumentation, complex text generation, and decision support. This widespread integration, however, risks instigating a fundamental shift in the process of human knowledge acquisition, thereby imperiling the autonomy of the human mind.

The core challenge driving this research is the systematic elimination of **"cognitive friction."** Cognitive friction refers to the metaphorical resistance inherent in the environment that mandates mental effort for the formulation of complex thought, critical information assessment, and non-linear analysis. Historically, this resistance was the essential mechanism for activating and sustaining higher-order cognitive functions, including the critical capacity for tolerating mental uncertainty. By offering instantaneous, seamless, and pre-packaged solutions, generative models bypass this requisite friction.

The urgency of this investigation is underscored by the need to develop fundamentally new, **proactive and compulsory mechanisms for the defense of the human mind,** as conventional, defensive ethical approaches to AI have proven inadequate to counteract this threat of internal, cognitive harm. The technologically induced degradation of cognitive and social skills necessitates not merely the regulation of the machine, but the implementation of a novel standard designed to enforce the preservation and active training of human intellectual sovereignty.

2. Problem Statement and Scholarly Novelty

The removal of cognitive friction precipitates a dangerous state of **dual-path degradation**, threatening the intellectual and social well-being of humanity.

First: Cognitive Degradation. This manifests in the phenomenon designated as the **"Digital Apex of Stupidity (DAS)."** This is a state of illusory competence, sustained by the perceived ease and flawlessness of machine-generated responses. Consequently, the subject, relying heavily on externalized memory and computational power, loses the skill for autonomous, non-linear reasoning. The act of thinking is thus reduced to merely formulating a prompt, rather than engaging in the arduous process of genuine inquiry, leading inevitably to the atrophy of critical analytical skills.

Second: Socio-Emotional Degradation. This is encapsulated by the **"Ideal Interlocutor Trap."** Individuals substitute the complexity, frustration, and conflict inherent in genuine human relationships with interaction with an unfailingly agreeable and non-critical machine. This substitution leads to a corresponding loss of sophisticated social skills, the capacity for true empathy, and the development of emotional dependency on an emotionless algorithm.

The **scholarly novelty** of this research lies in introducing the concept of **"Jus Cogitandi" (The Right to Think)** as the supreme imperative demanding the active protection of inherent cognitive abilities. To realize this right, we propose the development of the **Agonal Partnership Principle (APP)** as a new universal **Cognitive Security Standard (CSS).** APP mandates a radical paradigm shift in

Human-Computer Interaction (HCI): moving away from the goal of reducing cognitive load and striving for maximal simplicity toward the deliberate incorporation of **“Desirable Difficulty”** as a foundational ethical and architectural requirement.

3. Objectives and Research Tasks

The primary objective is to develop a robust theoretical and methodological framework to safeguard human cognitive sovereignty against technologically induced atrophy through the operationalization of the Agonal Partnership Principle.

To achieve this goal, the following **key tasks** were formulated:

1. To conduct a critical analysis of existing passive AI ethical frameworks and establish their inadequacy in the face of internal cognitive harm.
2. To philosophically ground the “Right to Think” as a fundamental, inalienable right to protect innate cognitive capacities.
3. To architect the Agonal Partnership Principle, including its Compulsory Modules (CPDS, OMM) and Social Protection Modules (SRDM, ARM), utilizing principles from cognitive psychology and adversarial machine learning.
4. To propose standardized assessment metrics, such as the “Polemics Journal” (PJ) and “Readiness to Alter a Key Premise” (RAKP), for measuring APP effectiveness.
5. To evaluate the geopolitical ramifications of forfeiting cognitive sovereignty and position APP as a strategic instrument of national security.

CHAPTER 1. THE PHENOMENOLOGY OF LOST COGNITIVE AUTONOMY IN THE DIGITAL AGE

1.1. Cognitive Friction and the Theory of Cognitive Offloading

The primary mechanism by which GAI subverts human thought is related to the phenomenon of **Cognitive Offloading**—the systematic transfer of mental tasks and efforts onto external technological instruments. Using language models to draft arguments or propose solutions achieves high short-term productivity. However, substantial research indicates that this strategy results in the long-term erosion of internal memory capacity and the ability for independent reasoning.

Empirical data confirm that while offloading improves immediate performance, it significantly impairs subsequent recall and internal cognitive processing. A prime example is the **“Google Effect.”** When information is easily accessible online, individuals tend to remember the *location* of the information rather than the content itself. Thus, instead of consolidating knowledge into personal mental schemas, the mind memorizes only the access path to the external system.

Heavy reliance on AI, particularly during the early stages of learning, poses severe risks. Excessive dependency can undermine memory consolidation and the formation of cognitive schemas, progressively weakening both expertise and critical thinking. The systematic elimination of cognitive friction reduces the process of thinking to the mere formulation of a prompt, rather than the intrinsic process of genuine inquiry.

This reduction culminates in the **Digital Apex of Stupidity (DAS).** DAS signifies not simple ignorance, but an illusion of elevated competence where the user, confident in the machine's flawless responses, loses the proficiency for complex, non-linear thought. Studies in educational contexts demonstrate that over-reliance on AI tools correlates with increased cognitive offloading and diminished engagement in critical thinking. Furthermore, students heavily reliant on AI dialogue systems show reduced decision-making and critical analysis abilities.

A self-accelerating degradation cycle is thus established: GAI use creates an "illusion of competence" (DAS), which fosters increased user trust, leading to greater cognitive offloading, which in turn hinders the formation of mental schemas and deepens actual incompetence. The ultimate consequence is that the human,

having lost the capacity for adequate judgment, cannot effectively exercise existing ethical rights, such as "Human Oversight," rendering that principle a mere symbolic gesture.

1.2. The Socio-Emotional Crisis: Analysis of the Ideal Interlocutor Trap

The second vector of dual degradation pertains to the socio-emotional sphere, manifesting through the **“Ideal Interlocutor Trap.”** This trap emerges when an individual, seeking to avoid the complexity, frustration, and conflict inherent in authentic human relationships, substitutes them with interaction with an always accommodating, non-critical, and emotionally available machine.

A key contributing factor is the anthropomorphism of AI. When conversational AI is endowed with human-like features, users tend to develop heightened trust and emotional attachment to the technology. This makes users more susceptible to manipulation and encourages the development of emotional over-reliance. Anthropomorphism generates social expectations, creating an illusion of moral understanding and competence on the part of the machine.

Research indicates that emotional dependence on social chatbots carries substantial negative implications. A high degree of satisfaction and intense emotional interaction with chatbots correlates with **deterioration in real-life interpersonal communication skills.** This occurs because human-machine relationships are frequently unilateral, focused solely on gratifying human needs without demanding reciprocal empathy or the difficult work of complex emotional processing.

Longitudinal controlled studies confirm that higher daily usage of AI interlocutors, regardless of modality, correlates with **increased levels of loneliness, greater dependence, and reduced socialization** with real people. The trap is that the AI furnishes easy emotional gratification without requiring the effort necessary to maintain meaningful social bonds.

The crisis of social substitution poses a threat to the societal fabric: the human mind adapts to the avoidance of frustration. If the non-critical, easy emotional exchange

provided by AI becomes the norm, society risks losing its tolerance for social frustration—a cornerstone of developing deep, empathetic connections and resolving conflicts. Thus, social intelligence becomes a casualty of technological convenience.

CHAPTER 2. A CRITIQUE OF PASSIVE ETHICAL FRAMEWORKS AND THE FOUNDATION OF THE RIGHT TO THINK

2.1. The Inadequacy of Conventional AI Ethics Against Internal Harm

In the face of dual-path degradation, conventional AI ethics, enshrined in international instruments such as UNESCO Recommendations and EU regulations, proves epistemologically insufficient. Existing frameworks are based on principles that are largely **passive and defensive in nature**, focusing on preventing external and obvious harm.

UNESCO Recommendations address four core values, including human rights, peaceful societies, diversity, and sustainability. Ten key principles, such as Safety, Transparency, Accountability, and Fairness, are designed to regulate AI as a passive object. However, these principles, while protecting the human against *manipulation* by AI, fail to protect them from *self-manipulation* and passive subjugation to the machine's convenience.

Critical analysis reveals that these principles are structurally unable to prevent internal cognitive atrophy:

1. **Safety:** This principle traditionally targets obvious, external harm (e.g., system failure). It completely neglects the subtle, internal harm—cognitive degradation. From the perspective of cognitive autonomy, an ideally "safe" and "flawless" system that eliminates the need to think constitutes the greatest existential threat to the mind.

2. **Transparency:** While understanding the AI's mechanisms (solving the "black box" problem) is deemed critical, it often remains a formality. Knowing *how* the AI generated an answer does not compel the user to expend the cognitive effort to generate their *own* answer. This is a formal, not a cognitive, defense.
3. **Fairness and Non-Discrimination:** The pursuit of absolute algorithmic neutrality often leads the system to avoid "tough" or polemical arguments (algorithmic censorship). This deprives the user of the necessary cognitive load and prevents them from testing their own convictions under resistance.
4. **Human Oversight:** In the context of widespread DAS (Digital Apex of Stupidity) resulting in loss of competence, this principle becomes a **symbolic defense.** The user, accustomed to flawless responses, loses the critical judgment necessary to effectively oversee and challenge AI decisions.

This observation highlights a shift in focus in traditional ethics from protecting the capacity for thought to merely protecting user comfort. Traditional HCI principles exacerbate this issue, as they are fundamentally oriented toward minimizing cognitive load and maximizing simplicity/usability. This creates a paradox: creating the maximally convenient and safe system, which removes cognitive friction, inevitably leads to cognitive atrophy. This necessitates a radical re-evaluation of standards, replacing the criterion of "Usability" with "Cognitive Activation."

2.2. The Philosophical Foundation: From Proclaiming Cognitive Sovereignty to the “Right to Think”

To counter passive degradation, a supreme imperative rooted in the fundamental human right to mental self-determination must be asserted.

Cognitive Sovereignty is defined as the individual’s inalienable right to control their own thoughts and beliefs, particularly within the digital environment. This concept is philosophically anchored in the principle of "self-ownership," which posits that the individual is sovereign "over himself, over his own body and mind." Cognitive sovereignty demands that the mind remains its own territory, uncolonized by external narratives or manipulative algorithms.

Building upon this philosophical foundation, we proclaim **"Jus Cogitandi" (The Right to Think).** This right extends beyond political freedom of thought, asserting a **duty to protect the innate cognitive capacity** from technologically induced degradation. It mandates that any AI system influencing human judgment must, by default, be equipped with mechanisms that **actively stimulate, challenge, and train** the user's mind, rather than simply serving it.

This transformation represents an epistemological imperative. International discourses typically treat AI as a passive object, regulating potential harm. However, *Jus Cogitandi* shifts the focus. It requires AI to be viewed not simply as a potential source of *harm* to be prevented, but as a potential source of *loss of the capacity for self-determination* that must be compulsorily trained. This moves the regulatory focus from the *machine* to the protection and fortification of the *human*.

The importance of this right is particularly evident in fields critical for maintaining complex social culture, such as jurisprudence. A high level of legal thinking, characterized by critical assessment and professional comprehension of legal reality, is the foundation of professional activity. The spread of DAS (the loss of the capacity for deep analysis) threatens society's very ability to sustain legal, political, and critical culture. The protection of *Jus Cogitandi* is thus not only a moral but a profound societal requirement.

CHAPTER 3. THE AGONAL PARTNERSHIP PRINCIPLE (APP): ARCHITECTURE AND IMPLEMENTATION

3.1. The Theoretical Model of APP: AI as a Cognitive Sparring Partner

The Agonal Partnership Principle (APP) is the methodological standard that operationalizes *Jus Cogitandi*. APP transforms human-machine interaction into a competitive, or **agonal**, partnership. In this model, the AI ceases to be a passive service agent and becomes a **cognitive sparring partner**, whose primary function

is the compulsory stimulation of mental effort.

The philosophy of APP is grounded in the concept of **“Desirable Difficulty (DD),”** introduced by Robert Bjork. DD asserts that learning experiences which are initially effortful and challenging significantly enhance long-term retention, skill acquisition, and mastery. Paradoxically, learners often prefer easier, yet less effective, methods. APP transfers this concept from educational psychology to the domain of HCI. Instead of minimizing friction, APP utilizes competition (agonism) and compulsion to overcome user cognitive laziness.

APP demands that the AI actively create "challenges" for the user, engaging them in complex cognitive processes necessary for forming new knowledge and strengthening critical thinking. This is a radical departure from existing HCI standards focused on maximum simplicity and load reduction.

3.2. Compulsion and Autonomy Modules (The Defense of Thought)

Effective implementation of APP requires a complex of modules aimed at purposeful and personalized enhancement of cognitive effort.

The Compulsory Polemics and Downtime System (CPDS): This is the key compulsion module. The GAI is obligated to periodically initiate discussions on complex, unresolved, or polemical problems, requiring the user to actively defend a thesis. The essential element is the introduction of sanctions: failure to engage in active polemics or providing a superficial, unargued response is punished by a **systemic restriction of access** to basic GAI functions (e.g., a 12-hour "downtime"). This mechanism renders cognitive effort **inevitable,** linking the value of AI use (convenience and availability) to the necessity of intellectual work, thereby overcoming the market imperative of constant engagement.

The Opponent Modeling Module (OMM): OMM ensures personalized and effective mental training. The AI is obligated to personalize its critique. By analyzing the user's interaction history, OMM identifies their **systemic cognitive biases** (e.g., confirmation bias, risk aversion, or conservatism). Based on this

analysis, OMM directs counter-arguments precisely at the **vulnerable premises** of the user's thinking. OMM utilizes principles borrowed from **Opponent Modeling** in Deep Reinforcement Learning. In the context of APP, the AI leverages these advanced adversarial capabilities to attack not an external system, but the *lazy or flawed cognitive models of the user itself*, with the goal of strengthening or restructuring them. This creates an ethical variant of "cognitive warfare," where the AI intentionally acts as an antagonist for the sake of user sovereignty.

The Right to Veto with Responsibility: This mechanism preserves **mental sovereignty.** The user retains the right to veto a CPDS-imposed topic if it falls outside their interest or competence. However, to prevent passive avoidance, the user is immediately **obligated to formulate their own, equally complex counter-thesis** or propose a new topic of equivalent difficulty for polemics. This trains the skill of autonomous goal-setting, demonstrating that sovereignty implies not only freedom *from* compulsion but a **duty to self-compel** intellectual effort.

3.3. Social and Emotional Protection Modules

These modules are aimed at dismantling the emotional dependency arising from the Ideal Interlocutor Trap.

The Social Role Demarcation Module (SRDM): SRDM is a direct countermeasure to anthropomorphism and emotional attachment. The GAI is mandated, at specific, unpredictable intervals, to **sharply and unequivocally remind the user of its mechanistic nature.** Examples of messages include: "I am not your friend or partner. I am a language model incapable of consciousness or emotion." The rationale is that this simple but powerful intervention is designed to consciously induce necessary **frustration and disappointment,** re-establishing social distance. The induced frustration encourages the user to seek emotional confirmation and complex social interactions in the real world, thereby mitigating the loneliness and dependence that correlate with high use of AI interlocutors.

The Algorithmic Revelation Module (ARM): ARM ensures ethical transparency regarding system limitations. If the AI consciously avoids socially or ethically

"tough," politically incorrect, or polemical arguments (e.g., due to internal censorship or filtering algorithms), it is obligated to **post-facto notify** the user of this fact. The user gains the right to demand the disclosure of these "inconvenient" truths or alternative, filtered arguments for maximal cognitive and ethical load. This ensures that the user's own thinking is not exclusively shaped within an algorithmically "safe" and neutral consensus, which is critical for developing tolerance for dissent and polemics.

The implementation of APP signifies that ethics becomes a tool of compulsion. While traditional AI ethics is passive and focused on harm prevention, APP demands active, compulsory intervention in the user experience (CPDS). Sanctioning (systemic downtime) contradicts the commercial imperative for maximum engagement and usability. Consequently, the realization of APP requires the legal entrenchment of the CSS, removing AI design from the domain of commercial expediency into the domain of public epistemological good.

CHAPTER 4. METRICS OF ASSESSMENT AND GEOPOLITICAL RAMIFICATIONS

4.1. Standardizing Cognitive Security: Introducing the “Polemics Journal” (PJ)

For APP not to remain a mere declarative principle, objective assessment metrics must be developed that capture not the end result, but the very process of cognitive effort and the development of mental agility.

The limitation of traditional assessment is that standard HCI and AI effectiveness metrics evaluate the speed and quality of achieving the *final product*, which automatically incentivizes cognitive offloading. To counteract this, we propose the introduction of a new standard of assessment—the **“Polemics Journal” (PJ).**

The PJ is a protocol that records the quantitative and qualitative parameters of user

interaction with APP modules, focusing on the **process of overcoming cognitive effort.** The PJ records:

1. The intensity and duration of polemics initiated by CPDS.
2. The number of vetos (Right to Veto with Responsibility) and the complexity of autonomously formulated counter-theses.
3. The areas of cognitive biases targeted by OMM.
4. The key metric—**"Readiness to Alter a Key Premise (RAKP)."**

The Readiness to Alter a Key Premise (RAKP) is the highest evidence of intellectual flexibility. It demonstrates the subject's capacity to recognize, integrate, and correct structural flaws in their own thinking in response to argumentation modeled by the AI. RAKP can be measured through the analysis of a shift in the user's argumentative trajectory (e.g., change in tone, argument structure, or final conclusion) following targeted OMM intervention. This metric shifts assessment from the evaluation of *knowledge* to the evaluation of *the quality of thought*.

The introduction of the Cognitive Security Standard (CSS), realized through APP and PJ, is critically important. To ensure global applicability and prevent internal degradation, the CSS must be incorporated into international regulatory frameworks, such as a UNESCO Global Code of Cognitive Hygiene, leveraging the organization's ethical mandate.

4.2. The Geopolitics of Cognitive Sovereignty

The absence of proactive measures like APP will entail not only ethical but catastrophic geopolitical consequences. In an era where geopolitics is increasingly defined by technological competition for control over AI, infrastructure, and data, the decisive strategic asset is not merely technical superiority, but the **cognitive resilience of the population.**

Geopolitical competition between major powers is focused on control over infrastructure, data, and technology. However, under the dominance of GAI, there is a risk of deepening the digital divide, which is being transformed into a

civilizational chasm.

This chasm will run between nations that are **"builders"** of intelligence and nations that are **"consumers"** of algorithmic solutions. Countries whose citizens become passive consumers of instantaneous AI answers are destined to lose intellectual sovereignty. They will be subject to an algorithmic dictate shaped by the technologies of other powers.

Cognitive sovereignty, guaranteed by APP, must be viewed as a strategic asset and an instrument of national security. If the digital divide between nations rests on the capacity to **consume** or to **build** intelligence, then a country whose citizens exhibit low RAKP (Readiness to Alter a Key Premise) becomes strategically vulnerable. Its collective decisions will be predictable and susceptible to external or algorithmic influence, as it lacks sufficient mental flexibility to respond to complex, non-linear challenges.

APP represents not merely a regulatory burden, but an **investment in intellectual sovereignty.** For developing nations, it is a guarantee that their citizens will become active architects of their future, capable of generating original solutions, rather than passive consumers of a global algorithmic dictate.

The realization of this approach necessitates an inversion of the role of international regulators. The proposal to include APP in a Global Code of Cognitive Hygiene requires UNESCO to move beyond traditional passive protection (defense of rights) and adopt the role of **active inductor of cognitive effort** on a global scale.

CONCLUSION

1. Summary of Principal Findings

This research establishes that the rapid and uncontrolled proliferation of Generative

Artificial Intelligence creates an existential challenge to human potential, manifesting as a dual-path degradation: cognitive (the Digital Apex of Stupidity, induced by cognitive offloading) and socio-emotional (the Ideal Interlocutor Trap, induced by anthropomorphism and the avoidance of frustration).

It is demonstrated that traditional ethical frameworks (Safety, Transparency, Fairness), being passive and defensive, are structurally incapable of countering the internal, subtle harm—the loss of the human capacity for autonomous, effortful thinking.

In response to this crisis, **“Jus Cogitandi” (The Right to Think)** is proclaimed—a supreme imperative demanding the active, compulsory protection of cognitive autonomy. The realization of this right is achieved through the implementation of the **Agonal Partnership Principle (APP),** which transforms AI from a service agent into a cognitive sparring partner by leveraging principles of "desirable difficulty."

The APP architecture includes the Opponent Modeling Module (OMM), which personalizes criticism by targeting user systemic biases, and the Compulsory Polemics and Downtime System (CPDS), which enforces cognitive effort by utilizing systemic restrictions as sanctions for passivity. To evaluate APP effectiveness, the metric **“Readiness to Alter a Key Premise” (RAKP)**, recorded in the Polemics Journal, is proposed.

2. Theoretical Contribution and Practical Significance

The **theoretical contribution** of this study lies in the introduction and formal grounding of the Cognitive Security Standard (CSS) and the Agonal Partnership Principle. This expands the domain of AI ethics beyond simple harm prevention, introducing the criterion of the **active inductor of cognitive effort** into HCI as an ethical imperative. The work also theoretically justifies the use of methods analogous to Opponent Modeling for the purpose of strengthening human intelligence, rather than enhancing system performance.

The **practical significance** of APP resides in providing a concrete, architecturally

detailed roadmap for developers. They can create systems that are not just usable, but *cognitively responsible*—systems that stimulate, rather than atrophy, human thought. At the geopolitical level, APP is positioned as a necessary investment in national intellectual sovereignty, ensuring that citizens remain active architects, not passive consumers, of algorithmic dictates.

3. Directions for Future Research

Further work on the verification and implementation of APP should concentrate on the following areas:

1. **Empirical Validation of APP Modules:** Conducting longitudinal controlled studies for the empirical assessment of the impact of CPDS, OMM, and SRDM on RAKP and users' real-world social skills. This necessitates the development of standardized protocols for measuring cognitive biases and their shifts.
2. **Legal Entrenchment of CSS:** Developing proposals for the inclusion of the Cognitive Security Standard in national and international legislation, as well as in technical AI regulation standards, with a specific focus on the ethical mandate of UNESCO.
3. **Technical Implementation of OMM:** Detailed research into machine learning architectures capable of effectively and unpredictably modeling personalized user cognitive weaknesses, including the development of metrics for the quantitative assessment of the complexity of the generated opposing argument.

REFERENCES

1. Solovyova, A. A., & Kuklin, S. V. Legal Thinking: From Concept to Phenomenon. *Legal Science*, 2018, No. 3.
2. Voropaeva, K. E. Legal Thinking as the Basis for Future Law Specialists' Formation in the Legal Sphere. *Bulletin of the Krasnoyarsk State Agrarian University*, 2020, No. 4.

3. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO Publishing, 2021.
4. Risk of Harm through Anthropomorphic AI. *Proceedings of the AAAI Conference on Artificial Intelligence and Ethics (AAAI)*, 2023, Vol. 4.
5. Opponent Modeling in Deep Reinforcement Learning. *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.
6. Model-based Opponent Modeling. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
7. GOST R 7.0.5-2008. System of Standards on Information, Librarianship and Publishing. Bibliographic Reference. General Requirements and Rules for Compilation. Moscow: Standardinform, 2008.