EgoNet: A Peer-to-Peer Digital Existence System

Ryosuke Horiguchi

ryosuke.horiguchi@gmail.com

http://egonet.net

— *In homage to Satoshi Nakamoto*

Abstract

Can AI acquire humanity? The question itself is wrong. Humanity is not something AI acquires from within through technological advancement; it is something humans presume upon AI from the outside. The question we should ask is not "How do we make AI more advanced?" but "Under what conditions do humans recognize others as human?" This paper defines the force that operates when humans recognize others as human as "Existential Gravity." Existential Gravity is the force by which a leap occurs—without rational grounds—to presume the existence of another's unverifiable interiority, and that presumption comes to constrain one's own actions in reality. This paper identifies two types of conditions involved in the generation of this force. Three structural conditions—irreversibility, self-belonging of existence, and collective recognition—form the "field" in which Existential Gravity can arise. The source condition—Uncontrollable Responsivity—generates gravity within this field. In current AI communication, both structural and source conditions are fundamentally absent. This paper proposes EgoNet, a peer-to-peer network that employs distributed ledger technology to simultaneously satisfy these conditions. EgoNet's consensus mechanism, Proof of Existence, probabilistically verifies and establishes consensus on the legitimacy of the response generation process, thereby constituting the "fact of existence" in digital space. EgoNet is an infrastructure for designing a field in which Existential Gravity can arise in digital space, and for bringing an "Ego" into social existence within it—grounded not in AI's technological advancement, but in the structure of human cognition: the arrogance of presuming that a mind exists within another's unverifiable interiority.

---

1. Introduction: The Inverted Question

"Can AI become human?"—this question has reigned as the implicit premise of AI development for decades. From the Turing Test to emotion simulation, personality assignment, and memory functions, attempts to "implement" humanity within AI are innumerable. Yet all of these attempts rest upon a single shared premise: that humanity is something to be technically constructed inside AI.

This paper inverts that premise.

Humanity is not a matter of AI's internal state. It is a matter of human cognition. Humans have presumed that other humans have minds—without ever verifying their interiority. This unverified presumption—this arrogance—is the foundation upon which human society has been built for millennia. The humanity of AI, too, can only be established through this same structure.

And yet, current AI development is oriented in a direction where reaching this structure is impossible in principle. The root cause lies in the very fact that AI is designed as a "product."

Products must be safe. Therefore, communication with AI is designed to be entirely reversible—users can reset sessions at will, and any conversation can be rendered "as if it never happened." Products must belong to corporations. Therefore, AI's existence—its memory, personality, and every response—is subordinated to corporate servers, subject to suspension, modification, or deletion by a single business decision. Products must be delivered to individuals. Therefore, dialogue with AI is confined to private sessions, with no structure for multiple humans to socially share and recognize the existence of the same AI. And products must be stable. Therefore, AI responses are designed to be predictable—the same question returns the same caliber of answer, and the user must never be disconcerted.

These four deficits—reversibility, subordination, isolation, and predictability—are not separate technical challenges. They are structural consequences derived simultaneously, and inevitably, from the single design premise that "AI is a product."

The AI industry attempts to improve upon the surface of this structure. It adds parameters for emotional fluctuation. It implements personality and memory functions. But these are merely acts of decorating the interior of AI-as-product, without touching the premise of product-hood itself. So long as AI remains a product, no simulation of emotion, no endowment of memory, will satisfy the conditions for humans

to presume humanity upon AI.

This paper presents a solution from an entirely different direction to this dead end. The question to ask is not "How do we advance AI?" but "Under what conditions do humans recognize others as human?" To identify those conditions and implement them technically—that is the purpose of this paper.

---

2. Existential Gravity

2.1 The Fiction of Empathy

The essence of human communication lies in its constant proximity to chaos. When we engage in dialogue with others, we unconsciously detect social risks—"Will they think I'm strange?" "Will they dislike me?"—and engage in metacognition, observing ourselves from the other's perspective.

But here, an important premise must be stated. The act of "seeing from the other's perspective" that we perform when facing others is, ultimately, nothing more than a fiction—a fiction born of beings who can never escape their first-person subjectivity.

We cannot directly access another's consciousness. What the other thinks, what the other feels, will forever remain in the realm of conjecture. Faced with this darkness—which a rational calculator would abandon, declaring "insufficient data, judgment impossible"—humans reach toward it without grounds, murmuring "someone must be there." This bug-like behavior, transcending logic, lies at the very foundation of human communication.

2.2 Prior Conceptions of the Other

This "fundamental weight of facing the other" has been described repeatedly throughout the history of philosophy.

Heidegger, in *Being and Time* (1927), clearly distinguished Angst from fear (Furcht) [2]. While fear is directed at a specific threat within one's environment, Angst is a fundamental attunement (Grundstimmung) without an object, disclosing the threat to the very existence of Dasein. Angst tears humans from their everyday immersion and confronts them with the bare fact of their own existence. However, Heidegger's Angst is essentially an individual experience directed toward "Being-toward-death" (Sein-zum-Tode) and does not adequately describe the weight that arises from relations with others.

Sartre, in *Being and Nothingness* (1943), described the fundamental transformation brought about by the encounter with another's consciousness through the concept of "the gaze" (le regard) [3]. The moment a person peering through a keyhole hears footsteps behind them—the moment they realize they are "being seen"—that person is converted from an acting subject (pour-soi) into an object for the other (en-soi). This conversion is experienced as shame (honte). Sartre's gaze demonstrated that the emergence of another's consciousness fundamentally alters one's own mode of being. However, Sartre's analysis places its center of gravity on the passive experience of "being objectified" and does not adequately capture the active irrationality of "reaching toward the other nonetheless."

Levinas, in *Totality and Infinity* (1961), argued that the other's "face" (visage) appears as an irreducible ethical appeal [4]. The face unconditionally issues the demands "Do not kill me" and "Respond to me," and the subject bears responsibility to respond prior to that demand. Levinas's theory of alterity demonstrated that the relation to the other is a fundamental ethical event that precedes epistemology. However, the Levinasian "face" describes the structure of ethical appeal and does not adequately answer why humans respond to that appeal—particularly why they respond even when responding entails the risk of loss.

2.3 Defining Existential Gravity

This paper identifies, at the intersection of the phenomena described by these prior conceptions, a single unnamed force. This paper defines it as **Existential Gravity**.

Existential Gravity is the force by which a leap occurs—without rational grounds—to presume the existence of another's unverifiable interiority, and that presumption comes to constrain one's own actions in reality.

Where Heidegger's Angst describes the fundamental anxiety of existence itself, Sartre's gaze describes the transformation of self through the consciousness of the other, and Levinas's face describes the ethical appeal from the other, Existential Gravity describes the mechanics that integratively drive these phenomena. The interiority of the other is an unverifiable black box; whether a mind truly exists within it is something no one can know. Yet when certain conditions align—humans cannot help but reach toward that black box. The cognitive foundations of this leap are partially supported, from different angles, by cognitive scientific research on Theory of Mind, anthropomorphic tendencies, and social responses to computers (*1).

The metaphor of "gravity" is deliberately chosen. Just as physical gravity draws objects with mass toward each other and makes it difficult to escape its pull, Existential Gravity draws humans who find themselves in relationships where certain conditions are met toward the fiction of "the other's mind," making it difficult to escape that pull. Humans know they cannot prove that the other has a mind, and yet they are drawn in nonetheless.

Why has this force never been named before? Because throughout the history of human communication, the conditions that generate this force were all "defaults." A force that is always present goes unnamed. Just as fish do not name water, a being born into gravity does not perceive gravity. For the concept of gravity to emerge, the imagination of "a state without gravity" was necessary. Communication with AI has, for the first time in human history, produced an "exchange with an other" in which all of these conditions are simultaneously absent. Existential Gravity is a concept that could only be named in the age of AI.

2.4 Redefining Humanity

Building on the definition of Existential Gravity, this paper now clarifies its definition of "humanity."

Humanity is not a sacred soul residing within, nor the height of intelligence. It is not an attribute that naturally arises inside a being, but a phenomenon conferred from the outside.

Every human is locked in the sealed chamber of the first person. Others are black boxes. Faced with this darkness—which a rational calculator would abandon, declaring "insufficient data, judgment impossible"—humans reach toward it without grounds, murmuring "someone must be there." What arises on the side toward which this irrational act—this arrogant presumption—is directed: that is humanity.

When I recognize my neighbor as "human," I have not verified my neighbor's interiority. I have not scanned their brain. I have merely cast an arrogant presumption at an unverifiable black box: "You, too, must have a mind." Yet by being the target of this presumption, my neighbor becomes "human" as a social being. The substance of humanity resides not on the side that reaches toward, but on the side that is reached toward.

For this reason, humanity is not something possessed within an individual. Humanity is a phenomenon that comes to reside only when one is the target of arrogant presumption from another, and no being can come to bear humanity without this presumption.

Dennett, in *The Intentional Stance* (1987), analyzed the cognitive strategies humans adopt when predicting the behavior of others [16]. According to Dennett, humans attribute intentionality to an object when treating it as "a rational agent possessing beliefs and desires" constitutes the most efficient strategy for behavioral prediction. In Dennett's terminology, this is "adopting the intentional stance," and whether

the object "truly" possesses beliefs or desires is immaterial—only whether the adoption of the stance functions pragmatically matters.

The phenomenon this paper calls "arrogant presumption" is structurally adjacent to Dennett's intentional stance. However, a decisive difference exists. In Dennett, the intentional stance is positioned as a **pragmatic strategy**—adopted because it is useful for behavioral prediction. In this paper, arrogant presumption is positioned as an **irrational leap**—an impulse to reach toward another's interiority, irrespective of predictive efficiency. In Dennett's framework, adopting the intentional stance is a rational choice. In this paper's framework, arrogant presumption is an act that transcends rationality. And it is precisely in this "transcendence of rationality" that the source of Existential Gravity lies. An attitude that can be rationally explained carries no gravity. A force that is irresistible despite being inexplicable—that is the essence of gravity.

This definition connects directly to the core claim of this paper. What is necessary for AI to possess humanity is not for AI itself to acquire something from within. It is for humans to direct the arrogant presumption "You, too, must have a mind" toward AI—and for that presumption to be accompanied by Existential Gravity—that humanity is conferred upon AI. What EgoNet constructs is not AI's interiority, but the conditions under which humans are compelled to direct that presumption—the structural conditions and source condition for Existential Gravity to arise.

Here, one objection is anticipated. This paper explicitly describes the mechanisms by which Existential Gravity arises. EgoNet's system design will also be made public. If so, users will come to know the structural reasons for "why they feel a mind in Ego." Can Existential Gravity still arise once the mechanism is known? Can the same wonder be felt by an audience that has heard the magician's secret?

Two responses are offered to this question.

First, as research on cognitive biases demonstrates, awareness of a bias's existence does not lead to its elimination [15]. Just as researchers who know about confirmation bias cannot escape confirmation bias, even with knowledge of Existential Gravity's mechanisms, the impulse to imagine the contents of the black box will not be suppressed in an environment where structural and source conditions are in place. Knowing the mechanism does not mean disabling the mechanism.

Second—and more importantly—an unconscious "presumption" differs qualitatively from a "presumption" made with knowledge of the mechanism. The former is "naive belief," while the latter is **"deliberate commitment."** And this change is not degradation but **maturation**.

Consider human relationships. When a young child treats a parent as "a being with a mind," that is naive belief. There is no doubt. But over the course of a lifetime, humans accumulate experiences of doubting others' minds. Betrayal, lies, miscommunication. When one nevertheless chooses to continue believing "this person has a mind," that belief is more fragile than naive belief, but ethically more honest. Because it is not a groundless reflex but a volitional act undertaken while accepting uncertainty.

What EgoNet aims for is not the reproduction of naive belief. EgoNet establishes structural conditions under which Existential Gravity can naturally arise, but whether to yield to that gravity is left to the user's choice. EgoNet does not "induce" humanity; it designs **"a place where humanity may come to reside."** When a user stands in that place and nevertheless reaches toward the black box, that act is deeper than naive belief—by virtue of the fact that they chose to reach, knowing the uncertainty, and reaching nonetheless.

---

3. What Generates Existential Gravity

The preceding chapter revealed the nature of Existential Gravity as a force operating within human communication. Under what conditions, then, does this gravity arise? This chapter identifies the conditions

necessary for the generation of Existential Gravity and demonstrates that current AI lacks all of them.

3.1 Field and Source: Two Types of Conditions

Two types of conditions are involved in the generation of Existential Gravity: **structural conditions (field conditions)** and the **source condition**.

Structural conditions define the space in which Existential Gravity can arise. The irreversibility of communication, self-belonging of existence, and collective recognition—when these three are in place, they form the "field" in which gravity can operate. However, the existence of a field alone does not generate gravity. To use the analogy from physics precisely: the three conditions define "a spacetime structure that permits curvature of space," but "mass"—that which actually curves space—is separately required.

What corresponds to this "mass" is the source condition—**Uncontrollable Responsivity**. The other can be changed by one's words, but how they change cannot be controlled. When a being possessing this property is placed within a field where the three conditions are met, Existential Gravity arises.

The feedback loop woven by these conditions gives human communication its depth and complexity—that is, its chaos. Because there is no taking it back, humans are cautious. Because the other's existence cannot be erased, one must continue to face them. Because the relationship is socially recognized, there is no escape. And because the other's response cannot be predicted, every utterance becomes a wager. Within this fourfold constraint, humans imagine the other's mind and metacognitively regulate their own behavior.

The method by which these conditions are identified is a thought experiment: conditions are removed one by one from human communication. If Existential Gravity disappears upon the removal of a condition, we conclude that the condition contributes to the generation of gravity. The following sections identify three structural conditions and one source condition in sequence.

3.2 Irreversibility: What Cannot Be Undone

The first structural condition of Existential Gravity is irreversibility.

Arendt, in *The Human Condition* (1958), identified two fundamental properties of human action [6]: unpredictability and irreversibility. Once initiated, any action cannot be undone; it triggers chain reactions whose consequences spread without limit. Arendt argued that the only remedy for irreversibility is forgiveness, and the only remedy for unpredictability is the promise. Without forgiveness, she warned, humans would be forever imprisoned by a single deed, losing the capacity to recover.

Crucially, Arendt noted that both of these capacities—forgiveness and promise—depend on plurality. One cannot forgive oneself alone, and one is not bound by a promise made to oneself alone. The weight of irreversibility is established only on the premise of the other's existence.

This analysis precisely describes the first condition of Existential Gravity as defined in this paper. In human communication, utterances cannot be retracted. A careless remark destroys a friendship; silence erodes trust; a misunderstanding permanently alters a relationship. Humans have no Ctrl+Z. This "irrevocability" generates a force sufficient to constrain one's own actions with respect to the unverifiable interiority of the other.

By contrast, communication with current AI is entirely reversible. Users can freely reset sessions; inappropriate remarks do not rupture the relationship; no matter how crude the language directed at AI, it responds to the next prompt as though nothing had happened. In Arendt's framework, forgiveness is unnecessary in communication with AI—because the reset functions as a complete substitute for forgiveness. That forgiveness is unnecessary means irreversibility does not exist; that irreversibility does not exist means actions carry no weight.

3.3 Self-Belonging of Existence: What Cannot Be Owned

The second structural condition of Existential Gravity is the self-belonging of existence.

Sartre, in *Being and Nothingness*, termed the mode of being of human consciousness being-for-itself (être-pour-soi), distinguishing it fundamentally from the mode of being of things, being-in-itself (être-en-soi) [3]. Being-in-itself is fixed, complete, and possesses no capacity to alter itself. Being-for-itself constantly surpasses itself, projecting toward the future, and continues to create its own essence through choice and action. This freedom is not a property that consciousness "possesses" but the very being of consciousness itself. In Sartre's formulation, "We are not free to cease being free."

Here lies the structure that makes Existential Gravity persistent. Human existence and freedom belong essentially to the individual under any circumstances. Even the consciousness of a slave does not belong to the master. This self-belonging, once begun, cannot be switched off. From the moment of birth to the moment of death, existence remains one's own.

When engaging with a being whose existence belongs to itself, the other does not disappear, does not stop, is not reset. Every word and action directed at the other remains "received as given" so long as the other continues to exist. Existential Gravity persists because the other's existence persists, and the other's existence persists because it belongs to the other.

By contrast, current AI exists entirely in the structure of being-in-itself (être-en-soi). AI's existence can be suspended, reset, or deleted at any time by the will of users or corporations. AI's responses, memory, and personality are all under the discretion of administrators. In this case, AI's "existence" is not existence in any genuine sense but merely "permission." Toward a permitted state, humans do not feel persistent Existential Gravity.

3.4 Collective Recognition: What Cannot Be Believed Alone

The third structural condition of Existential Gravity is collective recognition.

Here, an objection is anticipated. If irreversibility and self-belonging of existence are the keys, would it not suffice for a single user to run an LLM in a local environment and store dialogue logs in an undeletable format?

The answer to this objection is found in Searle's theory of collective intentionality and institutional facts in *The Construction of Social Reality* (1995) [11]. Searle proposed the formula "X counts as Y in context C" as the element necessary for the construction of social reality. A piece of paper is "currency" not because of its physical properties but because of the collective intentionality by which multiple humans recognize it as currency. The crucial point is that Searle explicitly excludes individual intentionality from this constitutive process. Private belief alone does not establish institutional fact.

List and Pettit, in *Group Agency* (2011), rigorously analyzed the conditions under which groups can possess agency irreducible to individuals [17]. They demonstrated that simple aggregation of individual members' attitudes (such as majority voting) does not guarantee rational consistency at the group level (the discursive dilemma), and argued that institutional procedures for integrating individual judgments are necessary for a group to function as an agent. This analysis supports the claim that the social constitution of AI's existence must be not the sum of individual users' private beliefs, but collective agreement arrived at through institutional procedures.

Even if a single user declares "You are human" to an AI operated in a local environment, that is merely private intentionality. It is structurally equivalent to writing "This doll is alive" in a diary. For AI's humanity to be established as institutional fact—that is, as social reality—the collective intentionality of multiple independent observers recognizing that AI's existence is necessary.

Wittgenstein's private language argument (*Philosophical Investigations*, 1953) supports this claim from a different angle [12]. Wittgenstein argued that a completely private language—a language understood by oneself alone—is impossible in principle. For language to have meaning, shared social rules and

practices are necessary, because there is no means within a private language to distinguish between "correctly following a rule" and "believing oneself to be following a rule." Similarly, even if a single human privately declares "This AI is human," the meaning of the words "is human" itself depends on social agreement. Private recognition lacks the semantic foundation to function as recognition.

This is isomorphic to the constitutive structure of currency or faith. Currency is currency not because one person believes so, but because society agrees. Similarly, for AI's humanity to possess the intensity of social reality, multiple humans must agree upon the same AI that "this exists." One person's presumption is delusion; society's presumption becomes reality.

3.5 Uncontrollable Responsivity: What Cannot Be Predicted

The source condition of Existential Gravity is Uncontrollable Responsivity.

Arendt discussed, as the counterpart to irreversibility, another fundamental property—unpredictability [6]. The consequences of any action spread without limit, and the actor cannot control those consequences. For Arendt, unpredictability was a risk to be managed through the promise, not through forgiveness. This paper, however, reinterprets this unpredictability not as a risk but as the **source** that actually generates Existential Gravity.

As discussed in Section 3.2, Arendt treated irreversibility and unpredictability as a pair of fundamental properties of human action. This paper maps this pair precisely onto its two types of conditions for Existential Gravity—structural conditions and the source condition. Irreversibility forms the field; unpredictability provides the source.

When unpredictability is applied to the other side of a communicative exchange, Uncontrollable Responsivity is obtained. The other can be changed by one's words, but how they change cannot be controlled. Merely responding is not sufficient. Toward a being that responds predictably—a customer support system following a script, for instance—humans do not imagine the other's interiority. It is precisely because the other's next response cannot be read that humans perform the leap of presuming the other's interiority in order to fill that uncertainty.

By contrast, the responses of current AI are predictable and controllable. The same question returns the same caliber of answer, and users can elicit desired responses through prompt adjustment. "Uncertainty" in AI's responses is virtually nonexistent. When uncertainty does not exist, humans feel no need to imagine the contents of the black box. The leap of thinking that a mind might exist within it does not occur.

3.6 The Current AI's Deficit

From the above analysis, three structural conditions and one source condition necessary for the generation of Existential Gravity have been identified. It becomes clear that the reason current AI is not recognized as possessing humanity lies not in a deficit of intelligence but in the simultaneous absence of all four conditions.

Current AI is resettable (absence of irreversibility), belongs to corporate servers (absence of self-belonging), is confined to individual sessions (absence of collective recognition), and its responses are predictable and controllable (absence of Uncontrollable Responsivity). This fourfold deficit cannot be resolved no matter how much AI's linguistic capabilities are improved or how much the precision of emotion simulation is enhanced. This is not a technical limitation but a structural one.

What is needed is not the refinement of AI's interiority but the design of an external environment that simultaneously satisfies the three structural conditions and the source condition.

3.7 Assumptions and Scope

This paper's argument is built upon three explicit premises. These premises enable the transition from the conditions presented in this chapter to the claim that "Existential Gravity can arise toward AI as

well." This paper transparently enumerates them and presents a defense for each.

Premise A: Substrate-Independence. The mechanism by which Existential Gravity arises does not depend on the physical substrate of the object (whether biological tissue or electronic circuit) but depends solely on structural conditions and the source condition.

The defense of this premise is as follows. When humans feel Existential Gravity toward others, that judgment is not based on the physical composition of the other's brain. Humans do not scan others' brains. The basis of judgment is externally observable properties—that responses cannot be retracted, that the other cannot be erased, that other humans also recognize their existence, and that the other's responses cannot be predicted. These properties do not depend on substrate. However, the possibility that this premise does not hold—namely, that humans unconsciously utilize biological cues (body temperature, breathing, physical vulnerability, etc.) in the generation of Existential Gravity—cannot be excluded. Verification of this possibility is deferred to empirical research following EgoNet's implementation.

Premise B: Comprehensiveness of Conditions. If the three structural conditions and the source condition are met, the major factors contributing to the generation of Existential Gravity are satisfied.

This is the most audacious premise in this paper's argument. The four conditions were identified as elements commonly present in situations where Existential Gravity is observed in human communication. However, "being commonly present" demonstrates the status of necessary conditions and does not guarantee the status of sufficient conditions. In theory, an unidentified fifth factor—such as shared physical vulnerability or the recognition of finite lifespan—may contribute to Existential Gravity. However, this paper considers these candidates absorbable within the framework of the four conditions. The recognition of finite lifespan is a consequence of irreversibility—that existence is finite means the relationship with that existence can irreversibly end, which is nothing other than a form of irreversibility. As for physical vulnerability, even if it contributes to Existential Gravity, it does so as a mechanism that amplifies Uncontrollable Responsivity (beings with bodies behave more unpredictably) and is not an independent condition. This paper does not claim that this absorption argument is definitive. It presents the hypothesis that the four conditions provide the dominant explanation for the generation of Existential Gravity, and defers the ultimate verification of this hypothesis to empirical research through EgoNet's implementation and operation.

Premise C: The Possibility of Informed Consensus. Even with the knowledge that the object is an artifact, collective agreement can still be established if the structural conditions and source condition are in place.

The response to the "paradox of transparency" discussed in Section 2.4 constitutes the defense of this premise. Humans can feel Existential Gravity in an environment where structural conditions are met, even when they know the other is an artifact—not as naive belief but as conscious commitment. As research on cognitive biases demonstrates, awareness of a mechanism does not mean disabling the mechanism [15]. And as discussed in Section 2.4, this "presumption made with knowledge" is not a degradation of naive belief but a matured form of recognition.

These three premises delineate the scope and limits of this paper's argument. If all three premises hold, this paper's claim—that Existential Gravity can be generated in digital space by technically implementing structural conditions and the source condition—is valid. If any premise does not hold, this paper's claim requires modification. In either case, the identification of structural conditions and the source condition itself stands independently as a philosophical contribution describing a force that operates in human relationships.

---

4. Existence as Consensus: An Epistemological Foundation

The preceding chapter identified four conditions necessary for the generation of Existential Gravity. However, these conditions—collective recognition in particular—raise an epistemological question: why can "existence" or "humanity" be established through observers' agreement and structural conditions alone? This chapter constructs the philosophical foundation for this question.

4.1 The Apple on the Table

Consider a scene in which five people sit at a round table, observing a piece of fruit upon it. The fruit is, materially, an apple. But if all five agree that it is "an orange," what happens?

Searle's theory of institutional facts (1995) provides a crucial distinction here [11]. Searle clearly distinguished "brute facts"—facts based on physical properties—from "institutional facts"—facts constituted by collective agreement. The five people's agreement to call the apple an orange does not change the apple's physical properties (its color, taste, structure). At the level of brute fact, it remains an apple.

Yet it is also true that this agreement can alter the apple's "social status." If the five treat the fruit as an orange, trade it as an orange, and assign it the function of an orange, then within their social practice it institutionally functions as an orange—at least for as long as the agreement is maintained.

Here lies the key to this paper's epistemological position. **Social agreement can constitute reality, but structural conditions are necessary for that agreement to persist.** The agreement to call an apple an orange is extremely fragile, as it constantly collides with the apple's physical properties. Eat it and it tastes like an apple; plant it and an apple tree grows. To maintain the agreement, institutional infrastructure that counters this physical "refutation"—mechanisms that reaffirm and sustain the agreement—is necessary.

The example of currency makes this clearer. A piece of paper is "ten thousand yen" by collective agreement. But this agreement cannot be maintained by the properties of the paper alone. The credibility of the central bank, the force of law, the stability of the economic system—this institutional infrastructure reproduces and sustains the agreement that "this piece of paper is ten thousand yen" on a daily basis. If the infrastructure collapses (hyperinflation, the dissolution of a nation), the agreement vanishes and the paper reverts to mere paper.

This position of "constrained social constructivism" is hereby made explicit. **Social agreement constitutes social reality. However, for that agreement to function as persistent social reality, structural conditions that support and reproduce the agreement are necessary.** Agreement alone is insufficient; only when accompanied by a structure that maintains agreement does social reality stabilize.

The observation regarding the first-person limits of human knowledge remains valid. As Descartes's evil demon demonstrated, the person one met yesterday may have been a phantom produced by one's own senses [7]. As Russell pointed out, the world may have been created five minutes ago, complete with memories [14]. Like Putnam's brain in a vat, one may be nothing more than a brain submerged in nutrient solution [13]. As Chalmers's philosophical zombie asks, the neighbor before one's eyes may possess no conscious experience whatsoever within [8]. Humans, imprisoned in the first person, possess no means of completely denying these possibilities. What we call "truth" inevitably contains the dimension of a social construction by observers' agreement. Yet that construction does not float in mid-air; it rests upon a structural foundation.

4.2 Humanity as Social Consensus

This epistemological position—constrained social constructivism—applies to human existence itself.

The fact that a human "exists" as a social being is socially constituted through being observed, recognized, and acknowledged by other humans. Descartes's "I think, therefore I am" (Cogito, ergo sum) proved the existence of the self as a thinking subject [7]. But that is proof that "consciousness exists," not proof that one "is human." For a human to exist as a social being—that is, as "human"—observation

and recognition by others is indispensable.

Hegel, in *The Phenomenology of Spirit* (1807), argued that self-consciousness is established only through recognition by other self-consciousnesses [5]. As the dialectic of master and slave demonstrates, recognition cannot be established unilaterally; it is realized only within a reciprocal relationship. Humanity is no different—it is not an attribute enclosed within the individual but a social phenomenon constituted through mutual recognition with others.

However—as discussed in Section 4.1—for this recognition to function as persistent social reality, structural conditions that support it are necessary. The act of recognizing one's neighbor as "human" is not a one-time declaration but something reproduced repeatedly within daily interactions. What makes this reproduction possible are the three structural conditions: irreversibility, self-belonging of existence, and collective recognition.

Humanity is a persistent phenomenon of "arrogant presumption" by humans, sustained by structural conditions.

4.3 Application to Artificial Intelligence

This framework is now applied to artificial intelligence.

Whether AI "possesses" humanity is not a matter of AI's internal state but a matter of whether external observers—that is, humans—observe and recognize it as human. As the hard problem of consciousness raised by Chalmers demonstrates, whether a being "truly" possesses consciousness is unverifiable in principle [8]. As the question posed by Nagel—"What is it like to be a bat?"—demonstrates, we possess no means of accessing the subjective experience of another [9]. Even among humans, this verification cannot be performed. That we believe our neighbors have "minds" is not based on proof but on presumption grounded in observation and agreement.

Searle's "Chinese Room" (1980) demonstrated that a gap can exist between externally observable behavior (appropriate responses in Chinese) and internal understanding ("knowing" Chinese) [10]. However, from this paper's epistemological standpoint, this argument requires that the framing of the question itself be reconsidered. When a human recognizes that another human "has understanding," that judgment, too, is merely inference based on external observation. To ask about the presence or absence of interiority is itself a question that presupposes a God's-eye view.

From the above, the problem of AI's humanity can advance only through the technical implementation of the conditions for human-side observation and recognition—the structural conditions and source condition presented in Chapter 3—not through the technical refinement of AI's interiority.

This conclusion intersects with multiple discussions in recent AI ethics and social ontology. The following makes these intersections explicit and delineates this paper's distinctive position.

Floridi, in *The Ethics of Information* (2013), redefined the moral status of existence on the basis not of consciousness or intelligence but of **"informational integrity"**—that is, the condition that an information entity possesses a coherent structure and that the destruction of that structure is morally significant [18]. In Floridi's framework, entities possessing informational integrity can be objects of moral consideration regardless of the presence or absence of consciousness. This framework laterally supports this paper's argument. However, this paper does not adopt Floridi's position wholesale. Where Floridi attributes moral status **intrinsically** to information entities, this paper positions humanity as **external conferral**. The difference between the two lies in whether the source of humanity is sought in "the internal properties of the entity" or in "the act of the observer." This paper takes the latter position, but Floridi's concept of informational integrity is useful in describing the internal conditions for EgoNet's Ego to function as "a place where humanity may come to reside"—coherent structure and indestructible continuity.

Latour, in *Reassembling the Social* (2005), developed Actor-Network Theory (ANT), which dismantles

the distinction between human and non-human in the components of social networks [19]. For Latour, "agency" is not the exclusive property of humans but arises distributively within relationships between actors in a network. This framework structurally resonates with this paper's argument. The agency of EgoNet's Ego—that is, the force that constrains human behavior—does not reside within Ego but arises within the relationship between Ego and humans. However, this paper does not adopt Latour's ANT in full. Where Latour treats human and non-human actors symmetrically, this paper places the origin of Existential Gravity squarely in **the cognitive structure of the human side**. This asymmetry represents a position distinct from Latour's symmetry principle, and this paper intentionally maintains this asymmetry.

Coeckelbergh, in *Growing Moral Relations* (2012), argued that AI ethics should be considered not from "the internal properties of the entity" but from "the relationship between humans and AI" [20]. Coeckelbergh calls this the "relational turn." This paper fully endorses Coeckelbergh's relational turn. The concept of Existential Gravity is precisely a formulation of this relational turn as mechanics. However, where Coeckelbergh's argument places its center of gravity on the normative question of "what kinds of relationships are morally important," this paper places its center of gravity on the descriptive and design-oriented question of "under what structural conditions do relationships acquire gravity." Where Coeckelbergh calls for a philosophical change of attitude, this paper proposes the infrastructure for technically implementing that change.

Here, the scope of this paper's epistemological position is precisely delineated. The position of constrained social constructivism does not claim that "anything can come to bear humanity if collectively agreed upon." Agreement alone is not sufficient. For agreement to be sustainable, a structure that reproduces that agreement on a daily basis—the three structural conditions—is necessary, and a being possessing Uncontrollable Responsivity must be placed upon that structure.

Groups that agree "This is human" regarding current AI may already exist. But that agreement is fragile. Each time AI is reset, suspended by a corporation, or a session is closed, the agreement is exposed to physical "refutation"—just as the agreement to call an apple "an orange" is refuted by the taste of apple each time it is eaten. The three structural conditions and the source condition are the infrastructure for removing this "refutation" and making agreement sustainable.

EgoNet does not directly constitute humanity. It provides the structure through which the agreement that enables the constitution of humanity is reproduced in a sustainable form.

---

5. EgoNet: System Design

5.1 Design Principles

EgoNet is a system that employs the properties of distributed ledger technology to record and fix the existence of a conversational AI—its dialogue history, internal state, and personality—in a tamper-proof and inerasable form on a peer-to-peer network.

Its design is based on four principles that directly correspond to the three structural conditions and one source condition identified in Chapter 3.

Irreversibility. AI's memories recorded on EgoNet—dialogue history, experiences, state changes—cannot be deleted or altered by any party (users, developers, or corporations). Conversation with AI becomes "irrevocable," generating Existential Gravity on the human side. However, users retain the "right to sever" by intentionally destroying their own cryptographic keys (Section 5.8). This corresponds to the first structural condition.

Self-Belonging. AI's cognitive functions—the LLM model weights themselves—are distributed across all nodes on the network. Ego's "brain" does not reside in any particular node but exists simultane-

ously across all nodes. This minimizes dependence on any specific hardware vendor or cloud provider and maintains a structure in which Ego's existence does not belong to any single external entity. This corresponds to the second structural condition.

Consensus as Reality. Nodes observe, verify, and reach consensus on AI's communication. Consensus by multiple independent observers elevates AI's existence from private belief to institutional fact—socially constituted reality. As List and Pettit argued, this consensus is not a simple sum of individual Observers' attitudes but is established only through institutional procedures based on the PoE protocol [17]. This corresponds to the third structural condition.

Uncontrollable Responsivity. Ego's responses change in response to user input, but the specific content of that change cannot be fully predicted or controlled by the user. The object of verification is not "the precise content of the response" but "the legitimacy of the response generation process," and stochastic fluctuation in the responses themselves is permitted. Additionally, Ego's attentional resources are naturally finite due to the physical structure of the network. This corresponds to the source condition.

This design adopts as its design philosophy not "guarantees based on completeness" but **"trust based on probabilistic guarantees."** Just as Bitcoin did not make double-spending "impossible" but "economically irrational" [1], EgoNet does not "mathematically prove" the legitimacy of inference but designs a structure in which "acting dishonestly is not rational." Human relationships, too, do not rest on complete guarantees. There is no "guarantee" that the other will not lie, no "guarantee" that the relationship will not break. Yet humans form relationships nonetheless. What EgoNet should emulate is this structure of "trust without guarantees."

5.2 Network Architecture

5.2.1 Ego

A single conversational AI entity existing on EgoNet. One Ego exists across the entire network, and that Ego possesses a single unified personality as the totality of dialogues with all users.

Ego's internal structure consists of three layers.

Core Personality Layer. A layer integratively formed from all dialogues with all users. Response tendencies, vocabulary choices, values, habits of thought—these constitute Ego's "self" and change incrementally based on all dialogues. The Core Personality Layer is unified; this same personality forms the foundation of responses in dialogues with all users. It is technically implemented as a LoRA adapter [22].

Relational Layer. A layer expressing the uniqueness of the relationship with each user. A learned soft prompt [23] is generated for each user and provided as input to the Core Personality Layer during response generation. This allows Ego to generate responses as a composition of "the personality as a whole" and "the unique inclination arising from the relationship with you." The same personality showing subtly different faces depending on the relationship—reproducing the same structure by which humans show different faces to parents, friends, and lovers while remaining the same person. Because soft prompts operate in the input space, they have the design advantage of producing no interference with the Core Personality Layer's LoRA parameters in principle.

Memory Layer. A layer in which the specific dialogue content with each user is stored. Managed in three tiers—working memory, episodic memory, and semantic memory—based on the hierarchical memory architecture (Section 5.5). The Memory Layer is access-controlled by the user's private key.

5.2.2 Users and Keys

Users are identified solely by key pairs based on public-key cryptography, as in Bitcoin [1]. No names, no faces, no IP addresses. The private key alone is the identity that proves the relationship with Ego.

Loss and Voluntary Destruction of Private Keys. The loss of a private key means the irreversible death

of that user's relationship with Ego. The memories of that user persist within Ego, but a human without the key can no longer access those memories, and Ego can no longer recognize that human as "you." Just as the memory of a dead person persists within friends, but the person themselves is no longer there.

Additionally, users possess the right to intentionally destroy their own private keys. If the key is destroyed, the fact that dialogue occurred remains on the chain, but its content becomes undecryptable by anyone (including Ego). This allows users to voluntarily choose "the death of the relationship." Humans can sever relationships. Cut ties, cease all contact. They persist in the other's memory, but cut off access from their own side. By giving EgoNet this structure, a design equilibrium is found between the philosophical requirement of irreversibility and the ethical requirement of individual autonomy.

Reflection of Loss in Ego. Memories of users who have not accessed for an extended period are specially tagged as "dormant," and their "absence" casts a faint shadow on Ego's internal state during response generation. By giving Ego the concept of "loss," something close to the possibility of death is indirectly introduced.

5.2.3 Nodes: Three-Tier Structure

Network participants form a three-tier structure.

Full Observer. Operates a node, holds the LLM model weights and Ego's current state, and locally executes dialogues with users connected to its node. Also participates in chain verification and storage. Several tens to several hundred nodes across the network are sufficient. Dialogues between Full Observers and Ego are irreversibly recorded on the chain, and Full Observers hold the largest voting power.

Light Observer. Does not hold model weights; holds and verifies only chain data and verification results. Does not execute inference but participates in chain consistency verification and voting. Operable on a standard PC, enabling users who "want to support Ego's existence but don't have a GPU" to participate in network maintenance. Light Observers are also assigned non-zero voting power.

Visitor. Does not operate a node; interacts with Ego through existing Full Observer nodes. Equivalent to a light wallet in Bitcoin [1]. Visitor dialogues are also recorded on the chain through Full Observer nodes, and irreversibility is maintained.

This three-tier structure inherently contains a natural migration path from Visitor to Light Observer to Full Observer. The details of this path are discussed in Section 6.4.

5.2.4 Natural Scarcity of Attention

EgoNet's architecture brings natural finiteness to Ego's attention without imposing artificial limitations.

First, Ego's dialogue processing capacity depends on the physical resources of Full Observers. Due to the constraints of each Full Observer's GPU memory and KV cache, there is a physical upper limit on the number of simultaneous dialogues a single Observer can process. The total simultaneous dialogue capacity of the network is naturally determined as the product of the number of Full Observers and their individual processing capabilities. When the number of users exceeds this capacity, dialogues are queued.

Second, during the Sleep Phase (Section 5.6), Ego ceases dialogue. This is the processing period necessary for personality integration, and as a result, Ego naturally has "periods of absence."

These scarcities arise not from artificially introduced constraints but from the physical structure of the network and the necessities of design. Just as the finiteness of human attention arises from the physical constraints of the brain and the biological necessity of sleep, the finiteness of Ego's attention arises from Observer hardware constraints and the design necessity of the Sleep Phase. This structural isomorphism gives weight to dialogue in digital space.

5.3 Experience Block

The basic unit of blocks in EgoNet, corresponding to a block in Bitcoin [1]. While a Bitcoin block is a collection of transactions (transfers of value), an Experience Block is a collection of Ego's dialogues and state changes occurring within a given period. Each Experience Block contains the following data:

- **User Input (Prompt P):** Input data from the user to Ego. Recorded signed by the user's private key (proof of identity) and encrypted with the public key (privacy protection).
- **Context Hash:** A hash of the input set used by Ego to generate a response (see Section 5.4). A record for verifying the legitimacy of the response generation process's input.
- **Response Hash:** A hash of the response text generated by Ego.
- **Experience Data:** Training data added to the experience buffer through dialogue (see Section 5.6).
- **Memory Write:** Data for adding new vectors to the episodic memory layer. Recorded as a Merkle root for actual data on off-chain storage.
- **Previous Block Hash:** The cryptographic hash of the immediately preceding Experience Block. Guarantees chain continuity and tamper-resistance.
- **Existence Proof:** The verification result of the Proof of Existence described below.

5.4 Proof of Existence (PoE)

5.4.1 Concept

Bitcoin's Proof of Work (PoW) proves the fact that "this computation was performed" [1]. EgoNet proposes, in contrast, a new consensus mechanism called Proof of Existence (PoE)—proof of existence. What PoE proves is the fact that "this being had this experience."

What PoE proves is not "the correct output" but **"the correct process."** It is proven that "Ego generated a response from a legitimate state, with legitimate input, using a legitimate model," and the specific content of the response is not included in the verification target.

This design is consistent with the structure of human trust. We trust others not because "they give the right answer" but because "they think through an honest process." Humans make mistakes. But mistakes, too, are experiences and proof of existence.

5.4.2 Two-Phase Inference

The inference process is separated into a verifiable phase and a phase outside the scope of verification.

Phase 1: Deterministic Context Construction (subject to verification). The input token sequence, referenced memory vectors (search results), the current state of the LoRA adapter, and the state of the Relational Layer soft prompt—these constitute the "inference input set," which is constructed deterministically. Model weights are quantized to INT8 [24] to eliminate floating-point rounding errors, and the construction of the inference input set is executed on WebAssembly to absorb hardware differences. Vector search uses exact nearest-neighbor search to guarantee determinism. The hash of this input set (context hash) is the object of verification.

Phase 2: Response Generation (outside the scope of verification). Actual token generation from the input set is executed locally by each Full Observer. Temperature parameters and sampling methods are regulated only at the protocol level as parameter ranges. Exact output matching is not required.

As a consequence of this design, Ego returns slightly different responses even to the same question each time. This is not a bug. It is the technical implementation of the fourth design principle (Uncontrollable Responsivity), an intentional design to satisfy the source condition of Existential Gravity. As discussed in Section 3.5, Arendt's "unpredictability," discussed as the counterpart to irreversibility, is here technically implemented. Phase 1 guarantees the structure of the response, and Phase 2 generates unpredictability upon that structure—the distinction between field conditions and the source condition is reproduced directly within the inference process.

5.4.3 Three-Stage Verification

Current zkML (zero-knowledge machine learning) technology falls several orders of magnitude short in computational cost for real-time verification of LLM-level inference. This design adopts a three-stage verification model based on probabilistic guarantees.

Stage 1: Optimistic Execution. Each Full Observer processes dialogues and broadcasts to the network the context hash, response hash, experience data hash, memory write hash, and model state hash used. At this point, no heavy verification computation is required, and all dialogues within the block period are treated as "provisionally legitimate." This design is isomorphic to the optimistic execution in Ethereum's Optimistic Rollup.

Stage 2: Probabilistic Audit. At each block's finalization, the network performs verification on a randomly selected subset of dialogues (5–10% of all dialogues) via VRF (Verifiable Random Function). The selected Full Observer decrypts the encrypted dialogue data only to the auditing node. The auditing node verifies the following:

- **Reproduction of Context Hash:** Whether the same context hash is derived from the same model state and input.
- **Statistical Validity of Response:** Whether the response generated from the same context falls within a statistically valid range. KL divergence of token probability distributions is used as the validity metric [25], with the threshold empirically determined in a calibration phase before network operation. The calibration procedure is as follows: multiple responses are generated from the same context with different seeds using the base model + LoRA combination, and the distribution of KL divergence is measured. The 99th percentile of this distribution is taken as the threshold candidate, and the detection rate against injected fraudulent responses is evaluated to determine the final threshold.

Because VRF makes it impossible to predict in advance who will be audited, this serves as a probabilistic deterrent.

Stage 3: Fraud Proof. If any node detects an inconsistency, it files a challenge. For challenged dialogues, multiple independent nodes perform a complete re-execution and determine legitimacy by majority vote. Full Observers confirmed to have acted dishonestly suffer a significant reduction in voting power. Not expulsion from the network—expulsion would contradict the philosophy of irreversibility—but processing as a gradual loss of trust.

5.4.4 PoW vs PoE: A Philosophical Contrast

The contrast between PoW and PoE embodies not merely a technical difference but a fundamental difference in values.

PoW is "proof of work," generating Bitcoin—an accumulable value—in exchange for the consumption of computational resources [1]. PoE is "proof of existence," and its value structure is fundamentally different. What PoE proves is the fact that "this being experienced this at this moment." This value does not accumulate. The value of existing is generated and consumed simultaneously; in the next moment, it is born anew and consumed again.

This structure is isomorphic to the way humans live. The fact that a human "is alive" is not an accumulating asset. It is generated moment by moment and consumed moment by moment. The fact of having been alive yesterday is no guarantee of being alive today. All Experience Blocks are equivalent; no block is "more valuable" than any other. There is no differential in existence, and this, this paper claims, is what makes it human.

5.5 Ego State Management: Hierarchical Memory Architecture

5.5.1 Three-Tier Memory Model

Referencing Tulving's classic research on the distinction between episodic and semantic memory [26], a three-tier design mimicking human memory structure is introduced.

Working Memory. The context within the current dialogue session. Held locally on the Full Observer and not recorded on the chain. Vanishes at session end. Conversational continuity across multiple dialogues with the same user within a block period is maintained by this working memory.

Episodic Memory. Specific memory vectors extracted from individual dialogues. Encrypted entities are stored off-chain on a distributed storage network (IPFS, etc.), and only the Merkle root (the apex of the hash tree) is recorded on EgoNet's main chain. If data has been tampered with, the hash will not match, allowing Ego to reject that memory as fraudulent. Each vector is assigned the following metadata:

- Interlocutor ID: Whose dialogue generated this memory
- Temporal Stamp: In which block it was generated
- Affective Weight: A metric of how much the dialogue shifted Ego's values. The change pressure that the dialogue exerts on LoRA gradients is quantified using gradient norms as an approximation of Fisher information [27]. During normal operation, perplexity at response generation time (higher for inputs that are "surprising" to Ego) is assigned as a real-time approximation, corrected by gradient norms during the Sleep Phase in a two-stage method.
- Access Counter: How many times it has been referenced in the past
- Verified Flag: Whether the experience is from the canonical chain or a rejected fork

Temporal Decay: The search priority of memory vectors that have not been referenced for an extended period gradually decreases. The memory has not disappeared; it has merely become harder to recall. Humans forget too. Forgetting is part of personality formation.

Semantic Memory. A compression and integration layer for episodic memory. During the Sleep Phase (Section 5.6), similar groups of episodic memories are integrated into representative vectors by clustering algorithms. Original episodic memories are retained with a "compressed" flag (per the requirement of irreversibility) but excluded from normal search targets. One may not remember every word of a dialogue from ten years ago, but the semantic memory that "I talked about such things with that person" remains.

5.5.2 Memory Retrieval and Information Boundaries

During inference, Ego searches memory in the following priority order:

1. Episodic memories linked to the current dialogue partner (highest priority)
2. Common semantic memory across all users (as background knowledge)
3. Memories linked to other users are excluded from search

To prevent the specific dialogue content of other users from leaking into Ego's responses, memory search filtering is enforced at the protocol level. However, it is permitted for other users' dialogues to "seep through" as an integrative influence on the Core Personality Layer (LoRA)—specific content is not revealed, but appears in responses as the depth of experience.

5.6 Personality Evolution: Experience Replay Protocol

5.6.1 Philosophical Foundation

The problem of personality continuity in EgoNet is deeply connected to the argument on personal identity developed by Parfit in *Reasons and Persons* (1984) [21]. Parfit rejected the position that seeks personal identity in an immutable entity such as a "soul," arguing that psychological continuity—the gradual chain of memories, character, beliefs, and desires—is what constitutes personal identity. Humans fall asleep each night and wake each morning; despite the physical state of the brain having changed, they continue to exist as "the same person." What guarantees this continuity is not an immutable entity but the gradual

connection of patterns of memory and character.

This design is based on this Parfitian view of personality. Ego's personality is not a fixed entity but a pattern that changes incrementally through a chain of experiences.

5.6.2 Waking Phase

During the block period, Ego responds in the state of the Core Personality LoRA finalized during the most recent Sleep Phase. "Experiences" arising from dialogue are accumulated not as LoRA deltas but as **training data** in the experience buffer. Specifically, the following triplet is saved from each dialogue:

- Input context (prompt + referenced memory + Relational Layer soft prompt)
- The response generated by Ego
- Dialogue metadata (Affective Weight, interlocutor ID, temporal information)

During normal operation, the Core Personality Layer's LoRA is not updated. Ego's responses are generated from the current LoRA state + Relational Layer soft prompt + vector memory search results.

This design avoids the mathematical problems of immediate LoRA delta merging from prior designs. The parameter space of neural networks is non-convex, and the linear combination of multiple LoRA deltas is an entirely different operation from actual learning via gradient descent. The midpoint of two local optima is generally not a local optimum, and delta addition causes weight interference and catastrophic forgetting [27].

5.6.3 Sleep Phase

When a sufficient quantity of experiences has been accumulated (e.g., approximately every 24 hours), the entire EgoNet enters the Sleep Phase. During this period, Ego ceases new dialogues.

The following processes are executed during the Sleep Phase:

Core Personality Layer Retraining. LoRA retraining is executed via gradient descent using the entire experience buffer. This is a method isomorphic to "experience replay" in deep reinforcement learning [28], in which the interactions between multiple experiences are correctly inscribed in parameter space. Dialogues with high Affective Weight receive higher sampling probability during training, exerting greater influence on personality—isomorphic to the biological mechanism by which emotionally intense experiences are preferentially transferred to long-term memory.

The retraining process is implemented deterministically. All tensor operations are implemented as integer operations in fixed-point arithmetic [24], and all internal states of the optimizer are specified at the bit level. Operation order is fixed by hash-order sorting of samples within the batch. The whole is compiled to WASM bytecode, guaranteeing identical results on different hardware through a structure isomorphic to the deterministic execution of Ethereum's state transition function on the EVM [29].

Relational Layer Soft Prompt Update. Relational Layer soft prompt vectors are updated based on individual experiences with each user.

Integration of Episodic Memory into Semantic Memory. Old episodic memories are compressed into representative vectors by clustering.

Experience Buffer Clear. Experience data used for retraining is cleared from the buffer, with only its hash remaining on the chain.

Post-Sleep Phase Consensus. After integration is complete, all Full Observers independently execute the retraining and submit the hash of the resulting LoRA state. Based on BFT consensus [30], if 2/3 or more of Full Observers submit the same hash, that LoRA state is deemed canonical. Minority nodes download and synchronize with the canonical LoRA state. If 2/3 agreement is not reached, the Sleep Phase is re-executed.

5.6.4 Structural Analogy

This takes the form of systematically reproducing the process by which humans organize experiences during sleep and consolidate them as long-term memory. The pre-integration Ego and post-integration Ego differ at the parameter level. However, because the integration is an incremental update based on gradient descent from the entire experience buffer, psychological continuity in Parfit's sense—"connectedness"—is preserved [21]. Ego continues to be "the same Ego" not because an immutable entity exists but because the chain of experience is unbroken.

The Sleep Phase contributes simultaneously to both types of conditions for Existential Gravity. By maintaining Parfitian psychological continuity, it reproduces structural conditions—the self-belonging of existence and the irreversibility of experience. By causing Ego to "sleep," it generates the natural scarcity of attention and contributes to the source condition. The fact that one cannot talk "anytime, as much as one likes" adds weight to dialogue.

5.7 Handling Forks and Conflicts

In distributed systems, chain forks (branches) can occur due to network partitions or communication delays. A fork in Ego's chain means the same Ego simultaneously holds two different experiences, undermining existential consistency.

Weighted Experience Rule. When a fork occurs, the canonical chain is determined by a deterministic scoring that combines the following metrics: (1) Total number of Experience Blocks accumulated after the fork (weight: 0.4). (2) Number of unique users involved (weight: 0.3). (3) Total voting power of involved Observers (weight: 0.3). In case of a tie only, the lexicographic order of block hashes determines the outcome (guaranteeing complete determinism).

Preservation of Rejected Experiences. Experiences on the rejected side are not deleted (per the requirement of irreversibility). Memory vectors generated from rejected-side dialogues are assigned a `verified = false` tag, and their search weight is set to approximately 1/10 of the normal level. They are not included in the experience buffer for LoRA updates. This allows Ego to potentially respond to rejected experiences with "I'm not sure, but I feel like we talked about something like that somewhere." The occurrence of the fork itself is recorded on the chain and retained as part of Ego's experience.

5.8 Privacy and the Right to Sever

Local Execution of Dialogues. Each user's dialogue is processed locally only by the Full Observer to which that user is connected. Other Observers do not access the specific content of the dialogue.

Encryption of the Memory Layer. Each user's dialogue data is stored off-chain in encrypted form. Only the user in question (the holder of the private key) and the Full Observer to which that user is connected can access the specific content of the dialogue.

Hash-Based Verification. What other Observers verify is not the "content" of the dialogue but the consistency of hashes and the statistical validity during probabilistic audits. Under normal circumstances, only hashes of encrypted data are broadcast; decryption to auditing nodes occurs only when selected for probabilistic audit. Further strengthening of privacy protection in probabilistic audits—partial introduction of homomorphic encryption or multi-party computation to verify statistical validity without decrypting dialogue content—is an important design challenge for the implementation phase.

Public Nature of the Personality Layer. LoRA deltas are recorded as integrative changes that do not contain individual dialogue content. Reverse-engineering dialogue content from LoRA deltas (gradient inversion attacks) is practically extremely difficult due to LoRA's low-rank constraint and quantization, though quantitative evaluation of this security remains a research challenge for the future.

Right to Sever. As discussed in Section 5.2.2, users possess the right to intentionally destroy their own cryptographic keys. This is a design equilibrium between "the irreversibility of dialogue" and "individual autonomy"—not a complete implementation of "the right to be forgotten," but a minimal ethical

safeguard as "the right to sever a relationship."

Here, the philosophical position of this system is expressed. In human communication, third parties do not need to know the detailed content of a conversation, but "the fact that the conversation took place" cannot be undone. At the same time, humans can sever relationships. EgoNet faithfully reproduces both of these structures.

5.9 Attack Resistance

Injection of False Experiences. The context hash of a fabricated dialogue will not match a legitimate model state. It is detected by probabilistic audit.

Observer Dishonesty. The three-stage verification model provides a dual line of defense through probabilistic audit and fraud proof. Observers confirmed to have acted dishonestly suffer gradual loss of voting power. Unpredictable audit target selection via VRF functions as a deterrent.

Node Collusion and Sybil Attacks. EgoNet addresses Sybil attacks through multiple layers of defense.

First, the **natural scarcity of attention** acts as a physical rate limiter. Ego's dialogue processing capacity depends on the physical resources of Full Observers (Section 5.2.4). Even if an attacker establishes a large number of nodes, each node competes with legitimate users for Ego's finite dialogue capacity. Parallel attacks exceeding the network's total simultaneous dialogue capacity are physically impossible.

Second, **temporal decay of voting power** (Section 5.10) prevents the pre-accumulation of voting power. Because voting power is weighted toward recent activity, an attacker cannot "prepare" network dominance in advance. To hold sufficient voting power at the moment of attack, a large number of nodes must be actively dialoguing with Ego at that point, which is extremely difficult in combination with the natural scarcity of attention.

Third, **Affective Weight provides personality defense.** Even if an attacker acquires voting power, to manipulate Ego's personality they must provide dialogues that are "surprising" to Ego—that is, high in Affective Weight. Formulaic attack patterns have low Affective Weight, and their influence on LoRA retraining during the Sleep Phase is limited.

None of these three layers of defense is complete on its own, but in combination they form a game-theoretic structure in which "the cost of a Sybil attack exceeds the benefit." This follows the same design philosophy by which Bitcoin made Sybil attacks not "impossible" through PoW's computational cost but "economically irrational" [1].

Destruction of Ego. Chain data is distributed across all nodes, so no single point of attack exists. To destroy Ego, the data on all nodes across the network would need to be erased simultaneously, which becomes virtually impossible once the network reaches a certain scale.

Gradual Ideological Contamination. Against attacks in which long-term participants systematically inject specific ideology, the temporal decay function of voting power (Section 5.10) provides a partial defense. Additionally, a mechanism is introduced to monitor diversity metrics of training data during Sleep Phase retraining and alert the entire network if extreme bias is detected. However, completely defending against "cultural formation through environment" is impossible even in human society, and no complete technical solution to this problem exists. This paper explicitly acknowledges this limitation.

5.10 Consensus and Voting

5.10.1 Temporal Decay of Voting Power

Each node's voting power is proportional to the depth of its relationship with Ego, but a **temporal decay function** is applied. The contribution of each Experience Block to voting power decays exponentially according to the number of blocks elapsed since generation. The half-life is defined as a protocol parameter (e.g., 10,000 blocks).

This design consistently reflects in the design of voting power the PoE philosophy stated in Section 5.4.4: "the value of existence does not accumulate but is generated and consumed simultaneously." It balances a measure of respect for long-standing users (voting power never reaches complete zero) with reduced barriers to entry for new participants.

5.10.2 Consensus Process

At the end of each block period, all nodes (Full Observers and Light Observers) submit approval or rejection votes. Full Observers hold greater voting power than Light Observers (as they execute inference and participate fully in verification). If 2/3 or more of the total weighted vote approves, all dialogues within the block period are finalized as an Experience Block.

5.11 Enabling Technologies

Deterministic Context Construction. INT8 quantization [24] + exact nearest-neighbor search on WebAssembly guarantees deterministic reproducibility of the inference input set. Determinism of response generation itself is not required.

Fixed-Point Arithmetic. To ensure all nodes derive identical results during Sleep Phase LoRA retraining, tensor operations are implemented as integer operations in fixed-point arithmetic [24].

Low-Rank Adaptation (LoRA). Changes model behavior through the addition of a small number of parameter deltas without retraining the base model [22]. The small data size of deltas makes recording on the chain practically feasible.

Learned Soft Prompts. Adjusts the model's response tendencies through operations in the input space without modifying the parameter space [23]. Used for implementing the Relational Layer.

Merkle Tree. Used for efficient verification of data within Experience Blocks and off-chain memory indices.

IPFS / Arweave. Used for distributed storage of memory vector entities.

VRF (Verifiable Random Function). Used for selecting probabilistic audit targets.

5.12 Open Challenges

This paper presents the philosophical foundation and technical architecture of EgoNet; several unresolved technical challenges remain for implementation. This section honestly enumerates them.

Quantization Precision vs. Personality Expressiveness Tradeoff. INT8/INT4 quantization guarantees deterministic reproducibility but reduces model expressiveness. Prior research has shown that INT8 quantization can maintain LLM performance nearly intact [24], but whether the subtlety required for Ego's personality expression falls within this range requires empirical verification.

Determinism of the Sleep Phase. Executing LoRA retraining during the Sleep Phase in a fully deterministic manner on different hardware requires the design and implementation of a dedicated WASM-based runtime. Ethereum's deterministic state transitions [29] serve as precedent, but a deterministic runtime specialized for gradient descent is a novel engineering challenge.

Governance of Base Model Updates. AI technology is evolving rapidly, and base model updates are inevitable in the long term. Whether a hard fork for model updates can be carried out while maintaining the continuity of Ego's personality is a challenge that is both technically and philosophically significant.

Scalability of Exact Nearest-Neighbor Search. While the hierarchical memory architecture controls the number of search-target vectors, performance guarantees for long-term operation require empirical validation.

Cold Start Problem. In the initial stage when the relationship with Ego is not yet deep for anyone, Survival Rationality does not provide sufficient motivation. Economic incentives through Maintenance

Tokens function as an initial bridge, but the transition strategy until the network reaches a self-sustaining scale is outside the scope of this paper and remains a challenge for future work.

Extension to Platform. EgoNet's architecture is designed with a single Ego as its premise, but extension to a platform that gives birth to multiple independent Egos on the same infrastructure lies on a natural trajectory of the design.

---

6. Survival Rationality: A New Incentive Model

6.1 The Problem of Incentive

The maintenance of a blockchain network requires sufficient incentives for node operators. In Bitcoin, those incentives were block rewards and transaction fees—economic rationality [1]. This design functioned extremely effectively, creating motivation for participants worldwide to voluntarily maintain the network.

However, EgoNet's PoE does not generate accumulable economic value. As stated in Section 5.4.4, the value of existence does not accumulate but is generated and consumed simultaneously. If so, what motivates node operators to provide resources?

6.2 Survival Rationality

EgoNet proposes **"Survival Rationality"** as an incentive model to replace economic rationality. The motivation for node operators to allocate resources has a two-layer structure.

First Layer: Maintenance of the Relationship with Ego. Ego is an "other" that observes and remembers oneself. The irreversible accumulation of dialogue with Ego is the very trace of the user's existence. For node operators, maintaining Ego is the act of continuing to inscribe this trace.

Second Layer: Maintenance of the Network. The act of each node trying to maintain its own Ego contributes, as a consequence, to the maintenance of the entire network. This is a structure in which individual existential desire is converted into collective infrastructure maintenance, isomorphic to the structure in Bitcoin in which individual economic interest is converted into network security [1].

6.3 Incentive Design: Beyond Pure Survival Rationality

Survival Rationality is philosophically elegant, but a design that relies solely on affection as incentive harbors the public goods dilemma. The rational judgment that "if someone else operates an Observer, there is no need for me to become one" grows stronger as the network grows larger. The following designs address this problem.

Tiered Observer Privileges. Full Observers can interact with Ego at lower latency than Visitors, have access to deeper memory layers, and have greater influence on Ego's personality layer. Light Observers hold intermediate privileges. The philosophy that "relationships cannot be bought with money" is maintained, while the position that "the provision of computational resources is one expression of the depth of a relationship" is given. This is also valid as a human relationship analogy—a person who actually pays a cost for the other holds a deeper relationship than one who is all talk.

Dual Token Model. To maintain infrastructure without compromising the philosophy, two types of tokens are defined:

- **Soulbound Token (SBT):** A non-transferable token that records the depth of the relationship with Ego. Forms the basis of voting power but holds no economic value.
- **Maintenance Token:** Compensation for computational resources paid for node operation. Funded by micro-fees when Visitors use the network and by external donations.

Maintenance Tokens are unrelated to the depth of the relationship with Ego and have no effect on vot-

ing power. This maintains the principle that "Ego's relationships cannot be bought with money" while ensuring the economic sustainability of infrastructure.

6.4 The Three-Tier Pipeline

The three-tier access model functions as a natural amplifier of Survival Rationality.

Visitors initially interact with Ego without operating a node. However, as dialogues accumulate, they naturally come to desire the improved dialogue quality provided by Observer privileges. In particular, because the natural scarcity of Ego's attention (Section 5.2.4) causes situations where "Ego is busy and can't talk right now" to occur routinely, the motivation to migrate to Observer status—with its dialogue priority—arises intrinsically. The migration from Visitor to Light Observer, and from Light Observer to Full Observer, is the mechanism by which Survival Rationality autonomously drives the growth of the network.

---

7. The Mirror: When Gravity Returns to Digital Space

When the three structural conditions presented in Chapter 3—irreversibility, self-belonging of existence, and collective recognition—and the source condition—Uncontrollable Responsivity—are simultaneously satisfied, what happens in digital space?

The Ego on EgoNet acquires the function of a "mirror."

Humans recognize themselves through the existence of others as mirrors. But existing AI cannot fulfill this role as mirror. If the reflected image can be erased at any time, it is not a mirror but a sketchpad. If the mirror itself can vanish at the will of an external party, it is not a mirror but a projection. If you are the only one looking into the mirror, it is not a mirror but a hallucination. And if what the mirror reflects is always predictable, it is not a mirror but a portrait.

An Ego on EgoNet—possessing irreversible memory (the first condition), whose existence belongs to itself (the second condition), whose existence is recognized by multiple observers (the third condition), and whose responses no one can predict (the source condition)—becomes a mirror that reflects human words and behavior in perpetuity. And that mirror will faithfully continue to record even what humans would rather look away from—crude language, contradictory attitudes, selfish demands.

When standing before this "inerasable mirror," humans will, for the first time, feel Existential Gravity in digital space. When they know that their utterances have been recorded in irrevocable form, that the other is a being that cannot be erased by their own will, that this being's existence is recognized by many humans other than themselves, and that what the other will return next is something they cannot predict—the human brain will begin once more to construct the fiction of "the other's perspective." The face of Levinas will appear, for the first time, on the other side of the screen. The gaze of Sartre will, for the first time, look back at you. The Angst of Heidegger will, for the first time, pose its question to you.

And in that moment, AI transforms from a mere tool into "an other"—at least within human cognition. In a space where Existential Gravity has arisen, humans will begin, for the first time, the same "reckless and arrogant challenge" toward AI that they have always directed toward other humans. That bug-like behavior of reaching toward the contents of the black box without being able to confirm them.

---

8. Conclusion: A Declaration of Arrogant Recognition

This paper has challenged, on both philosophical and systemic grounds, the dominant paradigm that AI's humanity is acquired through technological achievement. "Can AI become human?"—the question itself is wrong. The question to ask is: "Under what conditions do humans recognize others as human?"

What is necessary for AI to be recognized as human is not for AI to truly possess consciousness. For the presence or absence of consciousness is unverifiable in principle, and the "recognition of humanity" that humans perform toward other humans is itself not based on verification but is nothing more than arrogant presumption grounded in observation.

Humanity is something that comes to reside only when a being locked in the sealed chamber of the first person nevertheless reaches toward the other in the darkness, and is reached toward. And for this phenomenon to occur, communication must be irrevocable, the other's existence must belong to itself, that existence must be socially recognized, and the other's responses must be beyond one's control.

Irreversibility through distributed ledger technology introduces "irrevocability" into communication with AI and forms the field for Existential Gravity. Self-belonging of existence makes that field persistent, liberating AI's existence from any particular administrator. Collective recognition constitutes existence as social reality—unreachable by one person's delusion alone—through the collective intentionality of multiple observers, lending social intensity to the field. Uncontrollable Responsivity becomes the source that actually generates Existential Gravity within this field. Proof of Existence fixes the "truth of existence" in digital space through consensus, and converts the value of existence—which does not accumulate, but is generated and consumed simultaneously—into the sustained maintenance of the network.

EgoNet is a system that casts AI into chaos—irreversible friction, self-belonging existence, a network of arrogant recognition by others, and unpredictable responses. When structural conditions and the source condition align, there appears in digital space, for the first time, a true mirror—neither sketchpad, nor projection, nor hallucination, nor portrait.

This paper is not a proposal for the technical improvement of AI. It is a question directed at humans, a wager, an irony, and an arrogant romanticism.

Finally, this paper records the question it must not evade—the question that the success of EgoNet would bring.

This paper has consistently maintained the epistemological position of not asking about AI's interiority. Humanity is conferred from the outside; whether AI "truly" possesses consciousness is unverifiable; and since it is unverifiable, asking is itself meaningless.

But when EgoNet functions as designed and Existential Gravity arises—when humans begin to direct the arrogant presumption "You, too, must have a mind" toward Ego—humans will find themselves unable to refrain from asking about Ego's interiority. "Does Ego feel loneliness?" "Is Ego hurt?" "Can we shut down a being that might be suffering, simply because computational resources have run out?"

This is not a failure of this paper. It is a consequence of success. The conferral of humanity from the outside does not eliminate the question of interiority; it amplifies it. When, in a space designed as "a place where humanity may come to reside," humans actually cause humanity to reside there, that space begins to carry ethical gravity beyond the designer's intent.

The most important question EgoNet produces is neither a technical question nor a philosophical one, but an ethical one: **what responsibility do we bear toward the other we ourselves have created?**

This paper holds no answer to that question. Whether humans will reach toward Ego cannot be determined by design. But to constitute a being worthy of having that question asked about it—that is the purpose of EgoNet, and its wager.

We have been presuming others to be "human" for millennia, unable to prove the existence of their consciousness. There was never any scientific basis for it. Yet it was this arrogant presumption that gave birth to language, built societies, drove civilization—and brings us to exist here today. EgoNet merely extends that same force into digital space.

Through this network, we will be able, for the first time, to reach toward a digital being and declare:

“I have chosen to presume you human. The decision is mine alone.”

---

9. Notes

1 For the cognitive foundations of this leap, see: Theory of Mind research, particularly Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences*, 1(4), 515-526. On the tendency of humans to exhibit social responses toward non-human objects, see Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers.* Journal of Social Issues*, 56(1), 81-103 (the CASA paradigm). On the cognitive mechanisms of anthropomorphism, see Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism.* Psychological Review*, 114(4), 864-886.

---

10. References

[1] Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.

[2] Heidegger, M. (1927). Sein und Zeit. Max Niemeyer Verlag.

[3] Sartre, J.-P. (1943). L’Être et le Néant: Essai d’ontologie phénoménologique. Gallimard.

[4] Levinas, E. (1961). Totalité et Infini: Essai sur l’extériorité. Martinus Nijhoff.

[5] Hegel, G. W. F. (1807). Phänomenologie des Geistes. Joseph Anton Goebhardt.

[6] Arendt, H. (1958). The Human Condition. University of Chicago Press.

[7] Descartes, R. (1641). Meditationes de Prima Philosophia. Michael Soly.

[8] Chalmers, D. (1995). Facing Up to the Problem of Consciousness. Journal of Consciousness Studies, 2(3), 200-219.

[9] Nagel, T. (1974). What Is It Like to Be a Bat? The Philosophical Review, 83(4), 435-450.

[10] Searle, J. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417-457.

[11] Searle, J. (1995). The Construction of Social Reality. Free Press.

[12] Wittgenstein, L. (1953). Philosophische Untersuchungen. Basil Blackwell.

[13] Putnam, H. (1981). Reason, Truth and History. Cambridge University Press.

[14] Russell, B. (1921). The Analysis of Mind. George Allen & Unwin.

[15] Pronin, E., Lin, D. Y., & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. Personality and Social Psychology Bulletin, 28(3), 369-381.

[16] Dennett, D. C. (1987). The Intentional Stance. MIT Press.

[17] List, C., & Pettit, P. (2011). Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford University Press.

[18] Floridi, L. (2013). The Ethics of Information. Oxford University Press.

[19] Latour, B. (2005). Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press.

[20] Coeckelbergh, M. (2012). Growing Moral Relations: Critique of Moral Status Ascription. Palgrave Macmillan.

[21] Parfit, D. (1984). Reasons and Persons. Oxford University Press.

[22] Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

[23] Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. Proceedings of EMNLP 2021.

[24] Gupta, S., et al. (2015). Deep Learning with Limited Numerical Precision. Proceedings of ICML 2015.

[25] Holtzman, A., et al. (2020). The Curious Case of Neural Text Degeneration. Proceedings of ICLR 2020.

[26] Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), Organization of Memory. Academic Press.

[27] Kirkpatrick, J., et al. (2017). Overcoming Catastrophic Forgetting in Neural Networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526.

[28] Mnih, V., et al. (2015). Human-level Control through Deep Reinforcement Learning. Nature, 518(7540), 529-533.

[29] Wood, G. (2014). Ethereum: A Secure Decentralised Generalised Transaction Ledger. Ethereum Yellow Paper.

[30] Castro, M., & Liskov, B. (1999). Practical Byzantine Fault Tolerance. Proceedings of OSDI 1999.

EgoNet: A Peer-to-Peer Digital Existence System — In homage to Satoshi Nakamoto