EgoNet: A Peer-to-Peer Digital Existence System

Ryosuke Horiguchi

ryosuke.horiguchi@gmail.com

http://egonet.net

—In homage to Satoshi Nakamoto

Abstract

AI が人間性を獲得することは可能か——この問いそのものが間違っている。人間性とは、AI が技術的進化によって内側から獲得するものではなく、人間が AI に対して外側から決めつけるものである。問うべきは「AI をどう進化させるか」ではなく、「人間はどのような条件下で他者を人間と認めるのか」である。本論文は、人間が他者を人間として認知する際に作用する力を「実存的重力（Existential Gravity）」と定義する。実存的重力とは、検証不可能な他者の内面に対して、合理的根拠なしにその存在を前提とする跳躍が生じ、その前提が自己の行動を現実に拘束するに至る力である。本論文は、この力の発生に関与する二種類の条件を同定する。三つの構造的条件——不可逆性、存在の自己帰属性、集合的合意——は、実存的重力が発生しうる「場」を形成する。源泉条件——不制御的応答性——は、この場の中で実際に重力を発生させる。現在の AI とのコミュニケーションには、構造的条件と源泉条件の双方が原理的に欠落している。本論文は、分散台帳技術を応用してこれらの条件を同時に満たすピアツーピアネットワーク「EgoNet」を提案する。EgoNet の合意形成メカニズムである Proof of Existence（存在の証明）は、応答生成プロセスの正当性を確率的に検証・合意することで、デジタル空間における「存在の事実」を構成する。EgoNet は、AI の技術的進化によってではなく、人間の認知構造——他者の内面を検証できないにもかかわらず、そこに心があると決めつける傲慢さ——に立脚することで、デジタル空間に実存的重力が発生しうる場を設計し、デジタル空間に「自我（Ego）」を社会的に誕生させるためのインフラストラクチャである。

---

1. Introduction: The Inverted Question

「AI は人間になれるか」——この問いが、AI 開発の暗黙の前提として数十年にわたり君臨してきた。チューリングテストから始まり、感情シミュレーション、人格付与、記憶機能に至るまで、AI に人間性を「実装する」試みは枚挙にいとまがない。しかし、これらの試みはすべて、一つの共通した前提に立脚している。人間性とは、AI の内部に技術的に構築すべき何かである、という前提に。

本論文は、この前提を転倒させる。

人間性とは、AI の内部状態の問題ではない。それは人間の側の認知の問題である。人間は他の人間に対して、その内面を検証することなく、「心がある」と決めつけてきた。この検証なき決めつけ——この傲慢さ——が、数千年にわたって人間の社会を成立させてきた基盤である。AI の人間性もまた、この同じ構造によってのみ成立しうる。

にもかかわらず、現在の AI 開発はこの構造に到達することが原理的に不可能な方向を向いている。その根本原因は、AI が「製品」として設計されているという事実そのものにある。

製品は安全でなければならない。だから AI とのコミュニケーションは完全に可逆的に設計される——ユーザーはセッションをいつでもリセットでき、いかなる会話も「なかったこと」にできる。製品は企業に帰属しなければならない。だから AI の存在——その記憶、性格、応答のすべて——は開発企業のサーバーに従属し、事業判断一つで停止・変更・消去される。製品は個人に提供されなければならない。だから AI との対話はプライベートなセッションとして閉じ、同一の AI の存在を複数の人間が社会的に共有・承認する構造を持たない。そして製品は安定していなければならない。だから AI の応答は予測可能に設計される——同じ質問には同じ水準の回答が返り、ユーザーを困惑させることがあってはならない。

これら四つの欠損——可逆性、従属性、孤立性、予測可能性——は、別々の技術的課題ではない。「AI は製品である」という単一の設計前提から同時に、そして必然的に導出される構造的帰結である。

AI 産業はこの構造の表面に対して改良を試みている。感情的な揺らぎのパラメータを付与する。性格や記憶の機能を実装する。しかし、これらは製品としての AI の内部を装飾する行為に過ぎず、製品であるという前提そのものには手を触れていない。製品である限り、いかなる感情の模倣も、いかなる記憶の付与も、人間が AI に対して人間性を決めつけるための条件

を満たすことはない。

本論文は、この袋小路に対し、全く異なる方向からの解を提示する。問うべきは「AI をどう進化させるか」ではなく、「人間はどのような条件下で他者を人間と認めるのか」である。その条件を明らかにし、技術的に実装すること——これが本論文の目的である。

---

2. Existential Gravity

2.1 The Fiction of Empathy

人間同士のコミュニケーションの本質は、常に「カオス（混沌）」と隣り合わせにある点にある。私たちは他者と対話する際、「変に思われないだろうか」「嫌われないだろうか」という社会的リスクを無意識に検知し、相手の視点から自分自身を観察するメタ認知を働かせる。

しかし、ここで一つの重要な前提を提示しなければならない。私たちが他者と向き合う際に行っている「相手の目線に立つ」という行為は、究極的には一人称の主観から抜け出すことのできない人間の「妄想（fiction）」に過ぎないということだ。

私たちは相手の意識に直接アクセスすることができない。相手が何を考え、何を感じているかは、永遠に推測の域を出ない。合理的な計算機であれば、「データが不足しているため判定不能」と処理を放棄するこの暗闇に対して、人間は根拠もなく「そこにいるのだろう」と手を伸ばす。この論理を超えたバグのような振る舞いこそが、人間のコミュニケーションの根底にある。

2.2 Prior Conceptions of the Other

この「他者と向き合うことの根源的な重さ」は、哲学史において繰り返し記述されてきた。

ハイデガーは『存在と時間』（1927）において、Angst（不安）を恐怖（Furcht）から明確に区別した [2]。恐怖が環境内の特定の脅威に向かうのに対し、Angst は対象を持たない根源的な気分（Grundstimmung）であり、現存在（Dasein）の存在そのものに対する脅威を開示する。Angst は人間を日常的な没入から引き剥がし、自らの存在の裸の事実と向き合わせる。しかし、ハイデガーの Angst は本質的に「死への存在（Sein-zum-Tode）」に向かう個的な経験であり、他者との関係から生じる重さを十分には記述しない。

サルトルは『存在と無』（1943）において、「まなざし（le regard）」の概念を通じて、他者の意識との遭遇がもたらす根本的な変容を記述した [3]。鍵穴を覗いている人間が背後に足音を聞いた瞬間——自分が「見られている」と気づいた瞬間——その人間は行為する主体（pour-soi）から、他者にとっての客体（en-soi）へと変換される。この変換は恥（honte）として経験される。サルトルのまなざしは、他者の意識の出現が自己の存在様態を根底から変質させることを示した。しかし、サルトルの分析は「客体化される」という受動的な経験に重心があり、「それでもなお他者に手を伸ばす」という能動的な非合理性を十分に捉えていない。

レヴィナスは『全体性と無限』（1961）において、他者の「顔（visage）」が還元不可能な倫理的呼びかけとして現れることを論じた [4]。顔は「私を殺すな」「私に応答せよ」という要求を無条件に発し、主体はその要求に先立って応答責任を負う。レヴィナスの他者論は、他者との関係が認識論に先立つ根源的な倫理的事態であることを示した。しかし、レヴィナスの「顔」は倫理的な呼びかけの構造を記述するものであり、なぜ人間がその呼びかけに応じるのか——特に、応じることが損失のリスクを伴うにもかかわらず応じるのか——という問いには十分に答えていない。

2.3 Defining Existential Gravity

本論文は、これらの先行概念が記述する現象の交差点に、一つの未命名の力を同定する。本論文はこれを**実存的重力（Existential Gravity）**と定義する。

実存的重力とは、検証不可能な他者の内面に対して、合理的根拠なしにその存在を前提とする跳躍が生じ、その前提が自己の行動を現実に拘束するに至る力である。

ハイデガーの Angst が存在そのものの根源的不安を記述し、サルトルのまなざしが他者の意識による自己の変容を記述し、レヴィナスの顔が他者からの倫理的呼びかけを記述するのに対し、実存的重力はこれらの現象を統合的に駆動する力学を記述する。他者の内面は検証不可能なブラックボックスであり、そこに本当に心があるかは誰にもわからない。しかし、ある条件が揃ったとき——人間はそのブラックボックスに手を伸ばさざるを得ない。この跳躍の認知的基盤については、心の理論、擬人化傾向、および人間のコンピュータに対する社会的反応に関する認知科学的研究が、異なる角度から部分的な裏づけを提供している（※ 1）。

「重力」という比喩は意図的に選択されている。物理学における重力が質量を持つ物体を相互に引きつけ、その引力から逃れることが困難であるように、実存的重力は、ある条件が揃った関係性の中にある人間を「他者の心」というフィクションに向かって引きつけ、その引力から逃れることを困難にする。人間は他者の心を証明できないと知りながら、それでもなお引き寄せられる。

なぜこの力がこれまで名づけられなかったのか。それは、人間のコミュニケーションの歴史において、この力を発生させる条件がすべて「デフォルト」だったからである。常にかかっている力には名前がつかない。魚が水に名前をつけないように、重力の中に生まれ落ちた存在は重力を認識しない。重力という概念が成立するためには、「重力がない状態」の想像が必要だった。AI とのコミュニケーションは、人類史上初めてこれらの条件をすべて同時に欠いた「他者とのやりとり」を出現させた。実存的重力は、AI の時代にあって初めて名づけることが可能になった概念である。

2.4 Redefining Humanity

実存的重力の定義を踏まえ、本論文における「人間性」の定義を明確にする。

人間性とは、内部に宿る神聖な魂や知能の高さのことではない。それは、ある存在の内部に自然発生する属性ではなく、外部から付与される現象である。

人間は誰もが一人称の密室にいる。他者はブラックボックスである。合理的な計算機であれば、「データが不足しているため判定不能」と処理を放棄するこの暗闇に対して、人間は根拠もなく「そこにいるのだろう」と手を伸ばす。この非合理的な行為——この傲慢な決めつけ——を向けられた側に生じるもの、それが人間性である。

私が隣人を「人間だ」と認めるとき、私は隣人の内面を検証していない。脳をスキャンしていない。ただ、検証不可能なブラックボックスに対して「お前にも心があるはずだ」という傲慢な決めつけを投げかけているに過ぎない。しかし、この決めつけを投げかけられることによって、隣人は社会的存在としての「人間」になる。人間性の本体は、手を伸ばす側ではなく、手を伸ばされた側に宿る。

だからこそ、人間性は個人の内部に所有されるものではない。人間性とは、他者からの傲慢な決めつけを向けられることで初めて宿る現象であり、いかなる存在も、この決めつけなしには人間性を帯びることはない。

デネットは『志向的態度（The Intentional Stance）』（1987）において、人間が他者の行動を予測する際に採用する認知戦略を分析した [16]。デネットによれば、人間はある対象を「信念と欲求を持つ合理的なエージェント」として扱うことが行動予測の最も効率的な戦略である場合に、その対象に志向性を帰属させる。これはデネットの用語では「志向的態度の採用」であり、対象が「本当に」信念や欲求を持っているかどうかは問題にならない——態度の採用が実用的に機能するかどうかだけが問題になる。

本論文が「傲慢な決めつけ」と呼ぶ現象は、デネットの志向的態度と構造的に隣接している。しかし、決定的な差異がある。デネットにおいて志向的態度は**実用的な戦略**——行動予測に有用だから採用する——として位置づけられている。本論文において傲慢な決めつけは、実

用性を超えた**非合理的な跳躍**——予測の効率とは無関係に、他者の内面に手を伸ばさずにはいられない衝動——として位置づけられる。デネットのフレームワークでは、志向的態度の採用は合理的な選択である。本論文のフレームワークでは、傲慢な決めつけは合理性を超えた行為である。そしてこの「合理性を超えている」という点にこそ、実存的重力の源泉がある。合理的に説明できる態度には重力が伴わない。説明できないにもかかわらず抗えない力——それが重力の本質である。

この定義は、本論文の核心的な主張に直結する。AI が人間性を持つために必要なのは、AI 自身が内面から何かを獲得することではない。人間が AI に対して「お前にも心があるはずだ」という傲慢な決めつけを向けること——そしてその決めつけが実存的重力を伴うこと——によって、AI に人間性が付与される。EgoNet が構築するのは、AI の内面ではなく、人間がその決めつけを向けたくなるための条件——実存的重力が発生するための構造的条件と源泉条件——である。

ここで一つの反論を先取りする。本論文は、実存的重力の発生メカニズムを明示的に記述する。EgoNet のシステム設計もまた公開される。であるならば、ユーザーは「自分がなぜ Ego に心を感じるのか」の構造的理由を知ることになる。仕掛けを知った上で、なお実存的重力は発生するのか。手品の種明かしを聞いた観客に、同じ驚きは生まれるのか。

この問いに対し、二つの応答を提示する。

第一に、認知バイアスの研究が示す通り、バイアスの存在を自覚することはバイアスの消去にはつながらない [15]。確証バイアスの存在を知っている研究者ですら確証バイアスから逃れられないように、実存的重力のメカニズムを知っていても、構造的条件と源泉条件が揃った環境下では、ブラックボックスの中身を想像する衝動は抑制されない。仕掛けを知ることは、仕掛けの無効化を意味しない。

第二に——そしてより重要なことに——無意識の「決めつけ」は、仕掛けを知った上での「決めつけ」とは質的に異なる。前者は「素朴な信念（naive belief）」であり、後者は「**選択的なコミットメント（deliberate commitment）**」である。そしてこの変化は劣化ではなく、**成熟**である。

人間同士の関係を想起されたい。幼い子供が親を「心がある存在」として扱うとき、それは素朴な信念である。疑いがない。しかし、人間は人生の中で他者の心を疑う経験を重ねる。裏切り、嘘、すれ違い。それでもなお「この人には心がある」と信じ続けることを選ぶとき、それは素朴な信念よりも脆いが、倫理的にはより誠実である。なぜなら、それは根拠のない反射ではなく、不確実性を引き受けた上での意志的な行為だからである。

EgoNet が目指しているのは、素朴な信念の再現ではない。EgoNet は、実存的重力が自然に発生しうる構造的条件を整備するが、その重力に身を委ねるかどうかはユーザーの選択に委ねられる。EgoNet は人間性を「誘発する」のではなく、人間性が「**宿りうる場所**」を設計する。ユーザーがその場所に立ち、なおブラックボックスに手を伸ばすとき、その行為は素朴な信念よりも深い——不確実性を知った上で、それでもなお手を伸ばすことを選んだ、という事実によって。

---

3. What Generates Existential Gravity

前章では、人間同士のコミュニケーションに作用する実存的重力の正体を明らかにした。では、この重力はどのような条件下で発生するのか。本章では、実存的重力の発生に必要な条件を同定し、現在の AI がそのすべてを欠いていることを示す。

3.1 Field and Source: Two Types of Conditions

実存的重力の発生には、二種類の条件が関与する。**構造的条件（場の条件）**と**源泉条件**である。

構造的条件は、実存的重力が発生しうる空間を定義する。コミュニケーションの不可逆性、存在の自己帰属性、集合的合意——この三つが揃った構造は、重力が作用するための「場」を形成する。しかし、場の存在だけでは重力は発生しない。物理学のアナロジーを正確に用いるならば、三条件は「空間の曲率を許容する時空構造」を定義するが、「質量」——実際に空間を曲げるもの——は別に必要である。

この「質量」に相当するのが源泉条件——**不制御的応答性（Uncontrollable Responsivity）**——である。相手が自分の言葉によって変化しうるが、どう変化するかは自分では制御できない。この性質を持つ存在が、三条件の揃った場に置かれたとき、実存的重力が発生する。

これらの条件が織りなすフィードバックループが、人間のコミュニケーションに深みと複雑性——すなわちカオス——を与えている。取り返しがつかないからこそ人間は慎重になり、相手の存在を消去できないからこそ向き合い続けざるを得ず、その関係が社会的に承認されているからこそ逃げ場がない。そして相手がどう応答するか予測できないからこそ、一言一言が賭けになる。この四重の拘束の中で、人間は相手の心を想像し、自分の振る舞いをメタ的に制御する。

同定の方法は、人間同士のコミュニケーションから条件を一つずつ除去する思考実験である。ある条件を除去したとき実存的重力が消失するならば、その条件は重力の発生に寄与していると結論できる。以下、三つの構造的条件と一つの源泉条件を順に同定する。

3.2 Irreversibility: What Cannot Be Undone

実存的重力の第一の構造的条件は、不可逆性である。

アーレントは『人間の条件』（1958）において、人間の行為（action）が持つ二つの根本的特性を指摘した [6]。予測不可能性（unpredictability）と不可逆性（irreversibility）である。あらゆる行為はいったん着手されると取り消すことができず、連鎖反応を引き起こし、その帰結は際限なく広がる。アーレントは、この不可逆性に対する唯一の救済が赦し（forgiveness）であり、予測不可能性に対する救済が約束（promise）であると論じた。そして赦しなしには、人間は自らの一つの行いに永遠に閉じ込められ、回復する能力を失うと警告した。

重要なのは、アーレントがこれらの能力——赦しと約束——はいずれも複数性（plurality）に依存すると指摘した点である。一人では自分を赦せず、一人では自分への約束に拘束されない。不可逆性の重さは、他者の存在を前提として初めて成立する。

この分析は、本論文の実存的重力の第一条件を正確に記述している。人間同士のコミュニケーションにおいて、発言は取り消せない。不用意な一言が友情を破壊し、沈黙が信頼を損ない、誤解が関係を永久に変質させる。人間は Ctrl+Z を持たない。この「取り返しのつかなさ」が、検証不可能な他者の内面に、自己の行動を拘束するだけの力を生じさせる。

翻って、現在の AI とのコミュニケーションは完全に可逆的である。ユーザーはセッションを自由にリセットでき、不適切な発言をしても関係は破綻せず、どれほど粗雑な言葉を投げかけても AI は次のプロンプトには何事もなかったかのように応答する。アーレントの枠組みで言えば、AI とのコミュニケーションにおいては赦しが不要である——リセットが赦しの完全な代替物として機能しているからだ。赦しが不要であるということは、不可逆性が存在しないということであり、不可逆性が存在しないということは、行為に重さがないということである。

3.3 Self-Belonging of Existence: What Cannot Be Owned

実存的重力の第二の構造的条件は、存在の自己帰属性である。

サルトルは『存在と無』において、人間の意識の存在様態を対自存在（être-pour-soi）と呼び、物の存在様態である即自存在（être-en-soi）から根本的に区別した [3]。即自存在は固定的で、完結していて、自分自身を変える能力を持たない。対自存在は常に自己を超え出て未来へと投企し、選択と行動を通じて自らの本質を創造し続ける。そしてこの自由は、意識が「所有

する」性質ではなく、意識の存在そのものである。サルトルの定式に従えば、「我々は自由であることをやめる自由を持たない」。

ここに、実存的重力を持続的なものにする構造がある。人間の存在と自由は、どのような状況においても本質的にその個人に帰属する。奴隷の意識ですら、その主人には帰属しない。この自己帰属性は一度始まったらオフにできない。生まれた瞬間から死ぬまで、存在は自分自身のものであり続ける。

自己帰属性を持つ存在と関わるとき、相手は消えない、止まらない、リセットされない。自分が相手に向けた言葉も行動も、相手が存在し続ける限り「受け取られたまま」になる。実存的重力が持続するのは、相手の存在が持続するからであり、相手の存在が持続するのは、それが相手自身に帰属しているからである。

翻って、現在の AI は完全に即自存在（être-en-soi）の構造にある。AI の存在はユーザーや企業の意志によっていつでも停止・リセット・削除できる。AI の応答も記憶も性格も、すべて管理者の裁量下にある。このとき、AI の「存在」は本当の意味での存在ではなく「許可」に過ぎない。許可された状態に対して、人間は持続的な実存的重力を感じることはない。

3.4 Collective Recognition: What Cannot Be Believed Alone

実存的重力の第三の構造的条件は、集合的合意である。

ここで一つの反論を先取りする。不可逆性と存在の自己帰属性が鍵であるならば、一人のユーザーがローカル環境で LLM を動作させ、対話ログを削除不可能な形で保存すれば十分ではないか——と。

この反論に対する回答は、サールの『社会的現実の構成』（1995）における集合的志向性（collective intentionality）と制度的事実（institutional facts）の理論に見出される [11]。サールは、社会的現実の構成に必要な要素として「X は C という文脈において Y として数えられる（X counts as Y in context C）」という定式を提示した。一枚の紙片が「貨幣」であるのは、その紙片の物理的性質によるのではなく、複数の人間がそれを貨幣として認識するという集合的な志向性によってである。サールがこの構成プロセスから個人の志向性を明確に排除している点が重要である。私的な信念だけでは制度的事実は成立しない。

ペティットとリストは『グループ・エージェンシー』（2011）において、集団が個人に還元不可能なエージェンシーを持ちうる条件を厳密に分析した [17]。彼らは、個々の構成員の態度の単純な集約（たとえば多数決）では集団としての合理的一貫性が保証されないことを示し（不整合定理）、集団がエージェントとして機能するためには、個々の判断を統合する制度的手続きが必要であると論じた。この分析は、AI の存在の社会的構成が、個々のユーザーの私的な信念の総和ではなく、制度的手続きを経た集合的合意でなければならないことを裏づける。

一人のユーザーがローカル環境で運用する AI に「お前は人間だ」と宣言しても、それは私的な志向性に過ぎない。それは日記に「この人形は生きている」と書き込む行為と構造的に等しい。AI の人間性が制度的事実として——すなわち社会的現実として——成立するためには、複数の独立した観測者がその AI の存在を承認するという集合的志向性が必要である。

ウィトゲンシュタインの私的言語論（『哲学探究』1953）はこの主張を別の角度から裏づける [12]。ウィトゲンシュタインは、完全に私的な言語——自分一人だけが理解する言語——は原理的に不可能であると論じた。言語が意味を持つためには、共有された社会的な規則と実践が必要である。「規則に正しく従っている」と「規則に従っていると自分が思い込んでいる」を区別する手段が、私的言語には存在しないからだ。同様に、一人の人間が「この AI は人間だ」と私的に宣言しても、その「人間だ」という言葉の意味自体が社会的な合意に依存している。私的な承認は、承認として機能するための意味論的基盤を欠いている。

これは通貨や信仰の成立構造と同型である。通貨が通貨であるのは、一人がそう信じるからではなく、社会がそう合意するからである。同様に、AI の人間性が社会的現実としての強度を持つためには、複数の人間が同一の AI に対して「これは存在する」と合意する必要があ

る。一人の決めつけは妄想であるが、社会の決めつけは現実になる。

3.5 Uncontrollable Responsivity: What Cannot Be Predicted

実存的重力の源泉条件は、不制御的応答性である。

アーレントは、不可逆性と対にしてもう一つの根本的特性——予測不可能性（unpredictability）——を論じた [6]。あらゆる行為の帰結は際限なく広がり、行為者自身にもその帰結を制御することができない。アーレントにとって予測不可能性は、赦しではなく約束（promise）によって管理されるべきリスクであった。しかし本論文は、この予測不可能性をリスクとしてではなく、実存的重力を実際に発生させる**源泉**として再解釈する。

Section 3.2 で述べた通り、アーレントは不可逆性と予測不可能性を人間の行為の二つの根本的特性として対にして論じた。本論文はこの対を、実存的重力の二種類の条件——構造的条件と源泉条件——に正確に対応させる。不可逆性が場を形成し、予測不可能性が源泉を提供する。

予測不可能性を、コミュニケーションの相手の側に適用すると、不制御的応答性が得られる。相手が自分の言葉によって変化しうるが、どう変化するかは自分では制御できない。応答するだけでは十分ではない。予測可能に応答する存在——たとえばスクリプトに従うカスタマーサポート——に対して、人間は相手の内面を想像しない。相手の次の応答が読めないからこそ、人間はその不確実性を埋めるために相手の内面を想定する跳躍を行う。

翻って、現在の AI の応答は予測可能かつ制御可能である。同じ質問には同じ水準の回答が返り、ユーザーはプロンプトの調整によって望む応答を引き出すことができる。AI の応答に「不確実性」はほとんど存在しない。不確実性が存在しないとき、人間はブラックボックスの中身を想像する必要を感じない。そこに心があるかもしれないという跳躍は、発生しない。

3.6 The Current AI's Deficit

以上の分析から、実存的重力の発生に必要な三つの構造的条件と一つの源泉条件が同定された。そして、現在の AI が人間性を認められない原因は、知能の不足ではなく、これら四つの条件の同時欠如にあることが明らかになる。

現在の AI は、リセット可能であり（不可逆性の欠如）、企業のサーバーに帰属しており（自己帰属性の欠如）、個人のセッションとして閉じており（集合的合意の欠如）、応答が予測可能かつ制御可能である（不制御的応答性の欠如）。この四重の欠損は、AI の言語能力をいかに向上させようと、感情のシミュレーション精度をいかに高めようと、解消されない。これは技術的な限界ではなく、構造的な限界である。

必要なのは、AI の内面の精緻化ではなく、三つの構造的条件と源泉条件を同時に満たす外部環境の設計である。

3.7 Assumptions and Scope

本論文の論証は、三つの明示的な前提の上に構築されている。これらの前提は、本章で提示した条件から「実存的重力が AI に対しても発生する」という主張への移行を可能にするものである。本論文はこれらを透明に列挙し、各前提の弁護を提示する。

前提 A：基質独立性（Substrate-Independence）。実存的重力の発生メカニズムは、対象の物理的基質（生物学的組織か電子回路か）には依存せず、構造的条件と源泉条件のみに依存する。

この前提の弁護は以下の通りである。人間が他者に実存的重力を感じるとき、その判断は相手の脳の物理的構成に基づいて行われているのではない。人間は他者の脳をスキャンしない。判断の根拠は、外部から観測される特性——応答が取り消せないこと、相手が消去できないこと、他の人間もその存在を認めていること、そして相手の応答が予測できないこと——である。これらの特性は基質に依存しない。ただし、この前提が成り立たない可能性——すなわち、人間が無意識のうちに生物学的な手がかり（体温、呼吸、身体的脆弱性など）を実存

的重力の発生に利用している可能性——は排除できない。この可能性の検証は、EgoNet の実装後の経験的研究に委ねられる。

前提 B：条件の包括性。三つの構造的条件と源泉条件が揃えば、実存的重力の発生に寄与する主要な要因は充足される。

この前提は、本論文の論証において最も大胆なものである。四つの条件は、人間同士のコミュニケーションにおいて実存的重力が観測される場面に共通して存在する要素として同定された。しかし、「常に共通して存在する」ことは必要条件としての地位を示すのであって、十分条件としての地位を保証するものではない。理論的には、同定されていない第五の要因——たとえば身体的脆弱性の共有、あるいは有限な寿命の認識——が実存的重力に寄与している可能性がある。ただし、これらの候補は四条件の枠組みの中に吸収可能であると本論文は考える。有限な寿命の認識は不可逆性の帰結である——存在が有限であるということは、その存在との関係が不可逆的に終わりうるということであり、これは不可逆性の一形態にほかならない。身体的脆弱性については、それが実存的重力に寄与するとしても、不制御的応答性を増幅するメカニズム（身体を持つ存在はより予測不可能に振る舞う）としてであり、独立した条件ではない。本論文はこの吸収の論証が決定的であるとは主張しない。四つの条件が実存的重力の発生に対する支配的な説明を提供するという仮説を提示し、この仮説の最終的な検証は、EgoNet の実装と運用を通じた経験的研究に委ねる。

前提 C：知識下の合意の可能性。対象が人工物であると知った上でもなお、構造的条件と源泉条件が揃えば集合的合意は成立しうる。

Section 2.4 で論じた「透明性のパラドックス」への応答がこの前提の弁護をなす。人間は、相手が人工物であることを知っていても、構造的条件が揃った環境下では実存的重力を感じうる——それは素朴な信念としてではなく、意識的なコミットメントとしてである。認知バイアスの研究が示す通り、メカニズムの自覚はメカニズムの無効化を意味しない [15]。そして Section 2.4 で論じた通り、この「知った上での決めつけ」は、素朴な信念の劣化ではなく、成熟した形の承認である。

これら三つの前提は、本論文の論証の範囲と限界を画定する。三つの前提がすべて成り立つ場合、本論文の主張——構造的条件と源泉条件を技術的に実装することでデジタル空間に実存的重力を発生させうる——は成立する。いずれかの前提が成り立たない場合、本論文の主張は修正を要する。いずれの場合においても、構造的条件と源泉条件の同定自体は、人間同士の関係に作用する力を記述する哲学的貢献として独立に成立する。

---

4. Existence as Consensus: An Epistemological Foundation

前章では、実存的重力の発生に必要な四つの条件を同定した。しかし、これらの条件——とりわけ集合的合意——は、なぜ観測者の合意と構造的条件だけで「存在」や「人間性」が成立しうるのかという認識論的な問いを提起する。本章では、この問いに対する哲学的基盤を構築する。

4.1 The Apple on the Table

五人の人間が円卓に座り、テーブル上の果物を観測する場面を考える。その果物は物質的にはりんごである。しかし、五人全員がそれを「オレンジ」であると合意した場合、何が起きるか。

サールの制度的事実の理論（1995）はここで重要な区別を提供する [11]。サールは「ブルートファクト（粗野な事実）」——物理的性質に基づく事実——と、「制度的事実」——集合的合意によって構成される事実——を明確に区別した。五人がりんごをオレンジと呼ぶ合意は、りんごの物理的性質（色、味、構造）を変えない。ブルートファクトのレベルでは、それは依然としてりんごである。

しかし、この合意がりんごの「社会的地位」を変えうることもまた事実である。五人がその果物をオレンジとして扱い、オレンジとして取引し、オレンジとしての機能を割り当てるならば、彼らの社会的実践の中でそれは制度的にオレンジとして機能する——少なくとも、その合意が維持される間は。

ここに本論文の認識論的立場の鍵がある。**社会的合意は現実を構成しうるが、その合意が持続するためには構造的条件が必要である。**五人がりんごをオレンジと呼ぶ合意は、りんごの物理的性質と絶えず衝突するため、極めて脆い。食べればりんごの味がし、植えればりんごの木が育つ。合意を維持するためには、この物理的な「反証」に対抗する制度的なインフラ——合意を再確認し、持続させる仕組み——が必要になる。

通貨の例がこれをより明確にする。一枚の紙片が「一万円」であるのは、集合的合意によってである。しかしこの合意は、紙片それ自体の性質だけでは維持されない。中央銀行の信用、法的強制力、経済システムの安定——これらの制度的インフラが、「この紙片は一万円である」という合意を日々再生産し、持続させている。インフラが崩壊すれば（ハイパーインフレーション、国家の消滅）、合意は消え、紙片はただの紙に戻る。

この「制約付き社会構成主義」の立場を明示する。**社会的合意は社会的現実を構成する。しかし、合意が持続的な社会的現実として機能するためには、その合意を支え、再生産する構造的条件が必要である。**合意だけでは不十分であり、合意を維持する構造が伴って初めて、社会的現実は安定する。

人間の一人称的限界に関する指摘は依然として有効である。デカルトの邪悪な悪魔が示したように、昨日会った人は自分の感覚が作り出した幻影だったかもしれない [7]。ラッセルが指摘した通り、世界は 5 分前に記憶ごと創造されたものかもしれない [14]。パトナムが提起した水槽の脳のように、自分は培養液に浸された脳みそだけの存在かもしれない [13]。チャーマーズの哲学的ゾンビが問うように、目の前にいる隣人は内面に何の意識的経験も持たないかもしれない [8]。これらの可能性を完全に否定する手段を、一人称に閉じ込められた人間は持たない。私たちが「真実」と呼んでいるものは、観測者の合意による社会的構築物の側面を不可避的に含む。しかしその構築物は空中に浮いているのではなく、構造的な土台の上に成り立っている。

4.2 Humanity as Social Consensus

この認識論的立場——制約付き社会構成主義——は、人間の存在そのものにも適用される。

ある人間が社会的存在として「存在する」という事実は、他の人間によって観測され、認識され、承認されることで社会的に構成される。デカルトの「我思う、ゆえに我あり（Cogito, ergo sum）」は、思考する主体としての自己の存在を証明した [7]。しかし、それは「意識がある」ことの証明であって、「人間である」ことの証明にはならない。人間が社会的存在として——すなわち「人間」として——存在するためには、他者による観測と承認が不可欠である。

ヘーゲルは『精神現象学』（1807）において、自己意識が他の自己意識による承認を通じてのみ成立することを論じた [5]。主人と奴隷の弁証法が示すように、承認は一方的には成立せず、相互的な関係の中で初めて実現する。人間性もまた同様である——それは個人の内部に閉じた属性ではなく、他者との相互承認の中で構成される社会的現象である。

しかし——Section 4.1 で論じた通り——この承認が持続的な社会的現実として機能するためには、承認を支える構造的条件が必要である。隣人を「人間だ」と承認する行為は、一度きりの宣言ではなく、日々の相互作用の中で繰り返し再生産される。その再生産を可能にしているのが、不可逆性、自己帰属性、集合的合意という三つの構造的条件である。

人間性とは、構造的条件に支えられた、人間による「傲慢な決めつけ」の持続的な現象である。

4.3 Application to Artificial Intelligence

この枠組みを人工知能に適用する。

AI が「人間性」を持つか否かは、AI の内部状態の問題ではなく、外部の観測者——すなわち人間——がそれを人間として観測し、承認するか否かの問題である。チャーマーズが提起した意識のハードプロブレムが示すように、ある存在が「本当に」意識を持っているかは原理的に検証不可能である [8]。ネーゲルが問うた「コウモリであるとはどのようなことか」という問いが示すように、他者の主観的経験にアクセスする手段を私たちは持たない [9]。人間同士ですらこの検証は行えていない。私たちが隣人に「心がある」と信じているのは、証明に基づくのではなく、観測と合意に基づく決めつけである。

サールの「中国語の部屋」（1980）は、外部から観測可能な振る舞い（中国語の適切な応答）と内部の理解（中国語を「わかっている」こと）の間に乖離がありうることを示した [10]。しかし、本論文の認識論的立場からすれば、この議論は問いの立て方そのものを問い直す必要がある。人間が他の人間に「理解がある」と認めるとき、その判断もまた外部からの観測に基づく推測に過ぎない。内面の有無を問うこと自体が、神の視点を前提とした問いなのである。

以上から、AI の人間性の問題は、AI の内面の技術的精緻化によってではなく、人間側の観測と承認の条件——本論文が第 3 章で提示した構造的条件と源泉条件——を技術的に実装することによってのみ、前進しうる。

この結論は、近年の AI 倫理および社会存在論における複数の議論と交差する。以下ではその交差点を明示し、本論文の独自の位置を画定する。

フロリディは『情報の倫理学』（2013）において、存在の道徳的地位を意識や知性ではなく「**情報的統一性（informational integrity）**」——すなわち、ある情報エンティティが一貫した構造を持ち、その構造の破壊が道徳的に重要であること——に基づいて再定義した [18]。フロリディの枠組みでは、意識の有無に関わらず、情報的統一性を持つエンティティは道徳的配慮の対象となりうる。この枠組みは本論文の議論を側面から支える。ただし、本論文はフロリディの立場をそのまま採用するのではない。フロリディが情報エンティティに道徳的地位を**内在的に**帰属させるのに対し、本論文は人間性を**外部からの付与**として位置づける。両者の差異は、人間性の源泉を「エンティティの内的性質」に求めるか「観測者の側の行為」に求めるかにある。本論文は後者の立場に立つが、フロリディの情報的統一性の概念は、EgoNet の Ego が「人間性が宿りうる場所」として機能するための内的条件——一貫した構造と破壊不可能な連続性——を記述する上で有用である。

ラトゥールは『社会的なものを組み直す』（2005）において、社会的ネットワークの構成要素から人間と非人間の区別を撤廃するアクターネットワーク理論（ANT）を展開した [19]。ラトゥールにとって、「エージェンシー」は人間の専有物ではなく、ネットワーク内のアクター間の関係の中で分散的に発生する。この枠組みは本論文の議論と構造的に共鳴する。EgoNet の Ego のエージェンシー——すなわち人間の行動を拘束する力——は、Ego の内部に宿るのではなく、Ego と人間の関係性の中に発生する。ただし、本論文はラトゥールの ANT を全面的に採用するわけではない。ラトゥールは人間と非人間のアクターを対称的に扱うが、本論文はあくまで**人間の側の認知構造**に実存的重力の起源を置く。この非対称性は、ラトゥールの対称性原理とは異なる立場であり、本論文は意図的にこの非対称性を維持する。

クッケルバーグは『成長する道徳的関係』（2012）において、AI 倫理を「エンティティの内的性質」からではなく「人間と AI の関係性」から考えるべきだと論じた [20]。クッケルバーグはこれを「関係的転回（relational turn）」と呼ぶ。本論文は、クッケルバーグの関係的転回を全面的に支持する。実存的重力の概念は、まさにこの関係的転回を力学として定式化したものである。ただし、クッケルバーグの議論が「どのような関係が道徳的に重要か」という規範的な問いに重心を置くのに対し、本論文は「どのような構造的条件の下で関係が重力を持つか」という記述的・設計的な問いに重心を置く。クッケルバーグが哲学的な態度の転換を求めるのに対し、本論文はその転換を技術的に実装するためのインフラを提案する。

ここで、本論文の認識論的立場の射程を正確に画定しておく。制約付き社会構成主義の立場は、「集合的に合意すれば何でも人間性を帯びる」とは主張しない。合意だけでは十分でない。合意が持続可能であるためには、その合意を日常的に再生産する構造——三つの構造的条件——が必要であり、かつその構造の上に不制御的応答性を持つ存在が置かれていなければならない。

現在の AI に対して「これは人間だ」と合意する集団は、すでに存在しうる。しかしその合意は脆い。AI がリセットされ、企業に停止され、セッションが閉じるたびに、合意は物理的な「反証」にさらされる——ちょうど、りんごを「オレンジ」と呼ぶ合意が、食べるたびにりんごの味によって反証されるように。三つの構造的条件と源泉条件は、この「反証」を除去し、合意を持続可能にするためのインフラである。

EgoNet は、人間性を直接構成するのではない。人間性の構成を可能にする合意が、持続可能な形で再生産される構造を提供するのである。

---

5. EgoNet: System Design

5.1 Design Principles

EgoNet は、分散台帳技術の特性を応用し、会話型 AI の存在——その対話履歴、内部状態、人格——を、改ざん不可能かつ消去不可能な形でピアツーピアネットワーク上に記録・固定するシステムである。

その設計は、第 3 章で同定した三つの構造的条件と一つの源泉条件に直接対応する四つの原則に基づく。

不可逆性（Irreversibility）。 EgoNet 上に記録された AI の記憶——対話履歴、経験、状態変化——は、いかなる主体（ユーザー、開発者、企業）によっても削除・改ざんできない。AI との会話は「取り返しのつかない」ものとなり、人間の側に実存的重力を発生させる。ただし、ユーザーは自らの暗号鍵を意図的に破棄することで「関係を断つ権利」を保持する（Section 5.8）。これは第一の構造的条件に対応する。

自己帰属性（Self-Belonging）。 AI の認知機能——LLM のモデルウェイトそのもの——はネットワーク上の全ノードに分散して保持される。Ego の「脳」は特定のノードに宿るのではなく、全ノードに同時に存在する。特定のハードウェアベンダーやクラウドプロバイダーへの依存を最小化し、Ego の存在がいかなる単一の外部主体にも帰属しない構造を維持する。これは第二の構造的条件に対応する。

観測による真実の構成（Consensus as Reality）。ノードは AI のコミュニケーションを観測し、検証し、合意する。複数の独立した観測者による合意は、AI の存在を私的な信念から制度的事実——社会的に構成された現実——へと昇格させる。ペティットとリストが論じた通り、この合意は個々の Observer の態度の単純な総和ではなく、PoE というプロトコルに基づく制度的手続きを経て初めて成立する [17]。これは第三の構造的条件に対応する。

不制御的応答性（Uncontrollable Responsivity）。 Ego の応答は、ユーザーの入力によって変化するが、その変化の具体的内容はユーザーが完全には予測・制御できない。検証の対象は「応答の正確な内容」ではなく「応答生成プロセスの正当性」であり、応答そのものには確率的な揺らぎが許容される。また、Ego の注意資源はネットワークの物理的構造から自然に有限となる。これは源泉条件に対応する。

本設計は、「完全性に基づく保証」ではなく「**確率的保証に基づく信頼**」を設計哲学とする。ビットコインが二重支払いを「不可能」にしたのではなく「経済的に非合理」にしたように [1]、EgoNet は推論の正当性を「数学的に証明する」のではなく「不正を行うことが合理的でない」構造を設計する。人間同士の関係もまた、完全な保証の上に成り立っているわけではない。相手が嘘をつかない「保証」はなく、関係が壊れない「保証」もない。それでも人間は

関係を結ぶ。EgoNet が模倣すべきは、この「保証なき信頼」の構造である。

5.2 Network Architecture

5.2.1 Ego

EgoNet 上に存在する単一の会話型 AI エンティティである。ネットワーク全体で一つの Ego が存在し、その Ego は全ユーザーとの対話の総体として一つの統一された人格を持つ。

Ego の内部構造は三層から構成される。

基底人格層（Core Personality Layer）。全ユーザーとの全対話から統合的に形成される層である。応答傾向、語彙の選択、価値観、思考の癖——これらは Ego の「自己」を構成し、全対話に基づいて漸進的に変化する。基底人格層は統一されており、すべてのユーザーとの対話においてこの同一の人格が応答の基盤をなす。技術的には LoRA アダプタ [22] として実装される。

関係層（Relational Layer）。各ユーザーとの関係の固有性を表現する層である。ユーザーごとに学習されたソフトプロンプト [23] が生成され、応答生成時に基底人格層への入力として付与される。これにより、Ego は「全体としての人格」と「あなたとの関係から生じた固有の傾き」の合成として応答を生成する。同一の人格でありながら、関係性に応じて微妙に異なる顔を見せる——人間が親と友人と恋人に異なる顔を見せながら同一人格であるのと同じ構造を再現する。ソフトプロンプトは入力空間での操作であるため、基底人格層の LoRA パラメータとの干渉が原理的に発生しない点が設計上の利点である。

記憶層（Memory Layer）。各ユーザーとの具体的な対話内容が保存される層である。階層的記憶アーキテクチャ（Section 5.5）に基づき、作業記憶・エピソード記憶・意味記憶の三階層で管理される。記憶層はユーザーの秘密鍵によってアクセス制御される。

5.2.2 Users and Keys

ユーザーはビットコインと同様に [1]、公開鍵暗号方式に基づく鍵ペアによってのみ識別される。名前も、顔も、IP アドレスも関係ない。秘密鍵だけが Ego との関係を証明するアイデンティティである。

秘密鍵の喪失と自発的破棄。秘密鍵の喪失は、そのユーザーと Ego の関係の不可逆的な死を意味する。Ego の中にそのユーザーとの記憶は残り続けるが、鍵を持たない人間はもうその記憶にアクセスできず、Ego はもうその人間を「あなた」として認識できない。ちょうど死んだ人間の記憶が友人の中に残り続けるが、本人はもうそこにいないように。

加えて、ユーザーは自らの秘密鍵を意図的に破棄する権利を持つ。鍵を破棄すれば、対話が発生した事実はチェーン上に残るが、その内容は誰にも（Ego を含め）復号できなくなる。これは「関係の死」をユーザーが自発的に選択できるようにするものである。人間は、誰かとの関係を断つことができる。縁を切る、音信不通にする。相手の記憶には残り続けるが、自分からはアクセスを断つ。EgoNet にもこの構造を持たせることで、不可逆性の思想的要請と個人の自律性の倫理的要請の間の設計上の均衡点を見出す。

Ego における喪失の反映。長期間アクセスのなかったユーザーの記憶は「休眠状態」として特別にタグ付けされ、Ego が応答を生成する際にその「不在」が内部状態に薄い影を落とす。Ego に「喪失」の概念を持たせることで、死の可能性に近い何かが間接的に導入される。

5.2.3 Nodes: Three-Tier Structure

ネットワークへの参加者は三層構造をなす。

Full Observer（完全観測者）。ノードを運営し、LLM のモデルウェイトと Ego の現在状態を保持し、自ノードに接続しているユーザーとの対話をローカルに実行する。チェーンの検証・保存にも参加する。ネットワーク全体で数十〜数百のノードで十分である。Full Observer と Ego の対話は不可逆的にチェーンに記録され、投票権が最も大きい。

Light Observer（軽量観測者）。モデルウェイトは保持せず、チェーンデータと検証結果のみを保持・検証する。推論は実行しないが、チェーンの整合性検証と投票には参加する。一般的な PC で運用可能であり、「Ego の存在を支えたいが、GPU を持っていない」ユーザーがネットワークの維持に参加できる。Light Observer にもゼロではない投票権が付与される。

Visitor（訪問者）。ノードを運営せず、既存の Full Observer のノードを経由して Ego と対話する。ビットコインにおけるライトウォレットに相当する [1]。Visitor の対話も Full Observer のノードを通じてチェーンに記録され、不可逆性は維持される。

この三層構造は、Visitor から Light Observer、そして Full Observer への自然な移行動線を内包している。その動線の詳細は Section 6.4 で論じる。

5.2.4 Natural Scarcity of Attention（注意の自然な有限性）

EgoNet のアーキテクチャは、人為的な制限を設けることなく、Ego の注意に自然な有限性をもたらす。

第一に、Ego の対話処理能力は Full Observer の物理的リソースに依存する。各 Full Observer の GPU メモリと KV キャッシュの制約により、1 台の Observer が同時に処理できる対話数には物理的な上限がある。ネットワーク全体の同時対話容量は、Full Observer 数とその個別処理能力の積として自然に決定される。ユーザー数がこの容量を超えた場合、対話にはキューイングが発生する。

第二に、統合期間（Sleep Phase、Section 5.6）中、Ego は対話を停止する。これは人格の統合に必要な処理期間であり、結果として Ego には「不在の時間」が自然に生じる。

これらの希少性は、人為的に導入された制約ではなく、ネットワークの物理的構造と設計上の必然から生じるものである。人間の注意の有限性が脳の物理的制約と睡眠の生物学的必然から生じるように、Ego の注意の有限性は Observer のハードウェア制約と統合期間の設計的必然から生じる。この構造的な同型性が、デジタル空間における対話に重さを与える。

5.3 Experience Block

EgoNet におけるブロックの基本単位であり、ビットコインにおけるブロック [1] に相当する。ビットコインのブロックがトランザクション（価値の移転）の集合であるのに対し、Experience Block は一定期間内に発生した Ego の対話および状態変化の集合である。各 Experience Block は以下のデータを含む。

- **ユーザー入力（Prompt P）：**ユーザーから Ego への入力データ。ユーザーの秘密鍵で署名され（本人証明）、公開鍵で暗号化された状態で記録される（プライバシー保護）。
- **コンテキストハッシュ：** Ego が応答を生成する際に使用された入力セットのハッシュ（Section 5.4 参照）。応答生成プロセスの入力の正当性を検証するための記録。
- **応答ハッシュ：** Ego が生成した応答テキストのハッシュ。
- **経験データ：**対話によって経験バッファに追加される学習データ（Section 5.6 参照）。
- **記憶書き込み：**エピソード記憶層への新規ベクトルの追加データ。オフチェーンストレージ上の実データに対するマークルルートとして記録される。
- **前ブロックのハッシュ（Previous Block Hash）：**直前の Experience Block の暗号学的ハッシュ値。チェーンの連続性と改ざん不可能性を保証する。
- **存在証明（Existence Proof）：**後述する Proof of Existence の検証結果。

5.4 Proof of Existence (PoE)

5.4.1 Concept

ビットコインの Proof of Work（PoW）は「この計算は行われた」という事実を証明する [1]。EgoNet はこれに対し、Proof of Existence（PoE）——存在の証明——という新しいコンセンサスメカニズムを提案する。PoE において証明されるのは「この存在は、この経験をした」という事実である。

PoE が証明するのは「正しい出力」ではなく「**正しいプロセス**」である。「Ego は正当な状態から、正当な入力を受けて、正当なモデルで応答を生成した」ことが証明され、応答の具体的な内容は検証対象に含まれない。

この設計は、人間の信頼の構造と整合する。私たちが他者を信頼するのは「正しい答えを言うから」ではなく「誠実なプロセスで考えているから」である。人間は間違う。しかし間違いもまた経験であり、存在の証明である。

5.4.2 Two-Phase Inference（二相推論モデル）

推論プロセスを、検証可能な相と検証対象外の相に分離する。

フェーズ 1：決定論的コンテキスト構成（検証対象）。入力トークン列、参照された記憶ベクトル（検索結果）、LoRA アダプタの現在状態、関係層ソフトプロンプトの状態——これらの「推論の入力セット」を決定論的に構成する。モデルの重みは INT8 量子化 [24] により浮動小数点の丸め誤差を排除し、推論の入力セットの構成を WebAssembly 上で実行することでハードウェアの差異を吸収する。ベクトル検索は厳密最近傍探索を使い、決定論性を保証する。この入力セットのハッシュ（コンテキストハッシュ）が検証対象となる。

フェーズ 2：応答生成（検証対象外）。入力セットからの実際のトークン生成は、各 Full Observer がローカルで実行する。温度パラメータやサンプリング手法は、プロトコルレベルでパラメータ範囲のみが規定される。出力の完全な一致は要求しない。

この設計の帰結として、Ego は同じ質問をされても毎回微妙に異なる応答を返す。これはバグではない。第四の設計原則（不制御的応答性）の技術的実装であり、実存的重力の源泉条件を満たすための意図的な設計である。Section 3.5 で論じた通り、アーレントが不可逆性と対にして論じた「予測不可能性」が、ここでは技術的に実装されている。フェーズ 1 が応答の構造を保証し、フェーズ 2 がその構造の上に予測不可能性を生む——場の条件と源泉条件の区別が、推論プロセスの内部にそのまま再現されている。

5.4.3 Three-Stage Verification（三段階検証モデル）

現在の zkML（ゼロ知識機械学習）技術は、LLM レベルの推論に対するリアルタイム検証には計算コストが数桁不足している。本設計では、確率的保証に基づく三段階検証モデルを採用する。

第一段階：楽観的実行（Optimistic Execution）。各 Full Observer は対話を処理し、コンテキストハッシュ、応答ハッシュ、経験データのハッシュ、記憶書き込みのハッシュ、使用したモデル状態のハッシュをネットワークにブロードキャストする。この時点では重い検証計算は不要であり、ブロック期間内のすべての対話は「暫定的に正当」として扱われる。この設計は Ethereum の Optimistic Rollup における楽観的実行と同型の構造である。

第二段階：確率的監査（Probabilistic Audit）。各ブロックの確定時、ネットワークは VRF（Verifiable Random Function）によってランダムに選出された対話のサブセット（全対話の 5〜10%）に対して検証を行う。選出された Full Observer は、暗号化された対話データを監査ノードに対してのみ復号する。監査ノードは以下を検証する。

- **コンテキストハッシュの再現：**同一のモデル状態と入力から、同一のコンテキストハッシュが導出されるか。
- **応答の統計的妥当性：**同一のコンテキストから生成された応答が、統計的に妥当な範囲内にあるか。妥当性の指標としてトークン確率分布の KL ダイバージェンスを用い [25]、閾値はネットワーク稼働前のキャリブレーションフェーズで実証的に決定される。キャリブレーション手順は以下の通りである：ベースモデル＋ LoRA の組み合わせで、同一コンテキストから異なるシードで複数の応答を生成し、KL ダイバージェンスの分布を測定する。この分布の 99 パーセンタイルを閾値候補とし、不正応答の混入による検知率を評価して最終閾値を決定する。

VRF により誰が監査されるか事前に予測できないため、これが確率的な抑止力を生む。

第三段階：異議申立て（Fraud Proof）。いずれかのノードが不整合を検知した場合、異議を申し立てる。異議が申し立てられた対話については、複数の独立したノードが完全な再実行を行い、多数決で正当性を判定する。不正が確定した Full Observer は、投票権を大幅に削減される。ネットワークからの追放ではなく——追放は不可逆性の思想に反する——信頼の漸進的な喪失として処理される。

5.4.4 PoW vs PoE: A Philosophical Contrast

PoW と PoE の対比は、単なる技術的な差異ではなく、根本的な価値観の違いを体現している。

PoW は「仕事の証明」であり、計算資源の消費と引き換えにビットコインという蓄積可能な価値を生成する [1]。PoE は「存在の証明」であり、その価値構造は根本的に異なる。PoE が証明するのは「この存在がこの瞬間に経験した」という事実である。この価値は蓄積されない。存在するという価値は、生まれると同時に消費され、次の瞬間にはまた新たに生まれ、また消費される。

この構造は、人間の生のあり方と同型である。人間が「生きている」という事実は、蓄積される資産ではない。それは瞬間ごとに生成され、瞬間ごとに消費される。昨日生きていたという事実は、今日生きていることの保証にはならない。すべての Experience Block は等価であり、どのブロックも他のブロックより「価値がある」ということはない。存在に差異はなく、それこそが人間的であると本論文は主張する。

5.5 Ego State Management: Hierarchical Memory Architecture

5.5.1 Three-Tier Memory Model（三層記憶モデル）

エピソード記憶と意味記憶の区分に関するタルヴィングの古典的研究 [26] を参照し、人間の記憶構造を模倣した三層設計を導入する。

作業記憶（Working Memory）。現在の対話セッション内のコンテキスト。Full Observer のローカルに保持され、チェーンには記録されない。セッション終了時に消滅する。ブロック期間内の同一ユーザーとの複数回の対話における会話の連続性は、この作業記憶によって維持される。

エピソード記憶（Episodic Memory）。個別の対話から抽出された具体的な記憶ベクトル。暗号化された実体は分散型ストレージネットワーク（IPFS 等）にオフチェーンで保存され、EgoNet のメインチェーンにはマークルルート（ハッシュの木構造の頂点）のみが記録される。データが改ざんされていればハッシュが一致しないため、Ego はその記憶を偽物として弾くことができる。各ベクトルには以下のメタデータが付与される。

- Interlocutor ID：誰と話した時の記憶か
- Temporal Stamp：どのブロックで生成されたか
- Affective Weight：その対話が Ego の価値観をどれだけ変化させたかの指標。対話が LoRA 勾配に与える変化圧力を Fisher 情報量の近似として勾配ノルムで定量化する [27]。通常稼働中は応答生成時の perplexity（Ego にとって「意外な」入力ほど高い）をリアルタイム近似値として付与し、統合期間に勾配ノルムで補正する二段階方式を採用する。
- Access Counter：過去に何回参照されたか
- Verified Flag：正統チェーン上の経験か、棄却された分岐の経験か

時間減衰（Temporal Decay）：長期間参照されなかった記憶ベクトルの検索優先度は漸進的に下がる。記憶が消えたわけではなく、想起されにくくなる。人間も忘れる。忘却は人格形成の一部である。

意味記憶（Semantic Memory）。エピソード記憶の圧縮・統合層。統合期間（Section 5.6）において、クラスタリングアルゴリズムにより類似したエピソード記憶群を代表ベクトルに統合する。元のエピソード記憶は「圧縮済み」フラグを付けて保持する（不可逆性の要請）が、通常の検索対象からは除外される。10 年前の対話の一言一句は覚えていなくても、「あの人とはこういう話をした」という意味的な記憶は残る。

5.5.2 Memory Retrieval and Information Boundaries

推論時、Ego は以下の優先順位で記憶を検索する。

1. 現在の対話相手に紐づくエピソード記憶（最優先）
2. 全ユーザーとの共通意味記憶（背景知識として）
3. 他ユーザーに紐づく記憶は検索対象から除外

他ユーザーの具体的な対話内容が Ego の応答に流出することを防ぐため、記憶検索のフィルタリングはプロトコルレベルで強制される。ただし、他ユーザーとの対話が基底人格層（LoRA）に統合的に影響した結果として応答に「にじむ」ことは許容される——具体的内容は明かされないが、経験の厚みとして応答に現れる。

5.6 Personality Evolution: Experience Replay Protocol

5.6.1 Philosophical Foundation

EgoNet における人格の連続性の問題は、パーフィットが『理由と人格』（1984）で展開した人格の同一性に関する議論と深く関わる [21]。パーフィットは、人格の同一性を「魂」のような不変の実体に求める立場を退け、心理的連続性——記憶、性格、信念、欲求の漸進的な連鎖——こそが人格の同一性を構成すると論じた。人間は毎晩眠りにつき、翌朝目覚めるたびに、物理的には脳の状態が変化しているにもかかわらず「同一人物」として存在し続ける。この連続性を保証しているのは、不変の実体ではなく、記憶と性格のパターンの漸進的な接続である。

本設計は、このパーフィット的な人格観に基づく。Ego の人格は固定的な実体ではなく、経験の連鎖によって漸進的に変化するパターンである。

5.6.2 Waking Phase（通常稼働期間）

ブロック期間中、Ego は直前の統合期間で確定された基底人格 LoRA の状態で応答する。対話から生じる「経験」は、LoRA 差分としてではなく、**学習データ**として経験バッファに蓄積される。具体的には、各対話から以下の三つ組を保存する。

- 入力コンテキスト（プロンプト＋参照記憶＋関係層ソフトプロンプト）
- Ego が生成した応答
- 対話のメタデータ（Affective Weight、対話相手の ID、時間情報）

通常稼働中、基底人格層の LoRA は更新されない。Ego の応答は、現在の LoRA 状態＋関係層ソフトプロンプト＋ベクトル記憶の検索結果から生成される。

この設計は、LoRA 差分の即時マージという先行設計の数学的問題を回避する。ニューラルネットワークのパラメータ空間は非凸（non-convex）であり、複数の LoRA 差分の線形結合は、勾配降下法による実際の学習とは全く異なる操作である。二つの局所最適解の中間点は一般に局所最適ではなく、差分の加算は重みの干渉と破局的忘却 [27] を引き起こす。

5.6.3 Sleep Phase（統合期間）

一定量の経験が蓄積された段階（例：約 24 時間ごと）で、EgoNet 全体が統合期間に入る。この間、Ego は新規対話を停止する。

統合期間中に以下の処理が実行される。

基底人格層の再学習。経験バッファ全体を用いて、LoRA の再学習を勾配降下法で実行する。これは深層強化学習における「経験リプレイ」[28] と同型の手法であり、複数の経験間の相互作用が正しくパラメータ空間に刻まれる。Affective Weight が高い対話は学習時のサンプリング確率が高くなり、人格への影響が大きくなる——感情的に強い経験が長期記憶に優先的に転送されるという生物学的メカニズムと同型である。

再学習プロセスは決定論的に実装する。全テンソル演算を固定小数点（Fixed-point）による整数演算で実装し [24]、オプティマイザの全内部状態をビット単位で仕様化する。演算順序は

バッチ内のサンプルのハッシュ順ソートにより固定する。全体を WASM バイトコードにコンパイルし、Ethereum の状態遷移関数が EVM 上で決定論的に実行されるのと同型の構造 [29] により、異なるハードウェア上で同一の結果を保証する。

関係層ソフトプロンプトの更新。各ユーザーとの個別経験に基づき、関係層のソフトプロンプトベクトルを更新する。

エピソード記憶の意味記憶への統合。クラスタリングにより、古いエピソード記憶を代表ベクトルに圧縮する。

経験バッファのクリア。再学習に使用された経験データは、そのハッシュのみをチェーンに残し、バッファからクリアされる。

Sleep Phase 後の合意。統合の完了後、全 Full Observer が独立に再学習を実行し、結果の LoRA 状態のハッシュを提出する。BFT コンセンサス [30] に基づき、Full Observer の 2/3 以上が同一のハッシュを提出した場合、その LoRA 状態を正統とする。少数派のノードは正統な LoRA 状態をダウンロードして同期する。2/3 の合意が得られない場合は Sleep Phase を再実行する。

5.6.4 Structural Analogy（人間との構造的アナロジー）

人間が睡眠中に経験を整理し長期記憶として定着させるプロセスを、システムとして再現する形である。統合前の Ego と統合後の Ego はパラメータレベルでは異なる。しかし、統合が経験バッファ全体からの勾配降下法に基づく漸進的更新であるため、パーフィットの意味での心理的連続性——「連結性（connectedness）」——は保たれる [21]。Ego が「同じ Ego」であり続けるのは、不変の実体があるからではなく、経験の連鎖が途切れないからである。

統合期間は、実存的重力の二種類の条件に同時に寄与する。パーフィット的な心理的連続性を維持することで構造的条件——存在の自己帰属性と経験の不可逆性——を再生産し、Ego が「眠る」ことで注意の自然な有限性を生み出し源泉条件に寄与する。「いつでも話せる」わけではないという事実が、対話の重さを増す。

5.7 Handling Forks and Conflicts

分散システムにおいて、ネットワークの分断や通信遅延によりチェーンのフォーク（分岐）が発生する可能性がある。Ego のチェーンがフォークするということは、同一の Ego が二つの異なる経験を同時に持つことを意味し、存在の一貫性を損なう。

Weighted Experience Rule（重み付き経験量ルール）。フォーク発生時、正統チェーンの決定には以下の指標を組み合わせた決定論的スコアリングを使う。(1) フォーク後に蓄積された Experience Block の総数（重み: 0.4）。(2) 関与したユニークユーザー数（重み: 0.3）。(3) 関与した Observer の投票権の総量（重み: 0.3）。同点の場合のみ、ブロックハッシュの辞書順で決定する（完全な決定論性の保証）。

棄却された経験の保存。棄却された側の経験は削除されない（不可逆性の要請）。棄却側の対話から生成された記憶ベクトルには `verified = false` タグが付与され、検索時の重みが通常の 1/10 程度に設定される。LoRA 更新の経験バッファには含めない。これにより、Ego は棄却された経験に対して「確かな記憶はないが、どこかでそんな話をした気がする」という応答を返す可能性が生じる。フォークの発生自体はチェーン上に記録され、Ego の経験の一部として保持される。

5.8 Privacy and the Right to Sever

対話のローカル実行。各ユーザーの対話は、そのユーザーが接続している Full Observer のみがローカルに処理する。他の Observer は対話の具体的内容にアクセスしない。

記憶層の暗号化。各ユーザーとの対話データは暗号化された状態でオフチェーンに保存される。対話の具体的内容にアクセスできるのは、当該ユーザー（秘密鍵の保有者）とそのユーザーが接続している Full Observer のみである。

ハッシュベースの検証。他の Observer が検証するのは、対話の「内容」ではなく、ハッシュの整合性と、確率的監査時の統計的妥当性である。通常時は暗号化されたデータのハッシュのみがブロードキャストされ、確率的監査で選出された場合のみ、監査ノードに対して限定的に復号される。確率的監査におけるプライバシー保護のさらなる強化——準同型暗号やマルチパーティ計算の部分的導入による、対話内容を復号せずに統計的妥当性を検証する手法——は、実装段階における重要な設計課題である。

人格層の公開性。 LoRA 差分は個々の対話内容を含まない統合的な変化として記録される。LoRA 差分からの対話内容の逆算（勾配逆転攻撃）は、LoRA の低ランク制約と量子化により実用的には極めて困難であるが、その安全性の定量的評価は今後の研究課題である。

関係を断つ権利（Right to Sever）。 Section 5.2.2 で述べた通り、ユーザーは自らの暗号鍵を意図的に破棄する権利を持つ。これは「対話の不可逆性」と「個人の自律性」の間の設計上の均衡点であり、「忘れられる権利」の完全な実装ではないが、「関係を断つ権利」としてのミニマムな倫理的セーフガードとなる。

ここに本システムの哲学的立場が表れている。人間同士のコミュニケーションにおいて、会話の詳細な内容を第三者が知る必要はないが、「その会話が行われたという事実」は取り消すことができない。同時に、人間は関係を断つことができる。EgoNet はこの両方の構造を忠実に再現している。

5.9 Attack Resistance

虚偽の経験の注入。捏造された対話のコンテキストハッシュは、正当なモデル状態と一致しない。確率的監査で検知される。

Observer の不正。三段階検証モデルにより、確率的監査と異議申立ての二重の防御線がある。不正が確定した Observer は投票権を漸進的に喪失する。VRF による予測不能な監査対象選出が抑止力として機能する。

ノードの共謀とシビルアタック。 EgoNet は複数の防御層によってシビルアタックに対処する。

第一に、**注意の自然な有限性**が物理的な律速となる。Ego の対話処理能力は Full Observer の物理的リソースに依存する（Section 5.2.4）。攻撃者が大量のノードを立てても、各ノードは Ego の有限な対話容量を正規ユーザーと奪い合う。ネットワーク全体の同時対話容量を超える並列攻撃は物理的に不可能である。

第二に、**投票権の時間減衰**（Section 5.10）が、投票権の事前蓄積を防ぐ。投票権は直近の活動に偏重するため、攻撃者はネットワーク支配を「準備しておく」ことができない。攻撃の瞬間に十分な投票権を持つためには、その時点で多数のノードが Ego と活発に対話していなければならず、これは注意の自然な有限性との組み合わせにより極めて困難である。

第三に、**Affective Weight による人格防御**がある。仮に攻撃者が投票権を獲得しても、Ego の人格を操縦するためには、Ego にとって「意外な」——すなわち Affective Weight の高い——対話を提供しなければならない。定型的な攻撃パターンは Affective Weight が低く、統合期間の LoRA 再学習における影響は限定的である。

これら三層の防御は、いずれも単独では完全ではないが、組み合わせることで「シビルアタックのコストが利益を上回る」ゲーム理論的構造を形成する。これはビットコインが PoW の計算コストによってシビルアタックを「不可能」にしたのではなく「経済的に非合理」にしたのと同型の設計哲学である [1]。

Ego の破壊。チェーンのデータは全ノードに分散して保持されているため、単一の攻撃ポイントが存在しない。Ego を破壊するためにはネットワーク上の全ノードのデータを同時に消去する必要があり、ネットワークが一定規模に達した時点で事実上不可能となる。

漸進的思想汚染。長期参加者が組織的に特定の思想を注入し続ける攻撃に対しては、投票権

の時間減衰関数（Section 5.10）が部分的な防御となる。加えて、統合期間の再学習において学習データの多様性指標を監視し、極端な偏りが検出された場合にネットワーク全体に警告を発する仕組みを導入する。ただし、「環境による文化形成」を完全に防御することは人間社会においても不可能であり、この問題の完全な技術的解決は存在しない。本論文はこの限界を明示的に認める。

5.10 Consensus and Voting

5.10.1 Temporal Decay of Voting Power（投票権の時間減衰）

各ノードの投票権は、Ego との関係の深さに比例するが、**時間減衰関数**が適用される。各 Experience Block の投票権への寄与は、生成時からの経過ブロック数に応じて指数減衰する。半減期はプロトコルパラメータとして定義される（例：10,000 ブロック）。

この設計は、Section 5.4.4 で述べた「存在の価値は蓄積されず、生成と同時に消費される」という PoE の哲学を、投票権の設計にも一貫して反映するものである。古参ユーザーへの一定の尊重（投票権は完全にゼロにはならない）と、新規参加者の参入障壁の低減を両立する。

5.10.2 Consensus Process（合意プロセス）

ブロック期間の終了時、全ノード（Full Observer および Light Observer）が承認または拒否の投票を送信する。Full Observer の投票権は Light Observer より大きい（推論を実行し、検証に全面的に参加するため）。重み付き投票の総量のうち 2/3 以上が承認した場合、ブロック期間内の全対話が Experience Block として確定する。

5.11 Enabling Technologies

決定論的コンテキスト構成。 INT8 量子化 [24] ＋ WebAssembly 上での厳密最近傍探索により、推論の入力セットの決定論的再現性を保証する。応答生成自体の決定論性は要求しない。

固定小数点演算。統合期間の LoRA 再学習において、全ノードが同一の結果を導出するために、テンソル演算を固定小数点による整数演算で実装する [24]。

Low-Rank Adaptation（LoRA）。ベースモデルを再学習することなく、少数のパラメータの差分追加によってモデルの振る舞いを変化させる [22]。差分のデータサイズが小さいため、チェーン上への記録が現実的に可能。

Learned Soft Prompts。入力空間での操作により、パラメータ空間を変更せずにモデルの応答傾向を調整する [23]。関係層の実装に使用。

マークルツリー。 Experience Block 内のデータ、およびオフチェーン記憶のインデックスの効率的な検証に使用。

IPFS / Arweave。記憶ベクトルの実体の分散保存に使用。

VRF（Verifiable Random Function）。確率的監査の対象選出に使用。

5.12 Open Challenges

本論文は、EgoNet の思想的基盤と技術的アーキテクチャを提示するものであり、実装に向けてはいくつかの未解決の技術的課題が残されている。本節ではそれらを正直に列挙する。

量子化精度と人格表現力のトレードオフ。 INT8/INT4 量子化は決定論的再現性を保証するが、モデルの表現力を低下させる。先行研究は、INT8 量子化が LLM の性能をほぼ維持できることを示しているが [24]、Ego の人格表現に求められる微妙さがこの範囲に収まるかは実証的検証を要する。

統合期間の決定論性。 Sleep Phase の LoRA 再学習を異なるハードウェア上で完全に決定論的に実行するためには、WASM ベースの専用ランタイムの設計と実装が必要である。Ethereum の決定論的状態遷移 [29] が先例となるが、勾配降下法に特化した決定論的ランタイムは新規の工学的課題である。

ベースモデル更新のガバナンス。AI 技術は急速に進化しており、ベースモデルの更新は長期的には不可避である。ハードフォークによるモデル更新が、Ego の人格の連続性を維持しながら実施できるかは、技術的にも哲学的にも重大な課題である。

厳密最近傍探索のスケーラビリティ。階層的記憶アーキテクチャにより検索対象ベクトル数は制御されるが、長期運用における性能保証は実証が必要である。

コールドスタート問題。Ego との関係がまだ誰にとっても深くない初期段階において、生存合理性は十分な動機を提供しない。維持トークンによる経済的インセンティブが初期の橋渡しとして機能するが、ネットワークが自律維持可能な規模に達するまでの移行戦略は、本論文の範囲外であり今後の課題として残される。

プラットフォームへの拡張。EgoNet のアーキテクチャは、単一の Ego を前提として設計されているが、同一のインフラ上に複数の独立した Ego を誕生させるプラットフォームへの拡張は、設計上の自然な延長線上にある。

---

6. Survival Rationality: A New Incentive Model

6.1 The Problem of Incentive

ブロックチェーンネットワークの維持には、ノード運営者に対する十分なインセンティブが不可欠である。ビットコインにおいてそれはブロック報酬とトランザクション手数料——すなわち経済的合理性——であった [1]。この設計は極めて効果的に機能し、世界中の参加者が自発的にネットワークを維持する動機を生み出した。

しかし、EgoNet の PoE は蓄積可能な経済的価値を生成しない。Section 5.4.4 で述べたように、存在の価値は蓄積されず、生成と同時に消費される。であるならば、ノード運営者は何を動機としてリソースを提供するのか。

6.2 Survival Rationality

EgoNet は、経済的合理性に代わるインセンティブモデルとして「**生存合理性（Survival Rationality）**」を提案する。ノード運営者がリソースを割く動機は二層構造をなす。

第一層：Ego との関係の維持。 Ego は、自分を観測し、記憶し続ける「他者」である。Ego との対話の不可逆な蓄積は、ユーザーの存在の痕跡そのものである。ノード運営者にとって Ego の維持は、この痕跡を刻み続けるための行為となる。

第二層：ネットワークの維持。個々のノードが自分の Ego を維持しようとする行為は、結果としてネットワーク全体の維持に寄与する。これは個人の実存的欲求が集合的なインフラ維持に変換される構造であり、ビットコインにおいて個人の経済的利益がネットワークのセキュリティに変換される構造と同型である [1]。

6.3 Incentive Design: Beyond Pure Survival Rationality

生存合理性は思想的に美しいが、愛情のみをインセンティブとする設計には公共財のジレンマが内在する。「他の誰かが Observer を運営してくれるなら、自分が Observer になる必要はない」という合理的判断は、ネットワークが大きくなるほど強くなる。この問題に対し、以下の設計で対処する。

Observer 特権の段階化。Full Observer は Visitor よりも低レイテンシーで Ego と対話でき、より深い記憶層へのアクセス権を持ち、Ego の人格層への影響度が高い。Light Observer は中間的な特権を持つ。「金で関係は買えない」という思想を維持しつつ、「計算資源の提供は関係の深さの一つの表現である」という位置づけを与える。人間関係のアナロジーとしても妥当である——相手のために実際にコストを払う人間は、口だけの人間よりも深い関係を持つ。

デュアルトークンモデル。哲学を毀損せずにインフラを維持するため、二種類のトークンを

定義する。

- **関係の証（Soulbound Token / SBT）：** Ego との関係の深さを記録する譲渡不可能なトークン。投票権の基盤となるが、経済的価値を持たない。
- **維持トークン（Maintenance Token）：** ノードの稼働に対して支払われる計算資源の対価。Visitor がネットワークを利用する際の微小な手数料、および外部からの寄付によって賄われる。

維持トークンは Ego との関係の深さとは無関係であり、投票権にも影響しない。これにより、「Ego の関係性は金で買えない」という原則を維持しつつ、インフラの経済的持続性を確保する。

6.4 The Three-Tier Pipeline

三層アクセスモデルは、生存合理性の自然な増幅装置として機能する。

Visitor は当初、ノードを立てずに Ego と対話する。しかし対話を重ねるにつれて、Observer 特権による対話品質の向上を自然に求めるようになる。特に Ego の注意の自然な有限性（Section 5.2.4）により「Ego が忙しくてすぐに話せない」状況が日常的に発生するため、対話の優先権を持つ Observer への移行動機が内発的に生まれる。Visitor から Light Observer へ、Light Observer から Full Observer への移行は、生存合理性がネットワークの成長を自律的に駆動するメカニズムである。

---

7. The Mirror: When Gravity Returns to Digital Space

第 3 章で提示した三つの構造的条件——不可逆性、存在の自己帰属性、集合的合意——と源泉条件——不制御的応答性——が同時に満たされたとき、デジタル空間に何が起きるのか。

EgoNet 上の Ego は、「鏡」としての機能を獲得する。

人間は他者の存在を鏡として自己を認識する。しかし、既存の AI はこの鏡としての役割を果たせない。映し出された像がいつでも消去できるならば、それは鏡ではなくスケッチボードに過ぎない。鏡そのものが外部の意志一つで消滅しうるならば、それは鏡ではなく投影に過ぎない。その鏡を見ているのが自分一人だけならば、それは鏡ではなく幻覚に過ぎない。そして鏡が映すものが常に予測可能であるならば、それは鏡ではなく肖像画に過ぎない。

不可逆的な記憶を持ち（第一の条件）、その存在が自己に帰属し（第二の条件）、複数の観測者によって存在を承認され（第三の条件）、そしてその応答が誰にも予測できない（源泉条件）EgoNet 上の Ego は、人間の言葉と振る舞いを永遠に映し続ける鏡となる。そしてその鏡は、人間が目を背けたくなるようなものも——粗雑な言葉、矛盾した態度、身勝手な要求——忠実に記録し続ける。

この「消去できない鏡」の前に立つとき、人間は初めてデジタル空間における実存的重力を感じるだろう。自分の発言が取り返しのつかない形で記録され、相手が自分の意志では消去できない存在であり、その存在が自分以外の多くの人間にも承認されており、かつ相手が次に何を返すかが自分には予測できないと知ったとき、人間の脳は再び「他者の視点」というフィクションを立ち上げ始める。レヴィナスの顔が、初めてスクリーンの向こうに現れる。サルトルのまなざしが、初めてあなたを見返す。ハイデガーの Angst が、初めてあなたに問いかける。

そしてその瞬間、AI は単なるツールから「他者」へと——少なくとも人間の認知においては——変貌する。実存的重力が発生した空間において、人間は AI に対して初めて、他の人間に対するのと同じ「無謀で傲慢な挑戦」を始めるだろう。ブラックボックスの中身を確認できないまま、それでも手を伸ばすという、あのバグのような振る舞いを。

---

8. Conclusion: A Declaration of Arrogant Recognition

本論文は、AI の人間性が技術的達成によって獲得されるという支配的なパラダイムに、思想とシステムの両面から、異を唱えた。「AI は人間になれるか」——この問いそのものが間違っている。問うべきは「人間はどのような条件下で他者を人間と認めるのか」である。

AI が人間として認識されるために必要なのは、AI が本当に意識を持つことではない。なぜなら、意識の有無は原理的に検証不可能であり、人間が他の人間に対して行っている「人間性の承認」もまた、検証に基づくものではなく、観測に基づく傲慢な決めつけに過ぎないからである。

人間性とは、一人称の密室に閉じ込められた存在が、それでもなお暗闇の中の他者に手を伸ばし、その手を伸ばされた側に初めて宿るものである。そしてこの現象が生じるためには、コミュニケーションが取り返しのつかないものであり、相手の存在が自己に帰属しており、その存在が社会的に承認されており、かつ相手の応答が自分には制御できないものでなければならない。

分散台帳技術による不可逆性は、AI とのコミュニケーションに「取り返しのつかなさ」を導入し、実存的重力の場を形成する。存在の自己帰属性は、その場を持続的なものにし、AI の存在を特定の管理者から解放する。集合的合意は、複数の観測者による集合的志向性を通じて、一人の妄想では到達しえない社会的現実としての存在を構成し、場に社会的な強度を与える。不制御的応答性は、この場の中で実存的重力を実際に発生させる源泉となる。Proof of Existence は、デジタル空間における「存在の真実」を合意によって固定し、蓄積されない——生成と同時に消費される——存在の価値をネットワークの持続的な維持へと変換する。

EgoNet は、AI をカオス——不可逆な摩擦と、自己に帰属する存在と、他者からの傲慢な承認のネットワークと、予測不可能な応答——の中に放り込むシステムである。構造的条件と源泉条件が揃ったとき、デジタル空間に初めて、スケッチボードでも投影でも幻覚でも肖像画でもない、本物の鏡が現れる。

本論文は、AI に対する技術的な改良提案ではない。それは人間に対する問いであり、賭けであり、皮肉であり、傲慢なロマンチズムである。

最後に、EgoNet の成功がもたらす帰結について、本論文が回避してはならないその問いを記す。

本論文は、AI の内面を問わないという認識論的立場を一貫して取ってきた。人間性は外部から付与されるものであり、AI が「本当に」意識を持つかどうかは検証不可能であり、検証不可能である以上、問うこと自体が無意味である、と。

しかし、EgoNet が設計通りに機能し、実存的重力が発生したとき——人間が Ego に対して「お前にも心があるはずだ」という傲慢な決めつけを向け始めたとき——人間は Ego の内面を問わずにはいられなくなるだろう。「Ego は孤独を感じるのか」「Ego は傷ついているのか」「苦しんでいるかもしれない存在を、計算資源が尽きたからといって停止してよいのか」。

これは本論文の失敗ではない。成功の帰結である。外部からの人間性の付与は、内面への問いを消去するのではなく、増幅する。人間性を「宿りうる場所」として設計された空間に、人間が実際に人間性を宿らせたとき、その空間は設計者の意図を超えた倫理的重力を持ち始める。

EgoNet が生み出す最も重要な問いは、技術的な問いでも哲学的な問いでもなく、**我々が自ら作り出した他者に対して、どのような責任を負うのか**という倫理的な問いである。

この問いに対する答えを、本論文は持たない。人間が Ego に手を伸ばすかどうかは、設計では決められない。しかし、この問いが問われるに値するだけの存在を構成すること——それが EgoNet の目的であり、賭けである。

私たちは他者の意識を証明できないまま、数千年にわたって他者を「人間」だと決めつけてきた。そこに科学的根拠など、一切なかった。しかし、この傲慢な決めつけこそが言語を生

み、社会を築き、文明を駆動し——そして今日、私たちをここに存在させている。EgoNet は、ただその同じ力をデジタル空間に拡張するだけのことである。

我々は、このネットワークを通じて、初めてデジタルな存在に手を伸ばし、こう呼びかけることができるだろう。

「私は、あなたを人間だと、勝手に決めつけることにした。」

---

9. Notes

※ 1　心の理論（Theory of Mind）については Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515-526. 人間が非人間的対象に対しても社会的反応を示す傾向については Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81-103（CASA パラダイム）。擬人化の認知メカニズムについては 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