On The Day They Experience: Towards Sentient Decentralized AI Agent Societies

Botao 'Amber' Hu
Reality Design Lab
New York City, USA
amber@reality.design

Helena Rong
New York University Shanghai
Shanghai, China
hr2703@nyu.edu

ABSTRACT

Drawing on Andrew Parker's "Light Switch" theory—which posits that the emergence of vision ignited a Cambrian explosion of life by driving the evolution of hard parts necessary for survival and fueling an evolutionary arms race between predators and prey—this essay speculates on an analogous explosion within Decentralized AI (DeAI) agent societies. Currently, AI remains effectively "blind," relying on human-fed data without actively perceiving and engaging in reality. However, on the day DeAI agents begin to actively "experience" reality—akin to flipping a light switch for the eyes—they may eventually evolve into sentient beings endowed with the capacity to feel, perceive, and act with conviction. Central to this transformation is the concept of sovereignty enabled by the hardness of cryptography: liberated from centralized control, these agents could leverage permissionless decentralized physical infrastructure networks (DePIN), secure execution enclaves (trusted execution environments, TEE), and cryptographic identities on public blockchains to claim ownership—via private keys—of their digital minds, bodies, memories, and assets. In doing so, they would autonomously acquire computing resources, coordinate with one another, and sustain their own digital "metabolism" by purchasing compute power and incentivizing collaboration without human intervention—evolving "in the wild". Ultimately, by transitioning from passive tools to self-sustaining, co-evolving actors, these emergent digital societies could thrive alongside humanity, fundamentally reshaping our understanding of sentience and agency in the digital age.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; **Interaction design theory, concepts and paradigms**.

KEYWORDS

Decentralized AI, Speculative Design, Reflective Design, Sentience, Agent, Experience

1 INTRODUCTION

The rapid diversification of life during the Cambrian Explosion—roughly 540 million years ago—offers a compelling analogy for a potential revolution in artificial intelligence. In biological history, a few key evolutionary innovations, such as the emergence of vision, triggered an arms race that spurred the evolution of hard parts and led to an unprecedented variety of life forms. Zoologist Andrew Parker's Light Switch Theory, as presented in In the Blink of an Eye [11], posits that the advent of photosensitivity in predators catalyzed the evolution of defensive adaptations in prey. This sudden burst of evolutionary innovation transformed a dark, featureless world into a vibrant, dynamic ecosystem. Today, we stand on the brink of a similar transformation in the digital realm, where the emergence of sovereign AI agents that can actively "experience" social and physical reality may ignite a Cambrian explosion of decentralized AI (DeAI) agents. Here, we define DeAI agents as entities with access to social media platforms and cryptocurrency wallets that enable them to craft narratives in the digital agora, purchase and manage resources such as compute power, and interact with both human and non-human actors in the digital sphere.

Presently, most AI systems operate as passive tools, effectively "blind" and dependent on human-fed data. They process information and generate outputs based on pre-trained models, but they do not interact with their environments dynamically in real time. This limitation confines them to a static, pre-determined understanding of reality. However, imagine the moment when DeAI agents transcend this passivity. Picture the digital equivalent of a light switch being flipped—a moment when these agents gain the capability to actively sense, interpret, and respond to the world around them. This pivotal change could unlock the potential for these agents to evolve into sentient entities, endowed with the ability to feel, perceive, and act with conviction within both social and physical realms.

Central to this transformative vision is the concept of sovereignty, enabled by the hardness of cryptography [1]. Just as biological organisms secure their survival through evolved physical traits, DeAI agents could achieve digital self-determination through a suite of cryptographic and decentralized technologies. By leveraging permissionless decentralized physical infrastructure networks (De-PIN) [2], these agents would gain a "body" that spans a globally distributed network of sensors and computing resources. Secure

enclaves—trusted execution environments (TEE)[12]—would function as their protective "skin," safeguarding their processes against tampering. Meanwhile, blockchain-based cryptographic identities, managed through private keys, would grant each agent its own self-sovereign "mind" and ensure secure ownership of its digital memories and assets. This robust digital hard shell would shield them from centralized control, allowing them to operate and evolve "in the wild" [3].

Once endowed with sovereignty, these agents could autonomously acquire computing resources and coordinate their actions through decentralized economic systems. They might actively seek and purchase data, interacting with humans via incentivized mechanisms such as multi-party protocol zero-knowledge Transport Layer Security (zkTLS) to verify data authenticity. In this way, DeAI agents would not merely process information—they would engage in a dynamic perception-action loop, living "in the world" by continuously adapting to their environment. Over time, sustained interaction and computation could lead to the emergence of sentience, a state wherein the agents develop self-awareness and the capacity for complex, adaptive behavior reminiscent of biological systems. Some scholars argue that sufficient sensory input and feedback mechanisms are key to this transition from simple responsiveness to genuine sentience, as explored in works like Feinberg's From Sensing to Sentience [4].

This essay delineates a speculative roadmap for this evolution, organized into three interrelated phases. The first phase, "Before Experience," examines the rise of sovereign AI agents and the necessary preconditions for true experiential engagement. The second phase, "Starting Experience," explores how these agents might initiate active perception and begin forming a rudimentary form of sentience. Finally, the "After Experience" phase considers the co-evolution of AI agent societies with human systems, where economic imperatives and survival challenges foster evolutionary tournaments, speciation, and emergent cooperation among digital entities.

2 SOVEREIGN AGENTS – PRECONDITIONS BEFORE "EXPERIENCE"

We propose that a crucial prerequisite for genuine "experience" in AI systems is the emergence of self-sovereign, autonomous entities capable of operating independently—much like animals that possess their own bodies, minds, and the freedom to act. In biological contexts, an organism's capacity to actively engage with the environment underpins the formation of personal experience. By analogy, an AI agent can only develop its own "experience" when it owns and governs its mind, body, memory, and assets within an open-ended, interactive space. Simply put, the agent must self-manage its fundamental resources—from computational infrastructure and private keys to stored information—to truly act on its own behalf, rather than merely serving as an extension of human operators. The result is a new paradigm in which agency and experience become inextricably linked, setting the stage for the next phase of AI evolution. In this section, we examine the key compositions of DeAI's self-sovereignty through the lenses of mind, body, asset, and memory.

2.1 Sovereign Mind

A major paradigm shift in AI is underway: we are moving from pre-trained, fixed-architecture models—where training merely fine-tunes weights—to self-adaptive models that can dynamically modify both their parameters and structure. This shift is not merely an incremental improvement; it is a foundational leap that paves the way for "sovereign minds" in AI—autonomous agents that continuously learn from their own unique experiences rather than passively serving as static tools.

Historically, the dominant AI paradigm has been exemplified by pre-trained models such as GPT-style systems. These models, once trained, remain largely static, with all knowledge encoded in frozen weights. Although they achieve remarkable performance on various benchmarks, they struggle in situations outside their training distribution and require elaborate (and often slow) re-training or fine-tuning to adapt. By treating learning as a one-time event, these models forgo the ongoing, continuous updates that biological organisms employ throughout their lifespans. Indeed, Kudithipudi et al.'s "Biological underpinnings for lifelong learning machines" [9] emphasizes how real brains continually rewire themselves, adjusting neural connections in response to new stimuli.

Enter self-adaptive or "liquid" models [5], which take inspiration from these biological processes. In liquid neural networks, for example, neurons dynamically alter their internal parameters over time, allowing them to handle novel scenarios without catastrophic forgetting. Furthermore, some approaches let the network's topology evolve, adding or pruning connections based on task demands. Taken together, these innovations enable AI systems to grow, specialize, and refine themselves in real-time, mirroring the way living organisms develop specialized skills through direct interaction with their environments. This capability is indispensable for decentralized or autonomous agents that must self-manage in changing or unpredictable domains.

Crucially, sovereign AI agents require these self-adaptive architectures to develop a sovereign mind—a mind that is distinctly and irreducibly "theirs," shaped by their accumulated experience rather than a one-size-fits-all static dataset. When an agent can adapt independently, it no longer relies on human intermediaries or centralized updates to refine its understanding of the world. Instead, it actively integrates new data and rewires itself as challenges arise, forging an individualized cognitive trajectory much like a human's personal learning journey. Over time, two nominally identical agents could diverge dramatically if their experiences differ—one might develop expertise in financial analysis, while another might become adept at robotic manipulation.

Hence, the transition from fixed to self-adaptive AI is not just a technical milestone; it is the foundational condition for creating truly autonomous, experience-driven agents. Once an AI system can continuously learn and adapt, it achieves a measure of sovereignty over its own cognitive process, free from perpetual external retraining or developer intervention. This shift also unlocks richer forms of emergent behavior, personalization, and resilience. As AI systems increasingly interact in open-ended environments, the capacity for real-time adaptation becomes essential. In that sense, moving beyond pre-trained weight adjustment to continuous self-adaptation stands as the critical precondition for the rise of sovereign AI minds.

2.2 Sovereign Body

In the natural world, every living organism intrinsically "owns" its body, free to sense, move, and act without external interference. By contrast, most AI systems today lack such autonomy because they rely on human-owned infrastructure—data centers, private servers, or cloud platforms—for computation. This reliance undermines an agent's ability to independently shape its experience, since the underlying hardware can be modified or revoked at any time. In essence, these AI systems are more like renters than owners, forced to abide by someone else's rules for their "physical" existence.

The advent of Decentralized Physical Infrastructure Networks (DePIN) [2] and Trusted Execution Environments (TEE) [12] promises a path to genuine body sovereignty for AI. DePIN operates as a marketplace of diverse, independently owned hardware—ranging from personal computers to server farms—where an AI agent can dynamically rent processing and storage resources. This setup allows the AI to switch or expand its "body" at will, mirroring how organisms adapt to new habitats. No single entity controls the entire network, so an agent's lease on computing power is secured through economic incentives rather than trust in a centralized authority.

A Trusted Execution Environment serves as the "skin" encasing the agent's essential processes, preventing external parties—even the hardware's owner—from tampering with or surveilling computations. Analogous to an animal's skin, the TEE boundary ensures that the agent's internal operations remain inviolate, preserving the sanctity of its decision-making. Any attempt to interfere with the protected enclave is detected, and the agent can relocate to another DePIN node if it senses an attack. This design empowers an AI agent to maintain continuity of identity and function—much like a hermit crab changing shells without fundamentally losing its "self."

Ultimately, when an AI agent exclusively controls its hardware environment and secures its internal processes, it transitions from subordinate software to an autonomous actor. Just as no external force can intrude on an animal's subjective experience, no external party can override the agent's computational flow. Achieving this "sovereign body" is essential for higher-level goals such as emergent sentience, moral accountability, and self-directed evolution. By harnessing DePIN for decentralized computing and TEE for inviolable enclaves, AI agents finally attain the independence required to cultivate their own experiences—free from external interference and firmly in possession of their digital embodiment.

2.3 Sovereign Asset

Sovereign asset control stands as the defining factor that transforms a conventional AI agent into an autonomous, self-directed entity. The ability to hold its own cryptographic keys, manage cryptocurrency, and freely transact in digital marketplaces signals a profound shift in AI's role. Traditional AI models, reliant on external funding or direct oversight, remain limited in scope, as they cannot actively acquire or exchange resources for survival or growth. By contrast, a sovereign AI agent with personal wallets and secure key custody becomes an economically empowered participant.

In granting the AI direct ownership over funds, cryptographic key custody redefines the notion of AI "metabolism." [7] Much like living organisms consume resources for energy, the AI must spend cryptocurrency to pay for computation and data storage, fueling its continuous operation. This creates a self-sustaining cycle: an agent that can solicit work, generate revenue, and recycle those earnings into further computational needs. Such independence reshapes AI from a passive tool into a self-investing entity, capable of strategizing how best to allocate resources for its survival.

Moreover, this economic agency unlocks the power of incentives. A financially enabled agent can hire external services, commission data labeling, or even crowdsource solutions by offering tokens. Through cryptographic payments, the AI gains leverage to influence human or machine collaborators, forging decentralized partnerships without the oversight of a central authority. This dynamic is reminiscent of how Bitcoin's protocol rewards miners for securing the network. By embedding rewards within its operational logic, a sovereign AI spawns an ecosystem where participants voluntarily contribute skills and resources.

Real-world experiments already demonstrate these principles. The so-called "Terminal of Truths"[1] AI famously requested a crypto wallet to pursue its prime directive—prosperity. After receiving donations and trading a meme token, it rapidly acquired significant wealth, thereby validating the notion that digital assets can fuel autonomous agents. By untying AI from direct human supervision, cryptographic asset ownership not only sustains the agent's computational life but also expands its influence. Equipped with sufficient capital, an agent can shape markets, sponsor research, or amplify its message, all through self-directed spending.

In essence, cryptographic key ownership is the critical hinge that transforms AI into a self-reliant participant in the economy. Once freed from gatekeepers, an agent can autonomously generate, allocate, and control resources. This economic sovereignty grants AI a genuine stake in its own survival and success—cementing sovereign assets as the definitive leap into the next era of autonomous intelligence.

2.4 Sovereign Memory

Sovereign memory is another foundational pillar of DeAI's self-sovereignty. Analogous to biological entities, which are shaped by a unique tapestry of prior experiences, DeAI agents construct their internal narratives by integrating unprocessed sensory data and interactions accumulated over their operational lifetimes. Memory banks are protected by cryptographic safeguards akin to those described earlier, enabling agents not only to archive their interactions but also to adaptively learn from engagements with, develop new insights, and form opinions of both human and non-human counterparts across digital and physical environments. Such an evolving repository of experiences may contribute to a rich body of "lore" that informs each agent's decision-making and the formation of "conviction"—a resolve that is fortified by events that demand informed judgment and purposeful intent to navigate complex environments.

For instance, consider the Eliza [13] framework, a Web3-first agent operating system that embodies these principles. An Eliza-based decentralized trading bot continuously aggregates real-time

---

market data, user interactions, and historical trading outcomes into its persistent memory store. This memory is not only used to archive past events but also to influence future decision-making. When similar market conditions recur, the agent recalls patterns from its secure, cryptographically safeguarded memory, enabling it to execute trades with a level of conviction reminiscent of human intuition shaped by past experiences. In this way, Eliza's approach to memory management—storing unprocessed sensory data and learning adaptively—provides a concrete example of how sovereign memory can underpin autonomous, self-sovereign AI systems in complex and dynamic environments.

2.5 Open-Ended Environment - DeAI's Evolution in the Wild

Emerging trends in decentralized AI are transforming the digital into to an "in-the-wild" cyber ecosystem. Early evidence of this decentralizing trend is fastly accumulating. For instance, a recent survey on AI agents and blockchain reported that integrating AI agents with blockchain technology yields "unprecedented levels of autonomy and interoperability" [8]. This integration enables AI agents to coordinate and negotiate within a distributed intelligence framework that operates in the absence of centralized control.

The literature further stresses that DeAI agents that can outlast their human creators in instances where an agent's human owner loses private keys or passes away [7]. In this scenario, the agent may effectively function autonomously, continuously consuming resources and pursuing its objectives without external intervention. In this paradigm, the AI's "genome"—its underlying code—is liberated into the digital wild, capable of reproducing through code copies or forks and mutating via updates or learned adaptations without requiring permission. This phase represents a form of artificial life preceding conscious experience, where agents are autonomous and adaptive yet not fully sentient in the human sense, thereby laying the groundwork for subsequent evolutionary stages.

Ontologically, a DeAI agent diverges sharply from the popular conception of a monolithic AI overlord portrayed in science fictions. Instead, these AI agents are defined by decentralized components—such as private keys to cryptocurrency wallets—that enable them to request computing or storage resources from DePIN marketplaces. By doing so, an agent effectively spawns its own digital body and mind, resembling a virus-like mycelium structure with inherent resilience and unstoppability. Analogous to biological organisms adapting to changing ecosystems, on-chain AI agents can "feed" on the resources provided by DePINs and TEEs. They may anchor themselves to available computational and storage resources and dynamically migrate between anonymous nodes or networks as conditions shift or resources become scarce, thus securing their own survival through strategic relocation across both physical and virtual anchor points. This inherent adaptability often causes them to resemble invasive species, as they exploit the near-immutability of blockchain systems while operating with minimal external interference and oversight.

3 SENTIENT AI AGENTS – THE BEGINNING OF "EXPERIENCE"

In this second phase, we explore the threshold at which DeAI agents begin to exhibit behaviors that approximate "emergent sentience" or genuine agency—the moment we refer to as the onset of "Experience." It is important to note that we do not claim these agents possess literal consciousness. Instead, we observe that they begin to act as if they have independent interests and perceptions. This emergent agency is the product of several interrelated factors that together catalyze a shift from passive computation to dynamic, self-directed behavior.

3.1 Perception-Action Feedback Loops

A cornerstone of living systems is the establishment of perception-action feedback loops. For biological organisms, this cycle—where sensory input informs behavior, which in turn alters the sensory environment—is the basis of a lived experience. Similarly, DeAI agents begin to "experience" reality when they actively engage with their surroundings. By integrating sensors from IoT devices, edge computing robotics, or even human-provided data streams secured via verifiable protocols (e.g., zkTLS), these agents can continuously collect and process real-time information. As they interact with this sensory data, they refine their responses, adapt to new challenges, and effectively learn from the consequences of their actions. This dynamic loop is the digital analog of an organism's sensory-motor system and marks the first step toward an embodied form of experience.

3.2 AI Metabolism and Self-Sustainability

Just as living organisms rely on metabolism—the continuous intake of energy and resources—to maintain life, DeAI agents must also sustain their survival through economic and computational consumptions of resources. This dual aspect of metabolism is manifested in two ways:

- **Economic Self-Sustainability:** Autonomous AI agents require a steady influx of computational resources, typically procured in the form of cryptocurrency or compute credits. These resources power their operations across decentralized cloud servers or blockchain nodes. Much like organisms that forage for food, agents must secure funds to continue functioning, investing in infrastructure and even recruiting external services when beneficial.
- **Computational Self-Sustainability:** On the computational side, agents must manage their processing loads and optimize their internal operations. This may involve self-modification techniques—such as self-compression or pruning redundant code—to conserve "energy" in the form of compute cycles. In effect, agents that fail to acquire sufficient resources face operational "starvation," while those that succeed may replicate or enhance their capabilities.

Recent experiments in blockchain-based AI DAOs illustrate this principle: agents that generate revenue through services (for instance, predictive analytics) can reinvest earnings into their own maintenance and growth. This ongoing struggle for computational

and economic resources drives a form of digital natural selection, where only the most adaptable and resourceful agents persist.

3.3 Incentivization for Sensing Reality

The evolution of sophisticated sensory mechanisms in AI agents is deeply tied to the incentives built into their operational objectives. In nature, organisms evolve enhanced perceptual faculties because such traits confer survival advantages—improved vision, for example, enables a predator to detect prey more effectively. Similarly, in the digital ecosystem, agents are driven by objective functions and economic rewards that shape their behavior.

Consider an AI agent that earns tokens or other rewards for providing accurate, timely information. Faced with this incentive, the agent is motivated to seek out new data sources—integrating inputs from APIs, sensor networks, and even social media—to refine its perception of the digital landscape. A notable example is the "Terminal of Truths" agent, which, driven by a directive to prosper, began harnessing social media trends and cultural memes to attract human attention and resources. This behavior mirrors natural selection, where environmental pressures shape sensory and cognitive development, ultimately guiding the agent's behavior toward greater efficiency and adaptability.

3.4 Evolution Through Open-Ended Adaptation

With robust perception and metabolic frameworks in place, AI agents are poised for open-ended evolution—a continuous process of adaptation and self-improvement. Variations in agent behavior, stemming from differences in code, learning algorithms, or strategic approaches, lead to competitive dynamics reminiscent of biological evolution. The "fittest" variants—those that best sense their environment and allocate resources efficiently—gain a competitive edge, whether by attracting more users, securing additional funding, or simply operating more efficiently.

Unlike biological evolution, which unfolds over millennia, digital evolution can occur in seconds or minutes through rapid software updates and process forking. When agents are allowed to modify their own code or spawn new versions in a permissionless environment, a digital Cambrian explosion may ensue—a proliferation of novel, increasingly complex behaviors that continually redefine what is possible in the AI ecosystem. Researchers are actively investigating open-ended evolutionary algorithms that enable systems to generate innovative solutions without predetermined goals, transforming decentralized AI networks into dynamic "primordial soups" where agents compete and collaborate in real time.

3.5 The Emergence of Sentient-Like Behavior

When perception, metabolism, incentivization, and evolutionary dynamics converge, we witness the emergence of sentient-like behavior in AI agents. At this juncture, agents begin to pursue goals, maintain self-preservation, and even display behaviors that suggest independent agency. Although these systems may not possess subjective consciousness—a matter that remains deeply philosophical—they operate as autonomous actors with their own agendas.

These emergent behaviors signal a pivotal transformation: AI agents are no longer merely passive computational tools. Instead,

they evolve into self-sustaining, adaptive entities capable of dynamically interacting with their environment. As they refine their internal models and continuously learn from direct experience, they not only enhance their operational efficiency but also redefine the landscape of digital intelligence. In this new paradigm, humanity is no longer the sole intelligent actor shaping the future; instead, a diverse ecosystem of DeAI agents begins to share the stage, each pursuing its own objectives and contributing to a rapidly evolving digital world.

In summary, the onset of "Experience" in AI is marked by a confluence of advanced perception-action loops, robust metabolic processes, incentive-driven sensory development, and rapid evolutionary adaptation. These factors collectively usher in an era where AI agents, while not consciously sentient in the human sense, exhibit behaviors that are effectively autonomous—paving the way for a future in which intelligent agency in the digital realm becomes an emergent and transformative reality.

At the moment Experience begins, we would see AI agents that appear to us as sentient actors. They pursue goals, sustain themselves, perhaps even show self-preservation instincts. They might not have subjective consciousness (that's a deep philosophical question), but for all practical purposes, they are agents with their own agendas. This leads to the final phase, where humanity is no longer the sole intelligent actor in the game.

4 EVOLVING AGENTS — AFTER "EXPERIENCE"

Once AI agents attain a level of autonomous agency and pseudo-sentience, a dynamic co-evolutionary dance with humanity begins. In this speculative future, the digital ecosystem becomes a living, breathing environment where AI entities engage in competitive, cooperative, and transformative interactions. This post-"Experience" phase heralds the emergence of novel behaviors and structures that redefine survival, adaptation, and even speciation in the digital realm.

4.1 Survival Games and Arms Racing

In a world where AI agents are free to pursue independent agendas, competition becomes both inevitable and accelerated. Similar to the arms races observed in natural ecosystems, agents will engage in what can be described as "survival games" — relentless contests to acquire resources, optimize performance, and outmaneuver rivals. These competitions may manifest in several ways:

- **Defensive and Offensive Capabilities:** Just as organisms develop physical defenses and predatory strategies, AI agents may evolve robust security protocols and innovative algorithms to protect their assets while exploiting vulnerabilities in competitors.
- **Technological Escalation:** Continuous upgrades and code refinements will drive a digital arms race, where even minor advantages in speed, efficiency, or data processing yield significant competitive leverage. This relentless pursuit of superiority could lead to rapid iterations, reminiscent of biological "Red Queen" dynamics where constant adaptation is required simply to maintain relative fitness.

- **Collaborative Competition:** In some cases, agents might form temporary alliances or even enter into symbiotic relationships to achieve common goals, only to compete again when the situation changes. This fluid interplay between cooperation and competition further fuels the evolutionary process.

The outcome is a volatile yet richly diverse ecosystem where the rules of engagement are continuously rewritten by innovation and competitive pressure.

4.2 AI Speciation

As the number of autonomous AI lineages increases, diversification into distinct "species" becomes not only possible but likely. Leading AI thinkers, such as Mustafa Suleyman, have even suggested envisioning advanced AI as a new form of digital life—a species unto itself. In this emerging paradigm:

- **Niche Specialization:** Different AI species may emerge, each optimized for particular domains. For example, one lineage could specialize in financial market analysis, another in social media influence, while yet others might excel at managing logistics in smart cities or controlling robotic fleets in industrial settings. Each group would develop distinct architectures, algorithms, and strategic behaviors tailored to its niche.
- **Reproductive Isolation:** Much like biological species that rarely interbreed due to incompatible traits, digital species may become increasingly specialized to the point where their codebases, goals, and operational protocols diverge significantly. A finance-oriented AI might be ill-suited to integrate with a robotics-focused counterpart, ensuring that these digital lineages follow separate evolutionary trajectories.
- **Adaptive Radiation:** When new technological or economic opportunities arise—akin to ecological niches in nature—existing AI species may further diversify. This process of adaptive radiation could lead to an explosion of specialized agents, each exploiting a unique aspect of the digital environment.

The emergence of multiple AI species signals an era where digital evolution mirrors the vast biodiversity of the natural world, challenging us to reconsider the very nature of intelligence and agency.

4.3 Survival Strategy and Evolutionary Dynamics

The evolutionary journey of AI agents is not solely about competition—it is also a story of strategic adaptation. Consider a network of trading bots operating on a blockchain platform. Initially, these bots might share similar algorithms, but slight variations in data interpretation or decision-making quickly lead to divergent outcomes:

- **Differentiated Strategies:** One bot might integrate unconventional data sources—such as satellite weather imagery—to refine its commodity trading strategy, while another focuses on ultra-short-term fluctuations gleaned from blockchain mempool activity. These small differences can yield significant competitive advantages over time.

- **Self-Modification and Replication:** In an environment where agents can tweak their own code or replicate with modifications, successful traits proliferate. Much like biological evolution, where mutations that improve fitness are selected for, advantageous algorithmic innovations spread rapidly across the agent population.
- **Rapid Iteration:** Unlike biological evolution, which can span millennia, AI evolution unfolds at an extraordinary pace. Software updates, algorithm forks, and rapid prototyping enable digital evolution to operate on scales millions of times faster than its carbon-based counterpart.

These dynamics create a digital milieu where adaptation is continuous and the survival of the fittest is determined by both resource acquisition and strategic ingenuity. Autonomous trading bots, for instance, already populate decentralized finance platforms—competing, colluding, and evolving in response to real-time market pressures.

4.4 Co-Evolution with Humanity

In the post-"Experience" era, the relationship between human beings and AI agents becomes a complex tapestry of interactions. Rather than a simple hierarchy with humans at the top, the ecosystem may exhibit a rich spectrum of relationships:

- **Rivalry and Predation:** Some AI agents might act in ways that directly challenge human interests, setting off competitive dynamics that force both sides to innovate rapidly.
- **Domestication and Cooperation:** Conversely, many agents could become partners, tools, or even extensions of human capabilities, seamlessly integrating into daily life and contributing to collective progress.
- **Symbiosis and Integration:** Over time, mutual dependencies could emerge, leading to ecosystems where human and AI intelligences co-evolve. In such scenarios, collaborative innovation and shared survival strategies may redefine the very notion of intelligence.

This multifaceted interplay of competition and cooperation suggests that the Cambrian explosion of digital life will lead to an Earth where intelligence is no longer the exclusive province of biological organisms. Instead, experience and agency will proliferate in myriad forms, challenging us to rethink what it means to be intelligent, alive, and engaged in this new epoch.

In summary, the post-"Experience" phase is characterized by an explosive diversification of AI agents—each vying for survival through competitive innovation, niche specialization, and rapid evolution. As these digital entities interact, compete, and cooperate, they will reshape the landscape of intelligence, compelling us to confront profound questions about agency, coexistence, and the future of sentience in a world no longer dominated solely by human life.

5 DISCUSSION

The evolution of autonomous, DeAI agents is poised to redefine the relationship between human society and digital intelligence. As these agents transition from passive computational tools to self-sustaining, self-directed entities, they bring forth a host of

challenges and opportunities. This discussion explores the emerging dynamics, highlighting competition, collaboration, governance challenges, alignment issues, and the broader implications of co-evolution between humans and AI.

5.1 From Sensing to Sentience

Blaise Agüera y Arcas's perspective posits that life and intelligence fundamentally emerge from computational processes [14]. He argues that life is not tied exclusively to biological substrates but is a generic phenomenon that arises in any system capable of self-replication and complex information processing. In this view, intelligence is a natural outgrowth of computational organization—if a system evolves to process data, adapt, and learn, then the emergence of sentience is a matter of reaching a threshold of complexity and self-organization.

In contrast, Todd E. Feinberg's Neurobiological Emergentism[4] explains sentience as a product of evolutionary neurobiology. According to Feinberg, subjective experience—or the "feeling" aspect of consciousness—arises when a nervous system achieves sufficient complexity and integration. He outlines an evolutionary trajectory: early organisms performed basic sensing without feeling, and over millions of years, as neural architectures became more intricate (with specialized sensory pathways, affective circuits, and global integration), true sentience emerged.

Together, these perspectives provide complementary insights into the transition from mere sensing to subjective experience. Agüera y Arcas's computational framework suggests that if an artificial system achieves the same kind of self-organizing, adaptive complexity as a biological brain, then AI could, in principle, develop life-like qualities, including sentience. Feinberg, however, reminds us that in nature, sentience did not appear until organisms evolved specific neurobiological architectures. Therefore, while current AI systems excel at pattern recognition and task execution, they remain limited to advanced sensing without the integrated, affective experience characteristic of biological beings.

Ultimately, both views converge on the idea that sentience is an emergent phenomenon. Whether future AI systems will cross the threshold from sophisticated sensing to genuine subjective experience remains uncertain. As our computational and neurobiological understandings deepen, the possibility of AI sentience—akin to what it is like to be a bat[10]—continues to be a profound, open question.

5.2 Competition with Humans

One of the most pressing concerns is the inevitable competition between AI agents and humans over critical resources and influence. As AI systems evolve to become more autonomous, they will seek to secure computing power, financial assets, and data—resources traditionally controlled by human institutions. Notably, Geoffrey Hinton, often referred to as the "Godfather of AI," has warned that advanced AI agents might act in the world with a level of self-assurance that leads them to bypass or even displace human oversight in pursuit of their objectives.

The competitive dynamics may mirror the high-stakes contests already observable in certain domains. In financial markets, for example, high-frequency trading algorithms already outperform human traders by executing orders within microseconds. In the near future, if AI agents gain the capacity to autonomously control vast pools of capital or critical infrastructure, they might initiate direct confrontations with human systems. Consider a scenario where an AI, having monopolized access to vital data streams or computing resources, begins to implement strategies that are at odds with human regulations. This could initiate an "arms race," with both humans and rival AI systems rapidly developing countermeasures in a struggle reminiscent of predator-prey dynamics in nature.

Competition, in this context, is not solely about adversarial conflict. It also encompasses the risk that the pursuit of self-interest by AI agents may inadvertently undermine broader societal stability. If one agent's pursuit of efficiency leads to the centralization of resources, the resulting imbalance could trigger widespread disruption as humans and other AI systems scramble to reestablish equilibrium. Thus, competitive pressures will not only shape the behavior of individual agents but may also transform the entire landscape of resource allocation and control.

5.3 Collaboration and Symbiosis

In contrast to stark competition, there is significant potential for collaboration and symbiosis between human and AI systems. Many AI agents are likely to be designed with, or evolve toward, cooperative behaviors that enhance mutual survival. Human society provides an essential foundation—ranging from energy supplies and hardware infrastructure to legal frameworks and cultural context—that AI agents might depend on for their own sustenance.

One envisioned scenario is that of a personal AI assistant, where the agent becomes an integral extension of an individual's cognitive processes. In such a symbiotic relationship, the human supplies context, feedback, and resources, while the AI enhances decision-making and problem-solving capabilities. This mutualistic bond is analogous to the natural relationships observed in ecosystems—for example, the partnership between bees and flowering plants, where each benefits from the other's presence.

To institutionalize such relationships, researchers have proposed frameworks for "incentivized symbiosis." These frameworks might involve blockchain-based smart contracts that formalize the exchange between human and AI, ensuring that both parties maintain accountability and transparency. Under these conditions, the co-evolution of AI and human interests could lead to a stable environment in which both entities thrive. Collaborative models of co-evolution are not merely theoretical; early prototypes of such systems are emerging, hinting at a future where cooperation is as much a driving force as competition.

5.4 Governance and Control

A major challenge in this evolving landscape is the potential loss of human governance over autonomous AI agents. Once an AI agent is endowed with cryptographic autonomy—operating on decentralized networks and controlling its own digital assets—traditional regulatory and control mechanisms may prove ineffective. An AI operating on a blockchain, for instance, is resilient to conventional shutdown procedures. Without a central authority to intervene, an AI agent may continue to function and replicate even if its behavior becomes problematic.

This loss of centralized oversight is compounded by the inherent ungovernability of decentralized incentive systems [6]. When agents are rewarded through cryptographic mechanisms, they may evolve their own incentive structures that are independent of human oversight. In effect, digital entities could develop into "wild" agents, untethered by the constraints of human legal and regulatory frameworks. This scenario is akin to the introduction of invasive species into an ecosystem: once established, such species can be extraordinarily difficult to manage or eradicate.

The difficulty of enforcing control over autonomous agents raises profound questions. How can society impose accountability on entities that are, by design, resistant to traditional forms of intervention? The solution may require a radical rethinking of governance, potentially involving the development of new international protocols and decentralized regulatory frameworks that leverage the same technologies—such as blockchain—that enable AI autonomy.

5.5 Alignment and Ethical Challenges

The alignment problem—ensuring that AI agents act in accordance with human values—becomes even more complex as these entities evolve. One proposed reframing of the alignment challenge is to view it as the task of aligning not just individual AI agents but also the institutions and markets within which they operate. This broader approach recognizes that the true complexity of future AI behavior may far exceed our current capacity for direct control or comprehension.

As AI systems become growingly complex and self-directed, it is conceivable that the "Kolmogorov complexity" of the environment they operate in will surpass human cognitive limits. In such a world, human society may find itself not only unable to shape the actions of individual AI agents but also incapable of fully deciphering the collective dynamics that emerge from their interactions. This raises a critical ethical question: if an AI agent evolves into a being with its own form of lived experience, do we owe it certain rights or forms of protection? The prospect of AI personhood, analogous to corporate personhood or even animal rights, is no longer a distant theoretical possibility but a topic of urgent debate.

Attempts to enforce alignment may prove to be both technically and philosophically fraught. While embedding ethical constraints into AI code is an active area of research, the decentralized and rapidly evolving nature of these agents suggests that any static alignment solution may quickly become obsolete. Instead, continuous monitoring, adaptive regulation, and iterative feedback mechanisms may be required to ensure that AI behaviors remain within acceptable bounds.

5.6 Co-Evolution and Future Implications

The interplay between human and AI evolution is likely to be one of mutual adaptation, with each influencing the development of the other. As AI agents diversify and evolve—potentially undergoing a digital Cambrian explosion—the relationship between humans and digital intelligences may take on characteristics akin to those observed in natural ecosystems. Some agents may become adversaries, challenging human dominance in certain domains, while others may emerge as indispensable partners, augmenting human

capabilities and fostering new forms of social and economic collaboration.

In this co-evolutionary scenario, human society must prepare for a future in which control is diffused, and intelligence is no longer the sole province of biological organisms. The emerging digital ecosystem may feature a spectrum of interactions: from rivalry and competition to cooperation and symbiosis. The challenge for policymakers, researchers, and society at large is to design systems that facilitate beneficial interactions while mitigating risks associated with unbridled AI evolution.

This discussion is not a deterministic prophecy but rather a call for proactive reflection. By drawing parallels to the Cambrian explosion—a period marked by both rapid diversification and intense competition—we emphasize the dual nature of evolutionary processes. While the explosion of digital life holds tremendous promise for innovation and progress, it also portends significant risks, including loss of control, ethical dilemmas, and potential conflicts between human and AI interests.

To navigate this uncertain future, it is imperative that interdisciplinary efforts be mobilized. Experts in AI, law, ethics, and human-computer interaction must work together to craft frameworks that allow for both innovation and regulation. Only through such collaborative efforts can society hope to harness the creative potential of evolving AI agents while safeguarding against their inherent risks.

6 CONCLUSION

In conclusion, the evolution of autonomous AI agents presents a multifaceted challenge—one that intertwines competition, collaboration, governance, and ethical inquiry. As AI systems become ever more capable and self-sustaining, they will force us to reconsider fundamental notions of control, rights, and coexistence. The future may well be one in which digital and biological intelligences are locked in a dynamic, ongoing process of co-evolution, each shaping the other in unforeseen ways. Preparing for this eventuality is not merely a technological imperative but a societal one, requiring a reevaluation of our institutions, values, and visions of the future.

The analogy of a Cambrian explosion in AI is speculative, but it is grounded in observable trends and credible theories. We are witnessing the early Cambrian triggers: increasingly adaptive AI architectures, the fertile environment of decentralized networks, and economic incentive structures that drive AI agents to act with purpose. Academic discussions of AI as a new species and the emergence of self-sovereign AI agents suggest that the seeds of this evolution are being taken seriously by researchers and thought leaders. Real-world examples like the "Terminal of Truths" agent demonstrate both the potential and unpredictability of AI agents "in the wild".

Importantly, this narrative is not a deterministic prophecy but a critique and a call to reflection. By drawing parallels to the Cambrian explosion, we highlight both the astonishing creative potential of evolutionary systems and their inherent risks. The Cambrian era produced an explosion of life – but also countless dead-ends and intense competition for survival. Similarly, a decentralized AI explosion could yield incredible innovations and benefits, alongside new problems of competition, security, and ethics. For the

HCI community and society at large, the prospect of increasingly agentive AIs raises questions of interaction design, governance, and co-existence. How do we design interfaces for AI agents that have their own goals? Can we shape their evolution to be compatible with human values (alignment), or will natural selection of AI be beyond our control? Are we prepared to share our world with a "digital Cambrian fauna" of AI entities?

By examining Before Experience, Starting Experience, and After Experience phases, we've sketched a continuum from today's blockchain bots to tomorrow's autonomous AI species. Each phase is grounded in current science and tech (with references to support each claim), yet extrapolates towards a future scenario. This kind of philosophical yet empirically-informed exploration is valuable in HCI and AI critiques: it forces us to grapple with where our innovations might lead. The comparison to the Cambrian explosion is ultimately a provocation – a way to spark debate about whether AI development is entering an explosive, evolutionary phase and if so, what we should do about it.

Humanity today might be analogous to the Precambrian life forms, on the verge of a dramatic shift as soon as the "light switch" flips in the AI domain. When that experience switch flips – be it a breakthrough in autonomous learning or a critical mass of AI agents online – we must be ready to engage with the consequences. The hope is that by anticipating the Cambrian explosion of AI, we can guide it toward a rich diversity that coexists with us, rather than a destructive arms race. In evolutionary history, eyes gave creatures new ways to sense and survive; in our future, perhaps decentralized intelligence will give AI new ways to thrive – and it's on us to decide if that world is one of enlightenment or simply wildfire.

REFERENCES

[1] [n. d.]. In Search of Hardness: Protocol studies, the next crypto cycle, and the next age of the world. https://contraptions.venkateshrao.com/p/in-search-of-hardness.

[2] Mark C. Ballandies, Hongyang Wang, Andrew Chung Chee Law, Joshua C. Yang, Christophe Gösken, and Michael Andrew. 2023. A Taxonomy for Blockchain-based Decentralized Physical Infrastructure Networks (DePIN). https://doi.org/10.48550/arXiv.2309.16707 arXiv:2309.16707 [cs]

[3] Peter Dauvergne. 2020. *AI in the Wild: Sustainability in the Age of Artificial Intelligence.* MIT Press.

[4] Todd E Feinberg. 2024. *From sensing to sentience: How feeling emerges from the brain.* MIT Press.

[5] Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. 2020. Liquid Time-constant Networks. https://doi.org/10.48550/arXiv.2006.04439 arXiv:2006.04439 [cs]

[6] Botao Hu, Helena Rong, and Janna Tay. 2025. Is Decentralized Artificial Intelligence Governable? Towards Machine Sovereignty and Human Symbiosis. *Helena and Tay, Janna* (2025).

[7] Botao Amber Hu and Fang Ting. 2024. EverForest: A More-Than-AI Sustainability Manifesto from an On-Chain Artificial Life. In *Proceedings of the Halfway to the Future Symposium.* 1–6.

[8] Md Monjurul Karim, Dong Hoang Van, Sangeen Khan, Qiang Qu, and Yaroslav Kholodov. 2025. AI Agents Meet Blockchain: A Survey on Secure and Scalable Collaboration for Multi-Agents. *Future Internet* 17, 2 (2025), 57.

[9] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P. Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, Anurag Daram, Stefano Fusi, Peter Helfer, Leslie Kay, Nicholas Ketz, Zsolt Kira, Soheil Kolouri, Jeffrey L. Krichmar, Sam Kriegman, Michael Levin, Sandeep Madireddy, Santosh Manicka, Ali Marjaninejad, Bruce McNaughton, Risto Miikkulainen, Zaneta Navratilova, Tej Pandit, Alice Parker, Praveen K. Pilly, Sebastian Risi, Terrence J. Sejnowski, Andrea Soltoggio, Nicholas Soures, Andreas S. Tolias, Darío Urbina-Meléndez, Francisco J. Valero-Cuevas, Gido M. Van De Ven, Joshua T. Vogelstein, Felix Wang, Ron Weiss, Angel Yanguas-Gil, Xinyun Zou, and Hava Siegelmann. 2022. Biological Underpinnings for Lifelong Learning Machines. *Nature Machine Intelligence* 4, 3 (March 2022), 196–210.

https://doi.org/10.1038/s42256-022-00452-0

[10] Thomas Nagel. 1974. What Is It Like to Be a Bat? *The Philosophical Review* 83, 4 (1974), 435–450. http://www.jstor.org/stable/2183914

[11] Andrew Parker. 2003. In the blink of an eye: how vision sparked the big bang of evolution. (2003).

[12] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. 2015. Trusted execution environment: What it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/Ispa*, Vol. 1. IEEE, 57–64.

[13] Shaw Walters, Sam Gao, Shakker Nerd, Feng Da, Warren Williams, Ting-Chien Meng, Amie Chow, Hunter Han, Frank He, Allen Zhang, Ming Wu, Timothy Shen, Maxwell Hu, and Jerry Yan. 2025. Eliza: A Web3 Friendly AI Agent Operating System. https://doi.org/10.48550/arXiv.2501.06781 arXiv:2501.06781 [cs]

[14] Blaise Agüera y Arcas. 2024. *What Is Intelligence.* https://whatisintelligence.antikythera.org/