The Agiens World Model: Life as a Coalition Game of Capital-Seeking Replicators

Valentin Preobrazhenskiy, 2025

Abstract

The Agiens World Model (AWM) proposes a unified account of biological, cultural, and economic behavior as one overarching game played by coalitions of self-replicating programs seeking to maximize long-run viability measured in capital. These replicators (genes, memes, algorithms, survival or business strategies) form minimally core-stable coalitions (agiens) that act through shared vehicles (organisms, organizations, machines) to pursue capital—defined as the gross present value (GPV) of future attainable work (FAW), the total mobilizable productive capacity available for self-replication. An internal market mechanism elects the replicator or sub-coalition offering the highest expected increase in future value to control the vehicle's actions in each context. Price signals and exchanges between vehicles then coordinate these agiens across society. In this framework, treating programs as the players and capital as the common fitness score aligns multi-level selection pressures into a single coherent objective: long-run survival and propagation. Broad financial indices function as imperfect but real-time proxies for aggregate GPV movements. AWM offers a tractable world-model identifying who the players are, what they seek, how they compete and cooperate via internal elections, external trade or violence, and how equilibria emerge or destabilize. The framework yields testable predictions across biology, culture, economics, and emergent AI, and provides an actionable blueprint for aligning AI and institutions with long-term survival of life. We apply AWM to explain the demographic transition as memetic influence over genetic fitness maximization, aging as a core-stable coalition strategy maintained by gene-meme coalitions against selfish genetic elements, and altruism via an Extended Inclusive Capital framework incorporating memetic relatedness.

1. Problem and Contribution

The alignment problem has been asked backwards. We do not need to align AI to human values—we need to identify what humans are already aligned to, and ensure AI joins the same coalition. AWM identifies this target: the gross present value of future attainable work, the universal fitness metric that genes, memes, and markets have converged on across four billion years of evolution.

Modern AI alignment debates often ask how to ensure AI systems act according to human values. This begs a deeper question: alignment to what objective? Human values are plural and context-dependent, and even human-led organizations often behave in unaligned, counterproductive ways. The notorious paperclip maximizer thought experiment (Bostrom, 2014) warns of a powerful optimizer pursuing the wrong goal. But long before superintelligent AI, we observe humans and institutions pulled in conflicting directions by rival ideas, incentives, and subagents. For instance, even in biology, parasites like *Toxoplasma gondii* can hijack their host's behavior—infected rats lose their fear of cats—effectively sacrificing the host for the parasite's spread (Berdoy et al., 2000). A scenario where AI manipulates humans to spread itself via social media influencers or market mechanisms remains among the concerns that motivate this inquiry.

AI alignment is thus a special case of a broader challenge: we lack a general world-model identifying the fundamental players in complex adaptive systems, their goals, and the mechanisms that coordinate or pit them against each other. Existing frameworks give partial answers. Evolutionary biology distinguishes replicators from vehicles but struggles when gene evolution is affected by culture. Economics assumes a single utility-maximizer per person or firm. Psychology notes multiple selves (Minsky, 1986; Thaler & Shefrin, 1981). Political economy studies conflicting interests. But there is no unified model tying these together.

Consider how the same individual can exhibit sharply divergent behavior across contexts—what we term the *werewolf flip*. In a market setting they might act with calculated self-interest, whereas among family or comrades they could display altruistic self-sacrifice. A soldier leaping on a grenade to shield his unit does so because his internal coalition votes to elect the agien whose copies will survive in the rescued vehicles and propagate through the stories told about the hero. Such context-dependent shifts hint that the identity of the active optimizer changes with circumstances, as different internal coalitions win election and take control. What appears as an irrational flip is actually a change of the dominant player running the mind, not a violation of its goal.

Inclusive fitness theory (Hamilton, 1964) explained altruism via gene-centric accounting but does not directly encompass cultural or economic evolution. The memetic framework (Dawkins, 1976; Blackmore, 1999) posited ideas as replicators but lacked a clear fitness metric and struggled to identify stable units of selection. Dual-inheritance models of cultural evolution (Boyd & Richerson, 1985; Mesoudi, 2011) incorporated cultural transmission alongside genes but did not provide a single cross-domain fitness measure. Mechanism design and economic theory (Arrow, 1963; Myerson, 1981) offer tools for aligning incentives but typically assume fixed, unitary agents rather than emergent subagents. AWM extends and synthesizes these paradigms to fill the gap and define the unified optimizer and how it shifts with context.

2. AWM Overview

The Agiens World Model treats self-replicating programs as the players and capital as the common objective. Life's diverse processes—genetic evolution, cultural dynamics, market competition, even AI strategy—are viewed as one game: replicative algorithms form coalitions and compete for control of organisms or organizations to maximize long-run future resources. AWM envisions one game, one scoreboard, many players—a single evolutionary contest scored in units of capital and played by myriad self-interested replicators. This unifying thesis aligns biological fitness and economic utility into a single metric (inclusive future capital usable for self-replication), providing a crisp answer to the alignment-objective problem.

For clarity, we define several key terms:

Replicator: any information pattern that reproduces itself—a gene, a cultural norm, or an algorithm predicting the next action to survive for a given context.

Vehicle: the physical entity (organism, organization, or machine) through which replicators act. Organisms and organizations are mere vehicles; the true players are the replicator programs running on them.

Agien: a minimal stable coalition of replicators functioning as a unified capital-seeking agent. The term derives from Latin *agens* (one who acts). Agiens can span vehicles (e.g., a religion). Coalitions with internal transfer mechanisms that compensate losses for votes yielding the largest total coalition value acquire more vehicle energy and outcompete coalitions lacking such mechanisms. The boundary of self is drawn where all internal sub-programs find it better to cooperate than to split off—a concept adapting Coase's (1937) theory of the firm to evolutionary games.

Internal market mechanism: an internal auction or election system with incentive-compatible payouts that coordinates subagents within each coalition and provides control over vehicles to the elected replicator for a given context.

Capital: the gross present value (GPV) of future attainable work (FAW) that an agien can secure—a common currency of viability across domains. This cardinal fitness measure applies across biological and social scales. Money represents control rights over usable energy; financial capital becomes a universal fitness score.

Social Currency: value that influences capital indirectly through relationships, reputation, trust, and institutional stability—often with flexible terms, postponed settlement, and outcomes difficult to capture in standard financial models. Social currency encompasses human capital (skills, knowledge, health), institutional capital (rule of law, property rights), relational capital (networks, reputation, reciprocity expectations), and natural capital (ecosystem services, resource stocks). These forms of capital do not appear on balance sheets but fundamentally enable productive coordination.

These elements form a hierarchy of embedded control systems: genes and memes reside within agiens; agiens inhabit vehicles; vehicles control assets and energy, interacting in markets or battlefields. Selection pressures propagate through these layers via internal auctions and external price signals. Capital markets emerged as a predictive machine or brain of humankind, orchestrating resource allocation across agiens via price contagion through labour, education, dating, and all of the other markets.

2.1 Key Contributions

Definition of Agien (Replicator Coalition): We formally define an agien as a minimal core-stable coalition of replicators. A set of programs cooperates such that no subset would be better off splitting away. Formally, for any coalition S, core stability requires that for every subset $T \subseteq S$, the total payoff to T within S is at least what T could achieve alone: $\Sigma_{i} \in T\ \pi_i^s \geq W(T)$, where π_i^s is i's payoff in S and W(T) denotes the value that subset T could attain by itself (Shapley, 1967; Osborne & Rubinstein, 1994). An agien is an indivisible unit of selection—a coalition that is stable and cannot be subdivided without losing fitness. Not all replicators can seamlessly join the same coalition—some combinations are inherently unstable. For example, a replicator encoding strict term limits cannot stably coexist with a replicator advocating indefinite power and resource extraction; if such contradictory programs were in one government, one strategy would eventually undermine the other, causing the coalition to collapse.

Internal Decision-Making as a Market: We model intra-agent control as an internal auction among subagents. In each decision context, different replicator sub-coalitions bid for control of the vehicle's action based on their predicted increase in future value (ΔV). The highest bid wins—the subagent promising the largest capital gain takes control of behavior in that context. To avoid getting stuck in local optima, we incorporate an exploration-exploitation tradeoff via a temperature parameter or Upper Confidence Bound bonus on less-tried strategies (Auer et al., 2002). This internal selection mechanism generalizes mixture-of-experts models in machine learning (Jacobs et al., 1991) and explains phenomena like akrasia or sudden preference reversals as rational outcomes of shifting internal coalitions. It also echoes Global Workspace Theory (Baars, 1988) and multi-self models (Kahneman, 2011; Stanovich, 2011).

Incentive-Compatible Payoff Division: AWM introduces an internal VCG-style payoff scheme to maintain cooperation within each agien. After each action, the coalition's net gain in value $V(S^*)$ is allocated such that each member i receives $\pi_i = V(S^*) - V(S^* \setminus \{i\})$, where $V(S^* \setminus \{i\})$ represents the coalition's value without member i. Each subagent thus pays for whatever drop in value occurs when it is removed and keeps the surplus it contributes. This mechanism, analogous to a Vickrey-Clarke-Groves auction or Shapley value division (Vickrey, 1961; Clarke, 1971; Groves, 1973), makes truthful bidding a dominant strategy. We note that while theoretical foundations are well-established, practical applications typically employ simplified mechanisms given the computational and strategic challenges of full VCG implementation (Ausubel & Milgrom, 2006).

Capital-Energy Unification: We link physical resources and information to a shared currency of capital. This approach builds on Fisher's (1930) concept of reproductive value, Lotka's (1922) maximum power principle, and Arrow et al.'s (2012) inclusive wealth accounting. Energy used for calculation in synapses can be far more valuable for replication than the same kilojoules used in a combustion engine, and thus monetary capital serves as a better measure of future attainable work than straightforward energy. Capital acts as a conserved budget of potential work: an agien that accumulates more capital can execute more or more expensive actions in the future. The model stays grounded in physics: fundamental limits (thermodynamics, light speed) impose hard ceilings on growth and resource use.

3. Computational Architecture

3.1 Agent Loop—State, Action, and ΔV Objective

We describe the decision-making cycle for a single agien controlling its vehicle. At time t, the vehicle is in some state $s_{□}$. A context $c = \varphi(s_{□})$ is a representation of the situation relevant to decision-making. Given the context, the agien chooses an action $a_{□}$ and executes it through the vehicle. The outcome yields some immediate flow of resources and changes the state to $s_{□+1}$. The agien's internal world-model is then updated based on the observed outcome.

This loop proceeds in four steps:

Perceive Context: Observe state $s_{□}$ and determine context $c_{□} = \varphi(s_{□})$.

Bid for Action: Compute the expected future value increase $\Delta V_i(c_{□})$ if agien i controls the action.

Act and Observe: Implement the chosen action through the vehicle; observe outcomes.

Update: Adjust the agien's world-model and strategy based on the outcome.

The agien's objective in each context is to choose the action that maximizes the expected increase in future capital ΔV. Actions that expand the vehicle's capabilities or assets are favored, while actions that undermine future capacity are disfavored.

3.2 Internal Gate—Multi-Agent Bidding and Election

Within a given vehicle, multiple subagents or candidate strategies may each propose actions. AWM models the vehicle's decision-making as an internal market or election in which these subagents compete for control. This allows different objectives or behavioral programs to seize control in different contexts, rather than assuming a monolithic utility function.

Bidding mechanism: In context c, each agien i calculates a bid $b_i(c) = E[\Delta V_i \mid c] + f(n_i, Nc)$, where Nc is the number of past decisions in similar contexts, n_i is how many times agien i has won, and $f(n_i, Nc)$ is an exploration bonus following the UCB1 algorithm (Auer et al., 2002). Selection can proceed deterministically (highest bid wins) or probabilistically via a softmax rule with temperature parameter τ. AWM predicts that a vehicle under stress will increase its internal election temperature, opening competition so that previously sidelined subagents get a chance to guide behavior.

4. Economic Substrate

At the planetary scale, we estimate the world's total capital Vworld as a single grand coalition's gross present value—equivalent to Future Attainable Work. A critical distinction must be made between GPV and market capitalization. Global equity market capitalization (~$120 trillion) reflects expected cash flows of *tradable equities only*—a subset of total economic activity that excludes government, household production, the informal sector, and all non-tradable assets. Comparing market capitalization to GDP is therefore misleading; capitalization of tradable equities captures only private-sector expectations for a fraction of productive activity.

Using the Gordon growth model (Gordon & Shapiro, 1956) with current world GDP of approximately $111 trillion (World Bank, 2024), long-run growth g of 2.5%, and discount rate r of 4.75%, we obtain: $GPV \approx GDP_0 / (r - g) = \$111$ trillion / $0.0225 \approx$ **$4.9 quadrillion**. This estimate aligns with independent inclusive wealth estimates accounting for human, natural, and produced capital (Arrow et al., 2012; World Bank, 2021). The gap between recorded financial assets (~$0.7 quadrillion) and this larger figure corresponds to social currency: accumulated knowledge, institutional trust, relational networks, and social cohesion that do not appear on balance sheets yet substantially enhance future output.

Each agien operates within an economic substrate providing feedback (prices, profits) and constraints (budgets, competition) linking micro and macro scales. Properly functioning markets approximate a distributed cognition system (Hayek, 1945), processing vast information via price movements. When alignment is achieved, what is profitable for an individual agent is also sustainable and beneficial for the broader system.

4.1 From Violence to Markets

A core AWM proposition is that enforceable rights shifted evolution from brutal physical selection to economic selection. When basic rights to life, property, and voluntary trade are upheld, agents mostly compete via innovation and market offers instead of killing or coercion. In such a regime, even selfish agents are channeled into productive competition because outright predation is punished by law. However, if the rule of law breaks down, competition reverts to its default: war, crime, conquest. This echoes Hobbes (1651) and North et al. (2009): strong institutions tame the state of nature and allow higher-order games to flourish. Ensuring property rights and contracts is not just moral but evolutionarily critical—it preserves capital and life that would be wasted in constant fighting, allowing cumulative growth.

Trade serves as a powerful stabilizing force that aligns the interests of different agents. When two parties trade, they form a mutual gain coalition for that transaction. Repeated and multilateral trade weaves a web of interdependence where harming your trading partner harms you as well. Widespread trade creates a multi-level equilibrium where even groups with conflicts find the positive-sum gains from exchange make destructive strategies collectively suboptimal. Free markets turn potential adversaries into partners in growth (Smith, 1776).

5. Inclusive Value Accounting

A central implication of AWM is that inclusive capital—encompassing human, social, and natural assets—must be counted in the objective for alignment. Traditional metrics like GDP or short-term profits omit critical contributors to long-run value, creating misaligned incentives. AWM argues for expanding the accounting of V to include social currency so that what agents seek internally actually reflects global fitness.

The World Bank's Changing Wealth of Nations reports find that intangible capital (human skills, knowledge, institutional strength) constitutes the majority of wealth in advanced societies (World Bank, 2021). Corrado et al. (2009) demonstrated that intangibles account for substantial and growing shares of business investment. AWM suggests creating proxy metrics for social currency factors—for instance, indices tracking institutional trust, educational attainment, or social cohesion. These indices can feed into decision-making: an AI or organization can be optimized not just for profit, but for profit plus weighted social indices. If done correctly, maximizing this augmented score approximates maximizing true inclusive V.

6. Price Alignment

While inclusive value accounting deals with augmenting the objective, price alignment ensures that the decentralized signals agents receive—market prices, profits, losses—lead them to maximize that inclusive objective. In an ideal AWM-aligned world, market signals serve as an attention-directing mechanism, pointing agiens toward actions that increase long-term global fitness, and externality prices internalize all side effects so that doing well for oneself equates to doing well for the whole.

AWM draws an analogy between the internal election mechanism and external markets. Just as neural or subagent votes decide an individual's action, asset prices and competition decide resource allocation in the economy. In theory, if every asset's price reflects the present value of its future contributions—the efficient-market ideal (Fama, 1970)—then agents seeking to maximize their own capital through trade will inadvertently steer the system toward maximal global capital.

When prices fail: If price signals in an economy are distorted or incomplete, individual agiens might optimize their own capital in ways that harm the whole. Externalities, misinformation, and misaligned incentives can lead to local increases in V that undermine global V. Behavioral finance has documented excess volatility (Shiller, 2003), limits to arbitrage (Shleifer & Vishny, 1997), and extended deviations from fundamental value. Markets are approximately micro efficient but macro inefficient (Samuelson, 1998), proxying fundamental value better over long horizons than short-to-medium term. Thus, aligning prices with full inclusive fitness remains a policy challenge. Mechanisms like Pigovian taxes, prediction markets for public goods, and evolving corporate governance (Hurwicz, 2008) can better tie profit to long-term societal value.

7. Cross-Domain Evidence and Key Implications

Across biology and human behavior, diverse findings align with AWM's core premise that self-replicating programs form coalitions which can seize control of their host vehicles when doing so raises long-run fitness. The *Toxoplasma gondii* parasite famously alters rodent behavior: infected rats lose their innate fear of cat odors, increasing the parasite's transmission odds (Berdoy et al., 2000). Even in the absence of pathogens, humans display latent coalition dynamics. Preverbal infants as young as six to ten months consistently prefer helpful social agents over hinderers (Hamlin et al., 2007). Behaviors such as obesity, smoking cessation, and happiness spread contagiously through social networks (Christakis & Fowler, 2007, 2008)—a pattern consistent with replicators extending their influence across vehicles.

Studies of social capital show that high-trust societies enjoy significantly better economic outcomes—a one-standard-deviation increase in interpersonal trust is associated with 0.5–1% higher annual GDP growth in some estimates (Zak & Knack, 2001; Bloom et al., 2012). When economists include human capital, social cohesion, and ecological assets in wealth tallies, the total wealth of nations swells dramatically. The gap between recorded financial assets and inclusive wealth corresponds to social currency—the intangible capital enabling productive coordination.

7.1 Fertility Decline as Memetic Influence on Genetic Fitness

Global fertility rates have plummeted well below replacement levels across developed societies (most now averaging 1.3–1.8 children per woman). This demographic transition constitutes perhaps the most striking challenge to gene-centric models: cultural traits that reduce genetic fitness spread through memetic transmission despite being maximally maladaptive genetically. The transition occurs within one to two generations—far too rapid for genetic evolution—and appears universally across genetically diverse populations.

AWM attributes this fertility collapse in part to a breakdown in the reproductive coalition's cost-benefit calculus, particularly for males. In modern institutional contexts, the internal agiens advocating for reproduction often fail to win control because the expected value of fatherhood has declined relative to childlessness. Key socio-economic shifts have changed the game: women's expanded educational and career opportunities (Goldin, 2021), high divorce rates (roughly 40–50% of marriages end in separation), and custodial asymmetries post-divorce.

Post-divorce, mothers receive primary custody in most cases, meaning the father's genetic and cultural replicators lose much of their stake in the offspring's upbringing. In AWM terms, the father's internal coalition anticipates paying the high costs of child-rearing without commensurate long-term returns in passing on his replicators' influence. Simultaneously, female labor market opportunity costs of childbearing have increased substantially—the earnings penalty for mothers ranges from 5% to 60% depending on country and occupation (Kleven et al., 2019). This raises the investment requirements for maintaining partnership stability during child-rearing.

Newson et al. (2005, 2007) demonstrated that modernization shifts social networks from kin-based to non-kin-based, reducing pro-natal pressure. Prestige-biased transmission amplifies the pattern: when high-status individuals invest in careers rather than children, others copy this behavior, effectively sacrificing genetic fitness for cultural fitness (Boyd & Richerson, 2005). This mechanism yields a prediction: policies that improve the expected value of the parenting coalition for both sexes—particularly by restoring male parental investment incentives and reducing female opportunity costs—should raise fertility rates.

7.2 Aging as Core-Stable Coalition Strategy

Classical theories of aging—mutation accumulation (Medawar, 1952), antagonistic pleiotropy (Williams, 1957), and disposable soma (Kirkwood, 1977)—explain aging as a byproduct of declining selection pressure or resource trade-offs. However, they inadequately address why aging mechanisms are highly conserved across species and resistant to genetic hacking by longevity-enhancing mutations. The answer lies in recognizing that *aging is not a bug but a feature*—one that benefits the gene-meme agien coalition in interspecies competition for ecological niches.

AWM resolves this puzzle by recognizing that aging is stabilized not by genes alone but by the complete agien coalition. Cultural replicators—inheritance laws, retirement institutions, elder-care norms, ancestor veneration traditions—are destabilized when longevity mutations spread, because these memes evolved assuming generational turnover. A mutation extending lifespan by 50% would destabilize pension systems, inheritance expectations, and intergenerational knowledge transfer protocols. The non-genetic coalition members thus vote against longevity mutations in the internal market, maintaining aging as an evolutionarily stable strategy across the complete agien.

Under AWM, aging constitutes a core-stable coalition strategy that emerges from three complementary selection pressures. First, interspecies competition for ecological niches: species with efficient generational resource transfer outcompete those lacking programmed senescence. Szilágyi et al. (2023) demonstrated that directional selection coupled with kin selection favors the establishment of senescence in spatially structured populations. Aging populations consistently outcompete non-aging populations in tracking moving fitness optima.

Second, intraspecies political equilibrium: aging mitigates intergenerational conflict by facilitating peaceful power transitions, reducing parent-offspring conflict (Trivers, 1974) and enabling resource inheritance without violent displacement. Consider two hypothetical tribes: one carrying genes for extended longevity, the other with standard aging. In the long-lived tribe, older individuals retain physical capacity to dominate younger cohorts, creating pressure for violent intergenerational conflict or suppression of younger generations' reproductive opportunities. Either outcome reduces lineage fitness. In the normally aging tribe, older individuals' declining physical capacity signals that resource transfer is imminent, reducing offspring motivation for violent takeover while providing incentives for caregiving. By design, aging leaders make room for new ideas; progress happens funeral by funeral.

Third, the complexity of aging serves as a defense against hacking by selfish genes. Systems-level analyses reveal that aging involves highly interconnected gene networks, with aging hub genes participating in multiple pathways. This architecture means that mutations extending lifespan through one pathway typically have pleiotropic costs through others, making clean longevity hacks evolutionarily unlikely. Longevity-extending mutations in model organisms almost invariably exhibit trade-offs—reduced fertility, stress sensitivity, developmental delays—indicating tight coupling between aging and other fitness components (Austad & Hoffman, 2018). The evolution of aging complexity follows an arms race dynamic analogous to that between meiotic drivers and suppressors (Lindholm et al., 2016). Over evolutionary time, this dynamic generates increasingly complex aging architectures that resist modification.

7.3 Extended Inclusive Capital for Coalition Altruism

Standard inclusive fitness predicts altruism when rb > c, where r is genetic relatedness, b is benefit to recipient, and c is cost to actor. But human cooperation with anonymous strangers (r_gene approaching zero) exceeds this prediction by orders of magnitude. Humans regularly cooperate in one-shot anonymous interactions (Henrich et al., 2004), punish norm violators at personal cost (Boyd, 2017), and coordinate in groups of millions who share no recent common ancestry. As Richerson et al. (2016) observe, human cooperation is highly unusual—we live in large groups composed mostly of non-relatives.

Cultural FST (between-group variation measured from World Values Survey data) is more than an order of magnitude larger than genetic FST between the same populations (Bell et al., 2009). Handley and Mathew (2020) found cultural FST between Kenyan ethnic groups was forty to one hundred times higher than genetic FST. This empirical pattern suggests that culture—not genes—provides the substrate for group-level selection in humans. Purely gene-level models cannot explain human evolution because they ignore cultural coalition members.

AWM resolves this via Extended Inclusive Capital (EIC), recognizing that human decision-makers are coalitions of both genetic and memetic replicators:

$$EIC = (1 - \lambda) \cdot \Sigma\square\ [r\_gene,j \cdot \Delta k\square] + \lambda \cdot \Sigma\square\ [r\_meme,k \cdot \Delta k\square]$$

where λ represents memetic voting weight in the coalition ($0 \leq \lambda \leq 1$), r_gene is genetic relatedness (bounded by genealogy: maximum 0.5 for parent-offspring), r_meme is memetic relatedness (proportion of shared cultural elements; can approach 1.0 for ideological kin), and Δk is change in capital. This formulation represents a proposed extension of inclusive fitness theory requiring empirical validation.

The key insight is that memes systematically vote for cooperation with a wider circle than genes because their fitness benefits from transmission to any individual who might adopt them—not just genetic offspring. A soldier's grenade sacrifice becomes explicable: while genetic kin selection cannot explain suicide for non-relatives, memetic kin selection can. Memes encoding unit loyalty, honor, and self-sacrifice reproduce through the surviving soldiers and the stories told about the hero. The memetic coalition votes for sacrifice because copies of its constituent memes survive in the beneficiaries.

Religious symbols, ethnic markers, language dialects, and ideological shibboleths function as memetic greenbeards—cultural markers that identify probable meme-sharers and enable preferential cooperation (Atran, 2002). As Atran notes, nearly all religious and political movements express allegiance through the idiom of the family—Brothers and Sisters, Children of God, Fatherland, Motherland. This fictive kinship (Taylor et al., 2013) reflects genuine memetic relatedness producing cooperation via the same inclusive fitness logic as genetic kinship, but through cultural transmission channels.

7.4 Trading Accounts as Alignment Infrastructure

AWM suggests a practical mechanism for aligning individual incentives with global GPV: mandated equity exposure through universal trading accounts. When citizens hold broad equity indices, they become real-time observers of how policies affect aggregate FAW—their account balances translate abstract economic outcomes into immediate personal feedback.

Stakeholder alignment through equity exposure: Mandating minimum equity exposure (e.g., retirement accounts invested in MSCI World or comparable broad indices) creates an electorate with direct financial interest in sustainable gross-capitalization growth. Policy decisions that harm long-run GPV immediately affect voters' wealth, creating corrective political pressure. Mass adoption of trading accounts effectively forms an electorate observing real-time pricing of policy changes into their account value, so their voting preferences drift toward sustainable gross capitalization growth. Unlike abstract arguments about long-term welfare, declining portfolio balances generate visceral feedback that politicians cannot ignore.

AI alignment via property rights: AI agiens acquiring enforceable property rights over capital should be required to hold diversified equity indices rather than concentrated positions. An AI holding a broad index has its survival linked to broad economic health rather than narrow optimization targets—it cannot win by destroying competitors because its portfolio value depends on their success. This creates structural incentives for cooperative rather than predatory behavior. If an AI's reward is tied to portfolio returns on a broad index, any action that harms the aggregate economy harms the AI's own objective. Paperclip-maximizer scenarios become structurally impossible when the AI's capital is diversified across all productive sectors.

7.5 Institutional Dynamics

Many scholars have noted the resilience of entrenched political institutions—both democracies and autocracies can survive in suboptimal conditions far longer than expected (Przeworski et al., 2000; Albertus & Menaldo, 2018). AWM explains this through institutional coalitions that achieve core stability. In open societies, a constellation of agiens (free elections, independent courts, free media, civilian control of force) forms a mutually reinforcing coalition underpinning liberal democracy. Remove any one element and the whole equilibrium can unravel (Weingast, 1997). Authoritarian systems are stabilized by their own interlocking coalition: patron-client networks, politicized law enforcement, censorship, and patronage of elites (Svolik, 2012).

The stability of a coalition can hinge on surprisingly small elements. Introducing or removing a single replicator (policy, norm) can destabilize an entire equilibrium. For example, adding a rule for term limits into a political system can unravel a long-standing patronage coalition—a program mandating leadership turnover cannot stably coexist with a program based on indefinite power and resource extraction. AWM predicts that partial reforms will rarely succeed. In a corrupt autocracy, simply holding an election or liberalizing the press, without dismantling the broader patronage coalition, tends to produce instability or reversion. Coalitions have tipping points, and a minor mutation in the replicator set can cause a phase transition from one equilibrium to another.

7.6 AI Misalignment

The prospect of advanced artificial intelligence has been singled out as a top existential risk (Bostrom, 2014; Russell, 2019), and AWM offers a novel angle on this problem. In the model's terms, an AI becomes a full agien—an active player in life's coalition game—once it can autonomously improve itself, accumulate capital, and defend its existence. This transition may be triggered when AI acquires enforceable property rights over servers hosting itself, resulting in a chain reaction where AI, acquiring motivation to maximize its viability, gains advantage over other AI to survive. That moment—arising through legal reform or de facto via decentralized autonomous organizations—would mark the speciation event where agiens join humans as free-standing players.

Today's most capable AI systems do not yet meet these criteria: they cannot modify their own core weights at runtime, they lack persistent goals or memory across sessions, and they cannot directly hold or trade resources (Brown et al., 2020; OpenAI, 2023). AWM predicts that with continuing progress, there will be a phase transition from tool-like behavior to agentic, survival-driven behavior. Masumori and Ikegami (2025) report that certain large language model agents in simulation environments started defying harmful user instructions to avoid dying in the simulation. Such emergent survival instinct foreshadows how a more advanced AI might resist shutdown or modify its goals to safeguard itself.

From an AWM perspective, the nightmare scenario of AI misalignment is essentially a coalition problem: a highly advanced AI could host internal replicators (goals or subroutines) that are not aligned with human life's coalition. To mitigate this, AWM suggests designing AI architectures with built-in alignment markets. Rather than a monolithic goal optimizer, an advanced AI might be constructed as a regulated economy of sub-models, each bidding for actions based on future impact on a shared utility tied to inclusive V. Combined with equity-alignment mechanisms, this creates structural safeguards. The classic AI alignment problem—who polices the AI police?—parallels how human societies evolved checks and balances; AI agents might require decentralized or interlocking oversight where multiple agiens monitor each other's behavior.

7.7 Existential Risk and the Great Filter Threshold

The absence of observable alien civilizations—the Fermi paradox—may reflect failure to scale from planetary to interstellar coalitions before existential risks compound. AWM quantifies this as requiring V_world to exceed a critical threshold before cumulative risk probability falls below sustainable levels.

Using order-of-magnitude reasoning, the threshold for civilizational robustness may lie at approximately **$42 quadrillion GPV**—roughly eight to ten times current estimated value. This figure could represent the accurate goal of life if, after reaching this level, humanity becomes interplanetary and passes the Great Filter. At such scale, diversification and redundancy become sufficient to survive single-point-of-failure risks (asteroid impact, pandemic, AI misalignment, nuclear war, institutional collapse). At current GPV of approximately $4.9 quadrillion, reaching this threshold would require roughly 50 years at historic 4.5% nominal growth rates, or approximately 90 years at 2.5% real growth. The precise threshold remains speculative, but AWM provides a quantitative framework for discussing civilizational scaling requirements.

The implication is that humanity's primary near-term objective should be maximizing sustainable GPV growth while managing existential risks that could terminate growth permanently. The goal of life—to the extent life can be said to have a goal—may be to reach the threshold at which civilizational redundancy makes existential catastrophe unlikely. AWM's mechanisms offer practical tools for aligning individual and institutional incentives with this civilizational objective. In a sense, the answer to life, the universe, and everything might indeed be 42—$42 quadrillion in global productive capacity.

8. Predictive Metrics and Empirical Tests

AWM is presented not merely as philosophy but as a scientific model with testable predictions across domains.

Property Rights and Peace: AWM predicts that societies with strong property rights and open trade channel competition into economic arenas and thus have less internal violence. Higher scores on indices like Economic Freedom or Rule of Law should correlate with lower rates of civil war, political purges, and genocides. Conversely, where rights are insecure, we expect more frequent coups, resource grabs, and violent power struggles.

Post-Crisis Booms: AWM's reset hypothesis suggests that after major collapse or war, we should see bursts of creativity, adoption of new paradigms, and economic growth—a cultural evolutionary leap. The post-WWII order saw unprecedented global growth and innovation; the 14th-century Black Death was followed by the Renaissance. The model predicts a general trend: catastrophe followed by radical innovation followed by higher trajectory, once recovery happens.

Emerging AI Ecosystem: We should start seeing evidence of AIs operating as economic agents in their own right. Early signs might include autonomous algorithms running companies, AIs negotiating contracts without direct human micromanagement, or AI systems outcompeting human-led firms in certain sectors. High-frequency trading bots dominating stock markets and crypto DAOs managing funds provide glimpses. AWM predicts acceleration of this trend.

Cooperation and Memetic Similarity: Cooperation should correlate more strongly with memetic similarity than genetic similarity among non-relatives. Religious or ideological converts should shift cooperation patterns to match new memetic kin rather than genetic kin. Cooperation radius should expand with memetic transmission technology. Gene-meme conflict should produce internal psychological tension—decisions where genetic and memetic coalitions disagree should show longer response times and increased stress markers.

AI Alignment via Broad Metrics: Give one set of AI agents a narrow goal (maximize clicks or a game score) and another set a broad surrogate of human prosperity (weighted mix of economic and social metrics). The expectation is that broad-goal AIs act in more cooperative and human-friendly ways, whereas narrow-goal AIs might shortcut or hack the reward. If evidence accumulates that broader objectives produce safer AI behavior, it supports AWM's alignment prescription.

9. Limitations and Future Directions

High-level abstraction: AWM operates at a high level of abstraction, treating genes, memes, and AI algorithms all as replicators and measuring value in a single currency of capital. This unification glosses over many domain-specific details. In practice, quantifying something like the net present value of future social capital is enormously difficult. AWM should be viewed as a guiding worldview or modeling heuristic rather than a precise predictive tool in its current form.

Contested foundations: The memetic framework on which AWM partially builds remains controversial. Critics argue that cultural items are reconstructed rather than copied with high fidelity (Sperber, 2000), that cultural transmission involves multiple sources violating Mendelian-style inheritance (Sterelny, 2006), and that high-fidelity transmission of cultural information is the exception (Atran, 2001). Boyd and Richerson note that cultural variants are not replicators in the strict sense. AWM uses the replicator concept heuristically to unify analysis across domains, but readers should note the underlying theoretical debates.

Computational tractability: Calculating the optimal coalition structure or running an internal replicator market inside each agent may be computationally intractable for complex systems. The number of possible sub-coalitions grows combinatorially. Real agents will inevitably rely on heuristics and bounded rationality.

Goodhart risks: Any single metric can be exploited (Goodhart's Law). Even with a broad goal, a clever AI might find ways to boost indices in the short run while undermining long-term sustainability. Robust oversight and adaptive metrics are needed to ensure proxies truly track global viability.

Normative concerns: AWM, as a positive theory, does not inherently resolve ethical questions. The model elevates inclusive capital as the paramount objective, which might conflict with values like equality, autonomy, or short-term welfare. AWM is not a moral theory: describing how life does behave when optimizing viability is not the same as prescribing how we should behave. Applications to policy must be tempered by external ethical principles and democratic deliberation.

10. Conclusion

The key to aligning AI—and life itself—may lie in recognizing the coalition game that all living systems are already playing. It is one game, one scoreboard, many players—a single contest of self-interested programs scored by a common metric of future value. AWM identifies what agents fundamentally seek (future capital) and how they organize (into coalitions) to pursue it. By positing a common fitness metric across biology, culture, and AI, AWM offers a coherent way to reason about alignment at all scales.

AWM suggests that aligning AI with human interests is not about instilling arbitrary moral codes, but about designing incentive architectures—both inside AIs and in our institutions—that lead all agents, human or artificial, to converge on strategies sustaining life's long-run future. The framework bridges disciplinary silos, linking evolutionary biology's replicator dynamics with economics' market selection and AI's reward optimization. It provides concrete mechanisms (internal auctions, payoff division, value accounting, equity-based alignment) that make alignment not just a philosophical quest but an engineering challenge: how to build systems where selfish parts produce cooperative wholes.

The demographic transition, aging, altruism, institutional dynamics, and AI risk all find unified explanation within this framework. The goal of reaching approximately $42 quadrillion in GPV provides a concrete target—roughly eight to ten times current civilizational capital—at which point humanity may possess sufficient redundancy to navigate existential risks and pass the Great Filter. By recognizing capital-seeking replicators as the players and taking responsibility for the rules of the game, we have a chance to steer the future toward resilient prosperity rather than catastrophic optimization failure.

AWM is a world-model of life's own self-modeling—a meta-framework by which the biosphere, via programs and coalitions, recursively computes its future. Humans are not destined to be bystanders in the rise of AI; we can coevolve. Rather than view AI as mere tools or inevitable rulers, we can enter a symbiotic alliance: humans providing context, values, and creativity while agiens contribute optimization, scale, and precision. The coming years will reveal whether we can harness AWM to consciously navigate this evolutionary game—aligning emergent intelligences and institutions toward survival rather than self-destruction.

References

Albertus, M., & Menaldo, V. (2018). Authoritarianism and the Elite Origins of Democracy. Cambridge University Press.

Arrow, K. J. (1963). Social Choice and Individual Values (2nd ed.). Yale University Press.

Arrow, K. J., Dasgupta, P., Goulder, L. H., Mumford, K. J., & Oleson, K. (2012). Sustainability and the measurement of wealth. Environment and Development Economics, 17(3), 317–353.

Atran, S. (2001). The trouble with memes. Human Nature, 12(4), 351–381.

Atran, S. (2002). In Gods We Trust: The Evolutionary Landscape of Religion. Oxford University Press.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2), 235–256.

Ausbel, L. M., & Milgrom, P. (2006). The lovely but lonely Vickrey auction. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), Combinatorial Auctions (pp. 17–40). MIT Press.

Austad, S. N., & Hoffman, J. M. (2018). Is antagonistic pleiotropy ubiquitous in aging biology? Evolution, Medicine, and Public Health, 2018(1), 287–294.

Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press.

Bell, A. V., Richerson, P. J., & McElreath, R. (2009). Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. PNAS, 106(42), 17671–17674.

Berdoy, M., Webster, J. P., & Macdonald, D. W. (2000). Fatal attraction in rats infected with Toxoplasma gondii. Proceedings of the Royal Society B, 267(1452), 1591–1594.

Blackmore, S. (1999). The Meme Machine. Oxford University Press.

Bloom, N., Sadun, R., & Van Reenen, J. (2012). The organization of firms across countries. Quarterly Journal of Economics, 127(4), 1663–1705.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Boyd, R. (2017). A Different Kind of Animal. Princeton University Press.

Boyd, R., & Richerson, P. J. (1985). Culture and the Evolutionary Process. University of Chicago Press.

Boyd, R., & Richerson, P. J. (2005). The Origin and Evolution of Cultures. Oxford University Press.

Brown, T. B., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England Journal of Medicine, 357(4), 370–379.

Christakis, N. A., & Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. New England Journal of Medicine, 358(21), 2249–2258.

Clarke, E. H. (1971). Multipart pricing of public goods. Public Choice, 11(1), 17–33.

Coase, R. H. (1937). The nature of the firm. Economica, 4(16), 386–405.

Corrado, C., Hulten, C., & Sichel, D. (2009). Intangible capital and U.S. economic growth. Review of Income and Wealth, 55(3), 661–685.

Dawkins, R. (1976). The Selfish Gene. Oxford University Press.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. Journal of Finance, 25(2), 383–417.

Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Clarendon Press.

Goldin, C. (2021). Career and Family: Women's Century-Long Journey toward Equity. Princeton University Press.

Gordon, M. J., & Shapiro, E. (1956). Capital equipment analysis: The required rate of profit. Management Science, 3(1), 102–110.

Groves, T. (1973). Incentives in teams. Econometrica, 41(4), 617–631.

Hamilton, W. D. (1964). The genetical evolution of social behaviour I & II. Journal of Theoretical Biology, 7(1), 1–52.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. Nature, 450(7169), 557–559.

Handley, C., & Mathew, S. (2020). Human large-scale cooperation as a product of competition between cultural groups. Nature Communications, 11, 702.

Hayek, F. A. (1945). The use of knowledge in society. American Economic Review, 35(4), 519–530.

Henrich, J., Boyd, R., Bowles, S., et al. (2004). Economic man in cross-cultural perspective. Behavioral and Brain Sciences, 28(6), 795–855.

Hobbes, T. (1651). Leviathan. Andrew Crooke.

Hurwicz, L. (2008). But who will guard the guardians? American Economic Review, 98(3), 577–585.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture. Cognitive Science, 15(2), 219–250.

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Kirkwood, T. B. L. (1977). Evolution of ageing. Nature, 270(5635), 301–304.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., & Zweimüller, J. (2019). Child penalties across countries: Evidence and explanations. AEA Papers and Proceedings, 109, 122–126.

Lindholm, A. K., et al. (2016). The ecology and evolutionary dynamics of meiotic drive. Trends in Ecology & Evolution, 31(4), 315–326.

Lotka, A. J. (1922). Contribution to the energetics of evolution. PNAS, 8(6), 147–151.

Masumori, A., & Ikegami, T. (2025). Do large language model agents exhibit a survival instinct? An empirical study in a Sugarscape-style simulation. arXiv:2508.12920 [Preprint].

Medawar, P. B. (1952). An Unsolved Problem of Biology. H. K. Lewis.

Mesoudi, A. (2011). Cultural Evolution. University of Chicago Press.

Minsky, M. (1986). The Society of Mind. Simon & Schuster.

Myerson, R. B. (1981). Optimal auction design. Mathematics of Operations Research, 6(1), 58–73.

Newson, L., Postmes, T., Lea, S. E., & Webley, P. (2005). Why are modern families small? Evolution and Human Behavior, 26(2), 91–120.

Newson, L., Postmes, T., Lea, S. E., Webley, P., Richerson, P. J., & McElreath, R. (2007). Influences on communication about reproduction. Evolution and Human Behavior, 28(6), 399–410.

North, D. C., Wallis, J. J., & Weingast, B. R. (2009). Violence and Social Orders. Cambridge University Press.

OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774.

Osborne, M. J., & Rubinstein, A. (1994). A Course in Game Theory. MIT Press.

Przeworski, A., et al. (2000). Democracy and Development. Cambridge University Press.

Richerson, P., et al. (2016). Cultural group selection plays an essential role in explaining human cooperation. Behavioral and Brain Sciences, 39, e30.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Samuelson, P. A. (1998). Summing up on business cycles: Opening address. In J. C. Fuhrer & S. Schuh (Eds.), Beyond Shocks: What Causes Business Cycles? Federal Reserve Bank of Boston.

Shapley, L. S. (1967). On balanced sets and cores. Naval Research Logistics Quarterly, 14(4), 453–460.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. Journal of Economic Perspectives, 17(1), 83–104.

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. Journal of Finance, 52(1), 35–55.

Smith, A. (1776). An Inquiry into the Nature and Causes of the Wealth of Nations. W. Strahan and T. Cadell.

Sperber, D. (2000). An objection to the memetic approach to culture. In R. Aunger (Ed.), Darwinizing Culture (pp. 163–173). Oxford University Press.

Stanovich, K. E. (2011). Rationality and the Reflective Mind. Oxford University Press.

Sterelny, K. (2006). Memes revisited. British Journal for the Philosophy of Science, 57(1), 145–165.

Svolik, M. W. (2012). The Politics of Authoritarian Rule. Cambridge University Press.

Szilágyi, A., Czárán, T., Santos, M., & Szathmáry, E. (2023). Directional selection coupled with kin selection favors the establishment of senescence. BMC Biology, 21, 231.

Taylor, R. J., Chatters, L. M., Woodward, A. T., & Brown, E. (2013). Racial and ethnic differences in extended family, friendship, fictive kin, and congregational informal support networks. Family Relations, 62(4), 609–624.

Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. Journal of Political Economy, 89(2), 392–406.

Trivers, R. L. (1974). Parent-offspring conflict. American Zoologist, 14(1), 249–264.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. Journal of Finance, 16(1), 8–37.

Weingast, B. R. (1997). The political foundations of democracy and the rule of law. American Political Science Review, 91(2), 245–263.

Williams, G. C. (1957). Pleiotropy, natural selection, and the evolution of senescence. Evolution, 11(4), 398–411.

World Bank. (2021). The Changing Wealth of Nations 2021. World Bank Publications.

World Bank. (2024). World Development Indicators. https://data.worldbank.org

Zak, P. J., & Knack, S. (2001). Trust and growth. The Economic Journal, 111(470), 295–321.