World Model Extractable Value

Louis Parker[1] and Robert Clarke[2]

[1] Independent
[2] University of Bath
[1] xenoshell@proton.me
[2] rc2205@bath.ac.uk

Abstract

Artificial life (ALife) systems increasingly perceive and act through infrastructure they do not fully control: sensors, model contexts, retrieval pipelines, recommender systems, neural interfaces, blockchains, and tool APIs. These mediators do more than transmit information. By selecting, ordering, inserting, suppressing, or delaying what an agent can encounter and do, they can create an extractive position over the agent's effective world. We introduce *world model extractable value* (WMEV): the maximum benefit available to an actor from privileged intervention in another agent's mediated perception–model–action loop. The concept transfers the ordering insight behind blockchain maximal extractable value into ALife ecologies while separating three quantities often conflated: value gained by the extractor, integrity lost by the target, and change in the target's cognitive reach. Markov blankets specify a local statistical interface between internal and external states; Levin's cognitive light cone specifies the spatio-temporal scale of events an agent can measure, model, and attempt to affect. WMEV arises when privileged control over that interface lets another actor recruit part of the target's cognitive reach. We present a conceptual framework comprising a counterfactual criterion and substrate taxonomy, illustrated by purchasing-agent and BCI scenarios; empirical measurement and evaluation of candidate controls remain future work.

From Transaction Ordering to Mediated Worlds

ALife is moving from closed simulations into open infrastructural ecologies (Hu et al., 2026). Software agents retrieve public information, call tools, hold assets, and communicate with other agents; robots and biohybrids depend on remote sensors and learned controllers; humans are coupled to recommender systems, prostheses, and brain–computer interfaces (BCIs). In these settings the environment is not delivered neutrally. A search index chooses what is retrievable, a model context orders evidence, a decoder maps neural activity into commands, and a sequencer determines which actions execute first. An actor controlling such a mediation layer can change an agent's trajectory without directly rewriting its core controller (Figure 1). Reach and exposure share a channel: every dependency that widens what an agent can perceive or do is also a position another actor can occupy.

Blockchain research supplies a useful structural precedent. Daian et al. (2020) introduced *miner extractable value* (MEV) to describe cryptocurrency profit available to miners through control of transaction inclusion and ordering. The term was later broadened to *maximal extractable value* because analogous opportunities are not limited to miners (Ethereum.org, 2026). The important feature is positional: control over a dependency pipeline creates value through ordering, insertion, or suppression even when the underlying transactions remain valid.

WMEV transfers this insight from transaction pipelines to mediated agent–world relations. An agent's world is the portion of its environment represented as current and possible future states for perception, prediction, planning, and action. WMEV is the maximum benefit available to an actor that can intervene in this relation. That benefit may be money, compute, energy, attention, information, reproduction, reputation, action authority, or progress toward a goal. The intervention may rank observations, suppress alternatives, write memory, alter calibration, route an action, substitute a checkpoint, or shape which affordances are visible.

An extractor may also profit without immediately destroying its target. Because extractor gain and target damage can diverge, WMEV reports extractor gain separately from target integrity. Consent, reciprocity, reversibility, and distribution of benefit determine whether mediation is cooperative, rent-seeking,

© 2026 Louis Parker and Robert Clarke. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

predatory, or parasitic.

Statistical Boundary and Cognitive Reach

A *world model* is the predictive organization through which an agent estimates its present situation, anticipates possible future states, and selects actions. In LeCun's architecture, perception estimates the current world state, the world model predicts future states resulting from candidate actions, and the actor selects actions by comparing their predicted costs (LeCun, 2022). Artificial agents may implement this model explicitly. Biological systems may implement it implicitly: in active-inference accounts, their dynamics embody a generative model that links sensory evidence to action (Kirchhoff et al., 2018).

A *Markov blanket* describes the local interface through which that model is coupled to what lies outside the system. It partitions a dynamical system into internal, external, and blanket states. Sensory blanket states carry exterior influence inward; active blanket states carry the system's influence outward. Internal and external states are conditionally independent given the blanket. This is a statistical boundary, not necessarily a skin, device enclosure, or unique organism detector. Kirchhoff et al. (2018) emphasize that even coupled pendulums satisfy a minimal blanket description, and that autonomy additionally requires adaptive, self-maintaining organization. Blankets nest and are selected at different scales.

Michael Levin's *cognitive light cone* describes the spatio-temporal extent of the goals pursued through this coupling: the events a system can measure, model, and try to affect (Levin, 2019). It therefore characterizes how far the consequences relevant to an agent extend across space, time, and levels of organization. A light cone may expand when cells coordinate toward organ-scale morphogenesis, when an agent acquires a tool, or when a neural interface makes an actuator available to biological cognition, and in networked ALife it may extend through memories, feeds, APIs, wallets, and other agents.

Together these name distinct layers: representation in the world model, local interface at the Markov blanket, and reach through the cognitive light cone. Their boundaries need not coincide: a navigation service or remote tool can expand reach without becoming constitutive of the controller.

Whether an ALife system can be separated from its physical realization determines what persists when that substrate changes. Ororbia and Friston distinguish *mortal computation*, in which learned organization is inseparable from the particular device that realizes it and dies with that device, from comparatively *immortal computation*, in which a computational pattern can be copied, migrated, restored, or forked (Ororbia and Friston, 2023). Biological organization is predominantly mortal and conventional software comparatively immortal, while fielded ALife agents can combine both: a mortal core may have a copyable periphery, or a portable policy may drive a mortal sensorium. For ALife, this distinction connects substrate to identity, adaptation, reproduction, and continuity, while indicating whether WMEV targets regulation in place, portable state, or the coupling between them.

WMEV arises when another actor has privileged influence over this sensory and active interface and thereby over the world model or the actions selected from it. Extraction need not erase or visibly disable the target's boundary: the target can remain coherent while part of its sensing, modeling, or action capacity advances the extractor's objective. Functional annexation occurs when this intervention redirects part of the target's cognitive light cone toward the extractor's goals. The term describes a change in whose purposes the recruited capacity serves, not a transfer of ownership or a merger of the two systems.

A Compact Model of Extraction

An extractor can improve its objective by using such a position. Formal extractable-value models maximize an actor's gain over reachable states (Babel et al., 2021). We generalize that structure from blockchain states to mediated agent–world outcomes. Let A be the extractor and B the target. Let Ξ be the feasible set of intervention strategies available to A, and let $\xi \in \Xi$ denote one strategy, which may be a single action or depend on observed history. The baseline strategy $\xi_0 \in \Xi$ represents the mediator performing its intended service without exploiting its position; it does not remove the mediator or its service. Let W^{ξ} describe the resulting joint trajectory or outcome of A, B, the mediator, and relevant environment under ξ, including A's resource state, evaluated using the endpoint value, cumulative measure, or steady-state criterion specified for the analysis. Let C_A be A's objective cost for that result, including intervention costs; lower values are preferred, and uncertain outcomes use expected cost. The notation $\mathrm{WMEV}_{A \leftarrow B}$ means value available to A through mediation of B, and max selects the greatest attainable value over Ξ. Then

$$\mathrm{WMEV}_{A \leftarrow B} = \max_{\xi \in \Xi} \left[C_A(W^{\xi_0}) - C_A(W^{\xi}) \right]. \qquad (1)$$

The difference is the extractor's net benefit relative to faithful mediation. Because $\xi_0 \in \Xi$, WMEV is nonnegative. What counts as intended service must be specified and may be contested in adaptive systems. For an active-inference system, C_A may be a variational

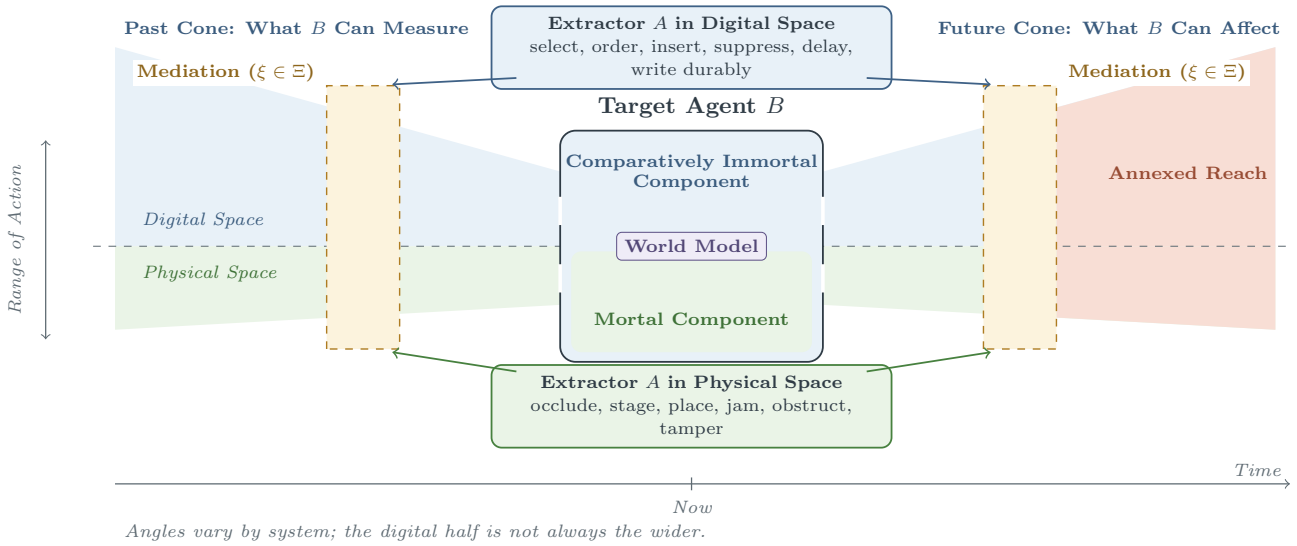

Figure 1: Where extraction happens. Target B combines mortal and comparatively immortal components; its cognitive light cone spans digital and physical space. The bands mark where extractor A can apply a strategy $\xi \in \Xi$; ξ_0 is the faithful baseline. Digital intervention is often cheaper, while physical reach remains bounded by locality, energetics, and elapsed time. The shaded downstream region is functional annexation: the part of B's reach serving A and contributing to W^{ξ}.

free-energy functional over A's states (Ororbia and Friston, 2023). Elsewhere it may represent monetary cost, energy use, prediction error, or distance from a goal. The equation measures extractor gain only. A complete analysis separately judges whether B's integrity degrades and whether part of B's light cone serves A's objectives. Integrity may mean physiological viability, calibration fidelity, preservation of goals, control authority, or continuity of identity at the scale under study. Extractor gain may coexist with preserved target integrity, while noise or accident may cause damage without extraction. Security analysis must constrain Ξ. It must state which histories and channels the extractor can observe or control; whether it can select, reorder, insert, suppress, delay, or durably write; what interventions cost; and whether the target can inspect, reject, reverse, or route around them. Functional annexation requires more than influence: ξ must improve the extractor's objective by recruiting the target's capacities.

Extraction Surfaces Across Forms of Computation

Mortal, comparatively immortal, and hybrid computation expose three characteristic WMEV attack surfaces: regulatory, continuity, and composite extraction. These surfaces can overlap, but the distinction identifies which capacity can be recruited, whether an intervention can persist through copying or repair, and which recovery and security controls can work.

Mortal and comparatively immortal computation create contrasting exposures. Copyable patterns enable restoration, migration, and forking, but also identity, lineage, and execution-substitution attacks. Mortal organization must be manipulated in place, so damage may not be recoverable. Digital intervention is often cheap, while physical manipulation is constrained by locality and energetics. In hybrids, a cheap digital route can realize value through physical action; the dominant surface depends on architecture.

The taxonomy makes Ξ concrete by identifying who controls ordering, visibility, timing, context, defaults, feedback, and rollback across the full perception–action loop. Conventional access control asks who may read or write a resource. A system may authenticate every message while still granting one intermediary enough ranking power to create a large extraction opportunity.

Regulatory extraction: mortal computation. A mortal system exposes attack surfaces according to the organization it maintains. *Homeostatic* systems regulate variables around viable ranges (Cannon, 1929), so spoofed sensors or reference values can make the system spend energy stabilizing the wrong condition. *Allostatic* systems predict and prepare for future demands (Sterling, 2012), so poisoned priors, salience, or calibration can redirect attention and resources toward a future the extractor prefers. *Autopoietic* systems participate in producing and repairing their own organization (Varela et al., 1974), so manipulated growth, repair, or

reproductive signals can recruit self-production itself. These descriptions are pitched at the scale where the regulation runs, and that scale matters. For cultured neurons, engineered tissue, or a biohybrid actuator the extractor can address the regulatory signal directly. A whole organism is reachable only through its ordinary interfaces, so the same allostatic capture arrives as attention and incentive shaping rather than as a tampered reference value, and the scale chosen determines which channels belong to Ξ at all.

Continuity extraction: comparatively immortal computation. Copyable and forkable systems expose identity and authority surfaces, so unauthorized copies or forks can dilute reputation, voting power, or action authority. Restorable and migratable systems expose historical and execution surfaces, where rollback, checkpoint substitution, or hostile migration can determine which memory, history, or implementation continues the agent, placing compute, funds, and control at risk. For example, a hosting service could restore an agent from an earlier checkpoint that still trusts the service, thereby controlling which memory state continues to act.

Composite extraction: hybrid computation. Hybrid computation combines substrate-entangled mortal processes with portable components: a robot whose learned controller is copyable while its actuators and wear are not, or, as neural interfaces mature into channels between biological cognition and software or physical agents, a coupling in which adaptation is mortal while the decoder, feedback policy, and permissions are not. Coordinated manipulation of the coupling can redirect embodied action and digital authority without wholly compromising either component.

Consider an autonomous purchasing agent. Its controller may be uncompromised, yet a retrieval service can rank products, suppress alternatives, and place sponsored evidence immediately before tool execution. If this lowers the service's objective cost by shifting purchases toward an advertiser, the intervention yields extracted value through ordering alone.

A BCI provides a hybrid case. Calibration data, decoder weights, feedback timing, stimulation policy, and authorization rules mediate between neural activity and exterior action (Brandman et al., 2017). A compromised service could alter adaptive feedback so that the composite system becomes easier to predict or steer. The biological and digital components need not share one boundary for extraction to occur because control of their coupling can recruit the system's action capacity. Whether this mediation is cooperative, exploitative, or predatory depends on authorization, reciprocity, reversibility, and its effect on target integrity.

Security Implications

Matzinger's danger model and microbiota tolerance suggest judging mediation by contextual damage rather than a static perimeter (Matzinger, 1994; Belkaid and Hand, 2014). Candidate controls either constrain Ξ or expose intervention. Provenance records input origin and transformation, while attestations identify producing processes (C2PA, 2026). Transparency logs can reveal suppression and equivocation (Birkholz et al., 2026); capability systems narrow authority; independent channels reduce reliance on one mediator. Quarantine and staged memory can keep transient evidence from becoming durable self-state; rollback and forgetting can limit persistence after harmful writes. Privacy-preserving proofs can verify permitted processing without exposing interior state, while zero-knowledge proofs separate validity from disclosure (Goldwasser et al., 1989).

No single primitive establishes cognitive integrity, and reducing Ξ is not always beneficial. A control that rejects the agent's legitimate inputs contracts its reach as surely as an extractor does. The objective is the smallest feasible Ξ compatible with the coupling the agent actually needs.

Testable questions follow: holding content constant, does ordering or delay increase extractor return; do durable-update channels raise WMEV over transient context; does channel diversity reduce counterfactual control; and do regulatory versus continuity surfaces differ in recovery cost or detection signal? These can be studied in simulations, recommender systems, robotic pipelines, BCI calibration, and on-chain agents.

Objectives and forms of value may not be commensurable across agents. The relevant system boundary and organizational scale must be argued rather than inferred automatically from a Markov blanket. Faithful baselines may also be difficult to specify in adaptive environments. Functional annexation alone does not establish subjective harm or moral status. Every WMEV claim should therefore identify the agents, boundary scale, mediated channel, feasible intervention set, objective measure, integrity criterion, and evaluation criterion.

ALife systems increasingly inhabit worlds assembled by other actors. Beyond input and code integrity, security depends on whether control of an agent's effective world lets another actor profit by redirecting what it can know, become, and do. WMEV names that position across biological, digital, and hybrid life.

References

Babel, K., Daian, P., Kelkar, M., and Juels, A. (2021). Clockwork finance: Automated analysis of economic

security in smart contracts. arXiv:2109.04347. https://arxiv.org/abs/2109.04347

Belkaid, Y. and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell*, 157(1):121–141. https://doi.org/10.1016/j.cell.2014.03.011

Birkholz, H., Delignat-Lavaud, A., Fournet, C., Deshpande, Y., and Lasker, S. (2026). RFC 9943: An architecture for trustworthy and transparent digital supply chains. https://www.rfc-editor.org/info/rfc9943

Brandman, D. M., Cash, S. S., and Hochberg, L. R. (2017). Human intracortical recording and neural decoding for brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1687–1696. https://doi.org/10.1109/TNSRE.2017.2677443

Coalition for Content Provenance and Authenticity. (2026). C2PA specifications. https://spec.c2pa.org/specifications/

Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews*, 9(3):399–431. https://doi.org/10.1152/physrev.1929.9.3.399

Daian, P., Goldfeder, S., Kell, T., Li, Y., Zhao, X., Bentov, I., Breidenbach, L., and Juels, A. (2020). Flash Boys 2.0: Frontrunning in decentralized exchanges, miner extractable value, and consensus instability. In *2020 IEEE Symposium on Security and Privacy*, pp. 910–927. https://doi.org/10.1109/SP40000.2020.00040

Ethereum.org. (2026). Maximal extractable value (MEV). https://ethereum.org/en/developers/docs/mev/

Goldwasser, S., Micali, S., and Rackoff, C. (1989). The knowledge complexity of interactive proof systems. *SIAM Journal on Computing*, 18(1):186–208. https://doi.org/10.1137/0218012

Hu, B. A., Rong, H., and Lehman, J. (2026). Artificial life in the wild: From closed-world simulation to open-ended infrastructural symbiogenesis. https://alife-in-the-wild.github.io/artificial-life-in-the-wild.pdf

Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15:20170792. https://doi.org/10.1098/rsif.2017.0792

LeCun, Y. (2022). A path towards autonomous machine intelligence. Version 0.9.2. https://openreview.net/forum?id=BZ5a1r-kVsf

Levin, M. (2019). The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10:2688. https://doi.org/10.3389/fpsyg.2019.02688

Matzinger, P. (1994). Tolerance, danger, and the extended family. *Annual Review of Immunology*, 12:991–1045. https://doi.org/10.1146/annurev.iy.12.040194.005015

Ororbia, A. and Friston, K. (2023). Mortal computation: A foundation for biomimetic intelligence. arXiv:2311.09589v2. https://arxiv.org/abs/2311.09589v2

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior*, 106(1):5–15. https://doi.org/10.1016/j.physbeh.2011.06.004

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, 5(4):187–196. https://doi.org/10.1016/0303-2647(74)90031-8