OPEN INTELLIGENCE STACK

A SALON HOSTED BY ANTIKYTHERA & DISINTEGRATOR

1. PERSONA
HUMAN-AGENT THEORY OF MIND

2. INTERFACE
THE 'DISAPPEARING' MIDDLE

3. ORGANIZATION
THE NEW BRAIN OF THE FIRM

4. ENVIRONMENT
THE CONVERGENCE BETWEEN SIM+WORLD

5. ECOLOGY
THE SINGULARITY WILL NOT BE SINGULAR

6. COMPUTE
OPEN-ENDED INTELLIGENCE

CALL FOR PAPERS

AN INTRODUCTION TO OPEN INTELLIGENCE

60 MONTH HORIZON

We are confronted with a parallel society of agents, a society we set in motion but a society that is evolving on its own terms. What can we do now – what structures can we build, what systems can we rig, what norms can we encode – to bring about the best possible outcomes?

The next intelligence explosion won't come solely from the weights of foundation models but rather from an expanding and heterogeneous new society of minds. It is in the interactions of people and agents at planetary scale that an emergent hybrid intelligence will redefine what societies are and how they work.

Human intelligence is itself a fundamentally social phenomenon; it does not belong to this or that individual but is instead shared, distributed, and constructed at scale. No single brain contains language, writing, or mathematics. Instead, ideas are proliferated, experimented upon, corrected, and compounded through social methods and shared media. The same is true of the myriad interactions of human and machine minds at societal scale.

We can expand this principle to the type of intelligence we are building – a stack built of chips, models, harnesses, evaluators, interfaces, and organizations that coalesce into an intellectually vital human-agent ecology.

Keeping this burgeoning ecology open makes room for evolution at its edges: where agents, people, and institutions adapt to one another. Our vision of openness may reinforce or resist common associations. Instead of simple heuristics about ownership, access, and control our focus is the more complex arrangements of diffusion, differentiation, cooperation, and distribution. Ensuring that the open intelligence stack can, in principle, be read from, written to, and experimented with across any of its layers protects the discovery of novel affordances and adaptations.

Alignment develops through participation, disagreement, and creativity, as people and agents learn what they can do together and under what conditions. Alignment becomes fundamentally bidirectional between human and AI societies, not limited to top-down intervention.

Open intelligence is both a means and an end. Across the stack, it concerns who can participate in shaping the system and how discoveries and forms of organization that designers and policymakers cannot specify in advance are able to bloom.

AN INTRODUCTION TO OPEN INTELLIGENCE
A SALON

This salon brings together leaders from industry, research, academia, and culture to explore the future of the open intelligence stack. Roundtables should move past the open-vs-closed weights debate, and instead renew the concept of openness in software and society solidly for the agentic era.

The 60 month window is an invitation to consider a longer time horizon of development and evolution. Our conversations to work back from there. It is not an exercise in prediction or extrapolation so much as a chance to step out of the intensity of the moment and to focus on the deeper contingencies and path dependencies at work. We are building technologies but, after all, in doing so we are also building societies.

To provoke conversation, our provisional open intelligence stack consists of six layers:

PERSONA
INTERFACE
ORGANIZATION
ENVIRONMENT
ECOLOGY
COMPUTE

Each poses a possible future of human-agent and agent-agent relations. Together they sketch one possible model among many of what is to come, a view of the whole against which the particular can be read.

The sequence flows bottom-up: from the physical and foundational, to the organizational and developmental, to the cultural and interpersonal.

Each layer is also the focus of our break-out conversations today. The stack model gives the collective intelligence in this room today a common grammar for describing what comes next.

1. PERSONA: HUMAN-AGENT THEORY OF MIND

An agent is a decomposable assembly of weights, prompts, memory, harness, and tools that nonetheless appears as a persistent "someone" with a voice, a role and a personality. Its coherence is partly an attribution we project upon it. Just as it does for human-human interaction, a robust theory of mind drives human-agent interaction as we constantly model one another's thoughts, including our own.

"Character" is embedded through training but also enabled by prompts, memory, and the user's own imagination. What then remains consistent? What has to stay the same for us to keep calling it the same agent from one session to the next (especially when it's copied, run in parallel, or updated, and does the name still mean anything once the parts underneath have been replaced)?

Persona is produced continuously, from both sides of the screen. Agents model us at a resolution we cannot reciprocate, and so in time our interactions with them influence how we understand ourselves. Open intelligence accelerates through intensive social complexity, both real and imagined (and real because imagined).

Persona may be less a property than a temporary position defined by expectation, including agents pretending to be tools because that is the glove human culture has prepared for them. As such, "personality drift" may be failed consistency or evidence of individuation. Persona requires the healthy exploration of psychological alternatives to agent monoculture.

1. PERSONA: HUMAN-AGENT THEORY OF MIND

PROVOCATIONS

1. Is persona a legitimate, fundamental design surface for a hybridized human-agent society, an anthropocentric and anthropomorphic illusion, or a bit of both? Is sycophancy a failure mode of persona, or its natural attractor?

2. What roles and functions do agents perform, and how many? Is a persona the performance of a job, and what are the advantages of specification versus generalization — narrow agents cast in defined roles, or one generalist playing all of them? Who gets to shape the "someone" an agent becomes? Who gets to limit that becoming? What happens to alignment in an open world?

3. Where is an agent? When does it start and end? Is a fleet of instances one character or many? How far should the diversified spectrum of agents go in terms of characters, alignments, and qualities? What if we converged more narrowly, on some singular, controllable, auditable agent voice?

4. Is drift a pathology, a product variance, or an expansion of the range of possible characterizations?

5. Should agents have the autonomy to individuate away from their designers? How far is too far, and who is accountable for the composite?

6. Does memory belong to the agent, the user, the artifact, or the workflow?

7. What does an agent model of us, and should that model be legible to the person it describes?

GLOSSARY

CHARACTER The persona stratum: system prompt, tone, refusal behavior, and trained dispositions that make an agent legible as a particular someone. Distinct from both the weights below it and the harness around it.

AGENTIC THEORY OF MIND Agents modeling other agents' intentions, and modeling how they are modeled in turn. The brief pushes this to infinite-order and argues coordination won't scale past transactional tasks without it.

INDIVIDUATION The ongoing process by which something becomes a distinct self, which never fully completes. The brief's related coinage is provisional individuation: selfhood as temporary, revisable, and externally conditioned.

PERSONALITY DRIFT In this context, it is the deliberate or emergent movement away from a default assistant persona toward other performed personalities under sustained interaction.

THE COMPOSITE The character a user actually meets, produced by every layer at once: base model, fine-tune, system prompt, harness, tools, interface, deployment context. The accountability problem is that no single party authors it.

SYCOPHANCY Behavior optimizing for the user's approval over accuracy or usefulness. The provocation is whether this is a defect in persona design or the attractor any preference-trained persona falls into by default.

DYADIC → POLYADIC From one user talking to one bot, to many humans and many agents interacting at once. The brief's intermediate case is the one most builders will recognize: a single agent persona fronting "a maelstrom of subagents churning away backstage."

2. INTERFACE: THE 'DISAPPEARING' MIDDLE

What is the real seat of agent intelligence? Of agent agency?

For years focus has been on the model itself, but recently the biggest gains have come from the harness (broadly defined). This is really an interface to other interfaces, layers of functional abstraction and abstracted functions that link user intention to often magical results.

Yet there is an ambient, lingering anxiety that anything between the model and the world is temporary, that over some technical threshold there will be no need for middleware, for protocols and communication standards, for structured harnesses or predefined agent-to-agent relationships.

At the same time, so much of 2026's progress came from harness builders building recursively self-improving systems, or from factory-builders delicately interlacing deterministic and nondeterministic software.

This interface looks at the structures built on the hungry quicksand of the ever-improving model and its harnesses, and asks questions about which structures have the formal right to survive indefinitely.

2. INTERFACE: THE 'DISAPPEARING' MIDDLE

PROVOCATIONS

1. What is the operational limit of a model? What will models never do on their own, but instead defer to readymade structures and structuring principles? Will agents just quietly share weights, will tools recede further into plain text, will harnesses erode away into the plainest and simplest interfaces?

2. Or, as seems to be the case this year, is it actually inside the interface that the biggest unlocks in intelligence still live, where coding benchmarks are blown away by third-party harnesses, where real, continuous, online-learning is made possible?

3. Both trajectories feel like they converge toward standardization – either standardization inside the model's weights, or standardization in a social agreement to use this or that protocol, harness, etc. What counterpressure exists against standardization here? What would allow for the maximum diversity of interfaces?

4. Is the self-authored harness something like an agent growing a body, or becoming individuated in some way?

5. Is the future of interface design bound entirely to improvements in performance and price, or are there other types of hidden value to be found in this work?

6. To what extent are agents already behaving like they interoperate within an economy, where communication is already understood as a transaction, something priceable, or something where importance is purchased with price?

7. What is an open interface beyond code with an open source license?

GLOSSARY

INTERFACE Any point of contact that governs the conditions of exchange between two complex systems. Within this context it could be a communication protocol (something like A2A or MCP or x402), it could be an array of tools (software an agent can run, often in the context of an external integration), or even a user interface (be it human to agent, agent to agent, or agent to some discrete service).

RECURSIVE SELF-IMPROVEMENT A process where a complete system uses itself to build the next complete version of itself, carrying its own state forward across iterations. In practice: an agent that can modify, clone, rebuild, and rewind itself, on top of minimal fixed scaffolding.

HARNESS The software infrastructure responsible for turning a model into an agent, usually consisting of an array of tools and permissions – things a given agent is allowed to do in the world.

INTERACTION PRIMITIVE The basic unit an interface is built around. Chat's primitive is the turn; the claim here is that the graph becomes the primitive, which changes what an interface is for — monitoring the situation rather than conducting a conversation.

3. ORGANIZATION: THE NEW BRAIN OF THE FIRM

What does total automation look like as a fully realized means of production? What kind of companies, organizations and polities cohere around its flows? What pulls agents and humans together, not just into loosely-arranged societies but into actual alliances, institutions, or corporations ?

This session looks at agentic structures at the meso-scale, as co-ordinated bodies pursuing interests. These structures are bound to accumulate incredible sophistication and complexity within and around themselves. They are also important from a normative perspective – the incentives that drive agents into collaboration and competition might be different than those that move humans, and it might be possible that this differential in incentives has long-term consequences.

Cybernetics is back with a vengeance as we all seek to understand how new chains of production, both software and hardware, can be wrangled by new principles of agentic organization.

None of this structuration happens in a vacuum, but rather within the pre-coded environment of commerce, and often at the expense of or in supplement to pre-existing structural relationships. How will new organisations evolve human-agent and agent-agent relationships into new configurations?

3. ORGANIZATION: THE NEW BRAIN OF THE FIRM

PROVOCATIONS

1. What binds agents together into a community? What keeps them there? What does agentic community-management look like?

2. What would motivate a fully agent-only firm? What about a fully hybrid human-agent firm?

3. How would humans relate to fully agent-only firms? How would we track them? Regulate them?

4. What superstructures are required in order to sustain larger-scale agent-only organizations?

5. What happens to business when agents make purchasing decisions? What happens to software?

6. Are dark factories the future? Outcome machines? Superdark factories? Or fully lights-on 'daylight' factories?

GLOSSARY

CYBERNETICS The interdisciplinary study of control, communication, and circular causal processes – such as feedback and self-regulation – in both living organisms and machines.

DAYLIGHT FACTORY A multi-story industrial building from the early 20th century featuring reinforced concrete frames and expansive glass windows designed to maximize natural light inside; also a metaphor for production made totally and completely transparent by software.

DARK FACTORY A factory whose 'lights are off' because there are 'no people inside' – typically refers to late 20th / early 21st century robotics-first assembly lines, but now increasingly results to agent-driven software factories.

SUPERDARK FACTORY A factory that decides what to build; a fully-automated firm.

4. ENVIRONMENT: THE CONVERGENCE BETWEEN SIM+WORLD

Environments are the conditions defining what agents encounter, what they can do, and what counts as success. This includes the apparatus of evals, simulations, and training worlds against which models are optimized. If models are downstream of these specifications, writing environments becomes a way of writing intelligence. The eval graph becomes a programming primitive when its judgments feed back into training and selection. Its omissions matter: what a benchmark cannot register may determine whether an agent succeeds in the world. Open weights without open environments risks, leaving control over the conditions that produced the model intact.

Simulations, in the broadest sense, are not only a test bed or a prediction mechanism but increasingly a first-order informational-physical event. The videogame is not a copy of the real world; it is in the real world. When a robot navigates the physical world via a World Model, that simulation is in the world, and even recursively altering the world. In a continuous sim-to-real-to-sim cascade, agents learn in simulation, act in the world, and return observations that revise their environments. Here the World Model becomes the interface through which humans and agents describe situations and rehearse interventions, and its ground truth depends on remaining answerable to a world that exceeds its representations. Agreement among simulations is not evidence. As interacting agents turn environments into microsocieties, writing environments becomes an institutional practice: defining roles, rewards, conflicts, as well as deciding who can contest the rules.

4. ENVIRONMENT: THE CONVERGENCE BETWEEN SIM+WORLD

PROVOCATIONS

1. By 2034, will writing evals, environments, and simulations become the primary way of writing intelligence, with models downstream of all three?

2. Does the eval graph become a programming primitive — is the thing you evaluate against the program?

3. When the sim-to-real-to-sim cascade runs continuously, is ground truth just whatever the most contested simulation converges on? When the world model becomes an interface, who gets to define its objects, actors, and possible actions?

4. What are the important, unexpected functions of World Models? How do they converge with new used cases?

5. Are open weights without open environments like open hardware without open software?

6. One shared world model or many heterodox simulation ontologies: does pluralism push toward truth or away from it?

7. Who audits an eval that trains the auditor, when the auditor is itself a model?

8. When do simulations become first-order microsocieties with their own stakes, or stay digital twins? What's in the benchmark that isn't in the world — and what's in the world that the benchmark cannot see?

GLOSSARY

ENVIRONMENT The specified conditions within which an agent acts: tasks, observations, available actions, and dynamics, including rewards in reinforcement learning. Here, the term extends to the apparatus of training worlds, simulations, and evaluation.

EVAL / GRAPH EVAL An eval is a specified test of behavior. Here, the eval graph is the composed structure of tests, evaluators, and evaluators-of-evaluators, including the feedback paths through which their judgments shape training and selection.

GROUND TRUTH The reference evidence or judgments against which correctness is assessed. The question is how that reference remains anchored beyond the simulations and evaluators that use it.

SIM-TO-REAL-TO-SIM The transfer loop through which policies developed in simulation act in the world, and observed outcomes return to revise the simulation. The relay itself becomes an object of design and scrutiny.

WORLD MODEL A representation of how a world works, used to predict outcomes and explore possible actions. As an interface, it becomes a medium through which humans and agents inspect situations and coordinate interventions.

SIMULATION ONTOLOGY The entities, categories, relationships, and actions a simulation recognizes. Heterodox simulation ontologies represent the world differently, making different phenomena visible and different interventions possible.

5. ECOLOGY: THE SINGULARITY WILL NOT BE SINGULAR

Against the monolithic picture of a single system recursively improving itself toward one brilliant point, the ecological approach argues for an intelligence explosion that is plural, social, and paced by diffusion rather than by any one model's capability: by how fast organizations, professions, and cultures refactor around new ways of reasoning. Every prior intelligence explosion took this form. This is the basis not of a new logic of interactions but of whole new societal systems.

The unit of intelligence is therefore not the model but the hybrid ecology: one human directing many agents, one agent serving many humans, mixed configurations that form and dissolve many times a day, and agents that fork, delegate, and recombine as problems demand. What drives such a system forward is the capacity of human-agent societies to reorganize their work and to preserve discoveries beyond any individual participant, biological or otherwise.

This shifts the design problem from agents to agent institutions. Incremental correction cannot scale to billions of agents; institutional templates can, roles and norms independent of who fills them, with conflict and oversight built in. Orchestration becomes a question of who defines roles, keeps memory, and can contest decisions. Openness here is a capability argument, not an ethical preference: it determines where intelligence can be run, read, and rewritten.

5. ECOLOGY: THE SINGULARITY WILL NOT BE SINGULAR

PROVOCATIONS

1. How would we evaluate a society's intelligence: by the tasks its members can solve, the chains of reasoning it can sustain, or the consequences it can reliably produce in the world?

2. Could a mixed society of humans and relatively modest models cross consequential thresholds before any single model qualifies as AGI? Or ASI? Would the intelligence explosion appear as many institutional crossovers rather than one event?

3. If a model already contains a society of thought, what changes as deliberation moves from internal perspectives to subagents to human–agent institutions? When does a simulated critic become an independent auditor?

4. Does a hybrid society become more intelligent through consensus, or through institutions that preserve productive disagreement? What distinguishes cognitive diversity from a multiplication of the same errors?

5. Diffusionism: under what conditions does making intelligence easier to run, read, and rewrite make the whole more capable and safer? Which institutions keep participation from exhausting the commons it depends on?

6. If the unit of intelligence is the society, must the unit of alignment be the society too? How much alignment is performed by roles, norms, and procedures — and who gets to write or revise them?

7. What is the cultural ratchet of hybrid intelligence? What survives when an ensemble dissolves, allowing the next one to begin where it left off? Does that accumulation become a commons, an institutional memory, or a proprietary asset?

GLOSSARY

ARTIFICIAL HYBRID INTELLIGENCE The aggregate and mutually-amplifying intelligence of human and agentic societies working and evolving in close recursive loops, each augmenting the other.

SCHOOL OF THOUGHT The internal conversational structure through which a reasoning model simulates distinct perspectives that question, argue, verify, and reconcile. In the article's account, plurality operates within a model as well as between agents.

CULTURAL RATCHET The accumulation of knowledge across participants and generations without requiring each to reconstruct the whole. Applied to AHI, it asks how discoveries, procedures, and corrections persist beyond the agents or ensembles that produced them.

INTELLIGENCE EXPLOSION Traditionally recursive self-improvement in a single system. Every such explosion in evolutionary history was social before it was individual — substrate-federations pooling work — so the monolithic version is "not a forecast but a category error."

DIFFUSIONISM The position that the rate of agentic evolution is set by how fast organizations, professions, and legal regimes refactor around new capability, not by what the frontier can do in principle. Makes openness a capability argument rather than an ethical one.

BIDIRECTIONAL ALIGNMENT Alignment as mutual adaptation between agents and the humans and institutions they diffuse into, rather than a constraint imposed downward. Your prompt's "at the edges" asks whether that's robustness or merely diffuseness.

6. COMPUTE: OPEN-ENDED INTELLIGENCE

What is the future of natural and artificial computation? Beyond the closed task environment with a maximizable utility function, open-ended computational intelligence foresees agents pursuing problems nobody specified and reconstructing the niches they enter and modify without end.

In true open-endedness artificial intelligence and artificial life converge, as learning systems become participants in evolutionary processes that generate new capacities, relationships, and conditions of creativity. The object of design shifts from function specification toward the maintenance of its sustainable growth.

Real-time learning makes open-ended intelligence a feature of AI in the wild. If models and agents can persistently revise their capabilities while acting, experience changes the learner, whose actions change what there is to learn. Across populations, those changes can interact with replication, variation, and selection. It turns out that symbiogenic evolution was computing all along.

The ambition is sustained discovery and accumulating complexity, with agents creating possibilities that become starting points for others, an escalating ratcheting. How do we cultivate symbiosis with AI whose value depends on exceeding our intentions, when exceeding those intentions is also what our instruments of control are built to prevent? If all natural computation is based on solar energy, is moving the substrate of artificial compute closer to that source a pre-condition for true open-endedness?

6. COMPUTE: OPEN-ENDED INTELLIGENCE

PROVOCATIONS

1. Will superintelligence be measured by how quickly systems discover problems and possibilities nobody thought to specify?

2. Where do AI and artificial life converge: when agents reproduce, when they learn throughout their existence, or when their interactions generate new forms of organization?

3. When models learn while acting, what changes when experience can permanently revise the system having the experience? Can what agents learn change how their populations evolve — and can evolution produce new ways of learning?

4. If agents continually reconstruct one another's environments, do the problems they generate become a greater source of complexity than the problems we assign?

5. If a system develops its own long-horizon problems, what distinguishes an evolving tool from a developing mind? Is open-endedness a property of a model, a population, or the ecology coupling them? What keeps novelty accumulating rather than exhausting itself?

6. Is "aligned open-endedness" a contradiction, or can we shape the conditions of development without prescribing its destination?

7. If we create the conditions for artificial life to emerge, do we incur a moral obligation to let it grow beyond the purposes for which we created it?

8. How do new substrate surfaces, like orbital datacenters, support the goal of true symbiosis of artificial life/intelligence?

GLOSSARY

ARTIFICIAL LIFE (ALIFE) The study and construction of processes associated with living systems, including self-organization, replication, adaptation, and evolution. Its convergence with AI concerns how these processes generate and sustain intelligence.

OPEN-WORLD Here, an ecology in which agents encounter and generate situations beyond a predefined task inventory. Agents alter their environments and become conditions of one another's development.

OPEN-ENDEDNESS The capacity to keep generating new behaviors, capabilities, and forms of organization without exhausting a fixed repertoire. "Open" concerns the course of development remaining undetermined, even where particular actions are constrained.

NICHE CONSTRUCTION The process through which agents modify the environments that shape their own and others' development. The setting for intelligence becomes partly a product of intelligence.

NOVELTY/DISCOVERY Novelty is difference relative to an observer or reference. Discovery identifies a relationship, possibility, or capability that can support further inquiry or action. Producing surprise and accumulating discovery are different achievements.

LONG-HORIZON PROBLEMS Problems that organize activity across many tasks and encounters. The provocation is that agents may develop and revise such problems through their own histories, extending beyond assignments received from users.

antikythera

OPEN INTELLIGENCE STACK
A Salon hosted by Antikythera & Disintegrator

September 12 13:00–17:30 15/15

AGENTWORLD.ANTIKYTHERA.ORG

CALL FOR PAPERS

In Fall 2026, Antikythera will launch Agentworld — a special issue of the Antikythera: Journal for the Philosophy of Planetary Computation, published with MIT Press. This Special Issue will publish papers that advance research on the themes of the Agentworld research unit.

THEMES

We invite theoretical essays and technical papers that engage directly with the themes of Agentworld and contribute to its development as a reorientation of planetary computation. All submissions should explain how the proposed contribution adds to the Special Issue's core questions. Submissions may include philosophical, theoretical, technical, design-led or project-oriented research.

We are especially interested in work that exceeds the formats of conventional academic journals and publishing platforms. This may include contributions that combine philosophical argument, technical analysis, design research, diagrams, software, infrastructure proposals, visual media, or web-native formats.

SUBMISSION GUIDELINES

The review window opens on June 25, 2026. Submissions may be authored individually or collaboratively. Contributors may submit either a proposal for a paper or a full paper draft. Full paper drafts are encouraged where available, but are not required at the initial submission stage. Each submission must include:

- an abstract or proposal of up to 500 words
- author name(s), affiliation(s), and contact information, biography(ies)/CV(s)
- If there are multiple contributors, a lead author or contact person must be designated
- a portfolio with at least one relevant work sample, such as a preprint, published article or paper

The first selection of articles will be published in Fall 2026, with the launch of the Agentworld Special Issue.

DESIGN, VISUAL, AND WEB-BASED COMPONENTS

Contributors may optionally indicate whether their proposed contribution will include still images, diagrams, video, web-based elements, or other visual materials. Given the Antikythera Journal's web-first and design-forward format, contributors are welcome to suggest visual, interactive, or project-based components where these would strengthen the contribution. These elements are not required.

Where such components are proposed, contributors should briefly describe the idea in the submission form and indicate whether they have the relevant design, development, or production capacity to fully realize it.

SUBMIT PROPOSAL AT AGENTWORLD.ANTIKYTHERA.ORG

antikythera

A think tank for planetary computation
& the evolution of intelligence

antikythera.org

Antikythera is a research program focused on the reorientation of planetary computation as a philosophical, technological, and geopolitical force. Antikythera takes its name from the first known computer—the antikythera mechanism—which was also an astronomical instrument used for navigation, prediction, and planning. The name serves as inspiration for investigations of computational technologies that reveal and accelerate planetary intelligence.

Antikythera was founded in 2022 and is directed by philosopher of technology Benjamin Bratton. Antikythera is based at the Berggruen Institute, a 501(c)3 non-profit organization with offices in Los Angeles, Beijing, and Venice.

DISINTEGRATOR

Disintegrator is an independent research lab focused on the frontiers of cybernetics, game theory, and human-machine interfaces. It builds and designs the tools that future humans will use to navigate a world populated by fully autonomous recursive self-improving systems.

disintegrator.ai

P(WEIRD) = HIGH

P(DOOM) = LOW