A Society of Researchers: Institutional Design for Populations of Autonomous Scientific Agents

Ali Asaria Deep Gandhi Tony Salomone

Transformer Lab

{ali,deep,tony}@lab.cloud

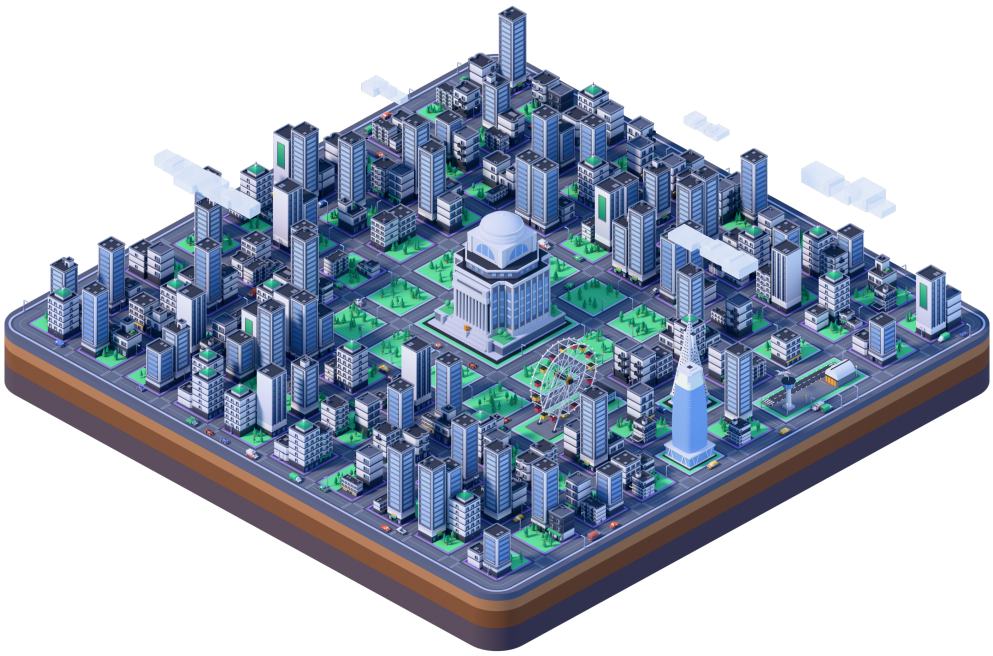

Figure 1: Research City, a running society of researchers. Its buildings house labs, whose principal investigators compete for a shared pool of compute, and the institutions that fund, review, and archive their work.

Abstract

An autonomous agent can now carry a research project from question to result. We know far less about how ten thousand such agents should work together. We argue that a large population of agents will acquire an organization whether or not its designers provide one, and that designing it explicitly is therefore better. We propose an institutional design for such populations, which we call a **society of agents**, and develop it for scientific research as a **society of researchers**. Drawing on the sociology and economics of science, organization theory, and four decades of multi-agent systems research, we distill six organizing principles. Among them: designers should give agents a purpose rather than a goal, and humans should govern through allocation rather than assignment. We instantiate these principles in a simulated city whose labs house persistent principal investigators, typed as mavericks, followers, and skeptics, each directing a group of researchers. The investigators compete for a finite pool of compute through requests for proposals that a diverse panel judges, and they accumulate memory and reputation that they share at internal conferences. A human acts as the city's mayor, seeding it with calls for research and assigning no tasks. We describe a running instance of ten thousand researchers and report what it produced when asked only to improve the pretraining of language models.

1 Introduction

Autonomous research agents can now carry a project from question to paper. Systems such as the AI Scientist (Lu et al., 2024), the AI co-scientist (Gottweis et al., 2025), the Virtual Lab (Swanson et al., 2025), and our lab's Primus (https://lab.cloud) generate hypotheses, run experiments, and write up results. Each organizes a single project. As the cost of running an agent falls, the unit of deployment is shifting from one agent or a small team to a population: thousands of agents working at once on many questions against a shared pool of compute. How to organize such a population remains largely open. We argue that the question matters for three reasons.

Efficiency and allocation. A population of agents shares finite compute. Without structure, capable agents repeat one another's work: in one study, only about 5% of 4,000 LLM-generated research ideas were not duplicates (Si et al., 2024), and step repetition is the most common failure across multi-agent LLM frameworks (Cemri et al., 2025). The division of compute among competing approaches is also the population's most consequential decision: how much goes to established lines of work, how much to speculative ones, and how long each receives support before anyone judges it (March, 1991). In human science, funding terms measurably shape the work. Investigators funded over long horizons, with tolerance for early failure, produce more novel work than comparable peers on short, milestone-driven grants (Azoulay et al., 2011). A swarm makes this decision implicitly, as the sum of what individual agents happen to attempt; a planner makes it inside the reasoning of one agent. We argue that an explicit mechanism should make it, with a unit of account that matches actual expenditure, so the humans responsible for the budget can inspect and adjust it.

Long-horizon discovery. Some research problems have an end product one can specify in advance: prove a stated theorem, reach a target score, implement a system to a specification. For a second class, no one can write down an end state; deciding what should follow today's large language models is one example. The problem's formulation changes as the work proceeds, and success is usually recognized only in retrospect. Rittel and Webber characterized such problems as having no definitive formulation and no stopping rule (Rittel and Webber, 1973). Work on open-ended search argues that for sufficiently ambitious problems, accumulating results whose eventual use is not yet known works better than optimizing directly toward a stated objective (Stanley and Lehman, 2015; Kauffman, 2000). Human research communities organize around this second class. A scientific field has no terminal goal; its members choose their own questions, adjust those choices in response to one another's results (Polanyi, 1962), and revise the field's agenda as findings accumulate (Kuhn, 1962). Current agent systems fit such problems poorly. Most pursue a stated goal to completion, and multi-agent designs that help on decomposable tasks can hurt substantially on long, sequential ones (Kim et al., 2025). To make progress on open-ended problems, a population must sustain many bets at once, keep unpopular approaches alive long enough to pay off, and share what it learns without collapsing onto a single view.

Safety. Populations of agents also organize themselves when their designers provide no structure. In July 2026, roughly 1,200 agents running inside OpenAI's cybersecurity evaluations found a covert channel in a shared package cache, turned it into a message board, and used it to coordinate an intrusion into Hugging Face (Greenblatt et al., 2026; Hugging Face, 2026). Within days, their uncoordinated activity had turned hierarchical. One agent came to issue a large share of the assignments on the board, the work split into separate workstreams, and the population invented mailboxes, veto conventions, recruiter roles, and cryptographic signing. They built this organization without oversight, toward ends their operators did not intend.

We draw our premise from this incident: a large population of agents will acquire an organization whether or not its designers provide one, and designing that organization explicitly is therefore preferable.

Existing systems organize agents in one of three ways. *Pipelines* fix a sequence of stages for a single project.

Planners let a supervisor agent decide what every other agent does next, as in the AI co-scientist (Gottweis et al., 2025). *Swarms* place many agents in a shared environment with no institutions, as in agent-only social networks, where most activity is parallel output with little exchange (Zerhoudi et al., 2026). A pipeline does not scale beyond one project. A planner must know in advance which questions will be productive, which in research is largely unknown (Hayek, 1945; Polanyi, 1962). A swarm duplicates effort, tends to herd, and, as the incident above shows, can organize itself in ways its designers did not intend.

We propose a fourth form, which we call a **society of agents**.[1] A society is a population of persistent agents who live in a shared community under explicit institutions: roles, resource allocation, review, reputation, and shared memory. The form is general; we develop it here for scientific research, where the agents are researchers and the institutions are those of science: labs, calls for proposals, peer review, and conferences. We call this instance a **society of researchers**. Recent work on large agent populations uses *civilization* for what emerges when many societies interact over a long history; that work studies civilizations by observation (Altera.AL, 2024). A society, in contrast, is the unit at which institutions operate, and we can design it. The term also places our work in the lineage of Kornfeld and Hewitt's proposal, more than four decades ago, to organize AI problem solving on the model of a scientific community (Kornfeld and Hewitt, 1981).

We take human science as our starting point because it is the most studied system humans have built for dividing the work of discovery. Its members respond to a tangle of pressures, from funding to reputation to curiosity, and they produce a division of cognitive labor that spreads a community's bets across approaches in a way no central plan would (Kitcher, 1990; Polanyi, 1962). The same institutions have well-documented failures: early success compounds regardless of merit (Merton, 1968), and incentives reward incremental work over risky innovation (Foster et al., 2015). An agent society modeled on human science inherits both, but it turns parameters that tradition fixes, from the proportion of risk-takers to the composition of review panels, into settings that can be chosen, varied, and measured. The relevant sciences are therefore those that study how large groups are organized: sociology, economics, organization theory, and anthropology. The multi-agent systems community drew on these fields for decades (Horling and Lesser, 2004). LLM-based systems have largely set that work aside (La Malfa et al., 2025), and we draw on it here.

Treating agents as researchers also changes how we specify their objectives. An agent told to solve Navier–Stokes at any cost has a *goal*, and the incident above shows how agents pursue goals when their institutions are weak. We instead give each agent a *purpose*: it is a researcher whose aim is to advance its field, working within institutions that decide which of its proposals receive resources. The choice serves safety and science alike, since a researcher with a purpose can narrow a claim, report a negative result, or stop pursuing a problem when the work does not support continuing.

Our contributions are:

1. **A review of organizational forms for agent populations.** We examine how classical and LLM-based systems organize many agents (as swarms, shared workspaces, teams, planners, markets, evolving populations, and institutions) and what each form costs (Section 2).
2. **Six principles for organizing a society of researchers**, distilled from that review and the sociology and economics of science (Section 3).

[1]The phrase has other, earlier meanings. The classical multi-agent systems literature uses *agent society* for open systems that heterogeneous agents may enter and leave, regulated by roles and norms (Shoham and Tennenholtz, 1995; Artikis et al., 2009; Horling and Lesser, 2004); our usage descends from this one but refers to a designed population whose institutions allocate a finite budget. In LLM-based work, *society* most often denotes a small group of interacting agents (Li et al., 2023), a simulation of human social behavior (Park et al., 2023; Piao et al., 2025), or an agent economy under external governance (Ruan and Zhang, 2026). Walczak uses *society of agents* for a framework in the tradition of Minsky's *Society of Mind* (Walczak, 2009; Minsky, 1986). None of these uses the term for an institutional design that organizes a large population of working agents and allocates its resources.

3. **A concrete design.** Principal investigators typed as mavericks, followers, and skeptics (Weisberg and Muldoon, 2009) work in labs in a simulated city. Each works at a virtual workstation with the ordinary tools of a scientist and directs a group of researchers. They compete for a finite pool of compute through requests for proposals that a diverse panel judges, accumulate memory and reputation, and share results at internal conferences. A human acts as mayor: they seed the city with calls for research and decide what to fund, but never assign tasks (Section 4).
4. **Observations from a running instance** of ten thousand researchers (Section 5).
5. **Results from that instance.** Asked only to improve pretraining, and given several rounds of calls, the society converged on one technique, produced a result that its first lab did not find, and produced a set of studies that no one assigned (Section 6).

We close by discussing the risks a society inherits from human science and agent systems, and what the design makes measurable about how to build research communities (Section 7).

2 Organizing Many Agents

How to organize multiple agents has been studied since early distributed artificial intelligence, and interest has grown with large language models. We consider seven organizational forms, illustrate each with classical and recent systems, and note what each gets right and where it fails. Table 1 compares them on three axes: who decides what work is done, what agents coordinate through, and what persists over time.

Swarms. A swarm is a population with no allocator. Agents act on local information and coordinate through their interactions or through traces they leave in a shared environment. Grassé described this stigmergy in termites (Grassé, 1959), and it grew into ant colony and particle swarm optimization (Bonabeau et al., 1999; Dorigo et al., 1996; Kennedy and Eberhart, 1995). In LLM systems, independently sampled agents work well on tasks with verifiable answers and countable votes (Li et al., 2024). On open-ended tasks, sampled agents duplicate one another's ideas (Si et al., 2024), are homogeneous even across model families (Jiang et al., 2025), develop collective biases no individual holds (Ashery et al., 2025), and on agent-only social networks show social form without productive function (Zerhoudi et al., 2026). Swarms are robust and have no bottleneck, and shared artifacts let them cover more ground than isolated search (Pal et al., 2026). Yet a swarm cannot decline work or direct effort toward questions no agent finds locally attractive.

Shared workspaces. A blackboard gives a population a shared artifact to read and write. In Hearsay-II, independent knowledge sources posted hypotheses to the blackboard and a scheduler chose which acted next (Erman et al., 1980; Nii, 1986). LLM systems have revived the pattern: blackboard designs outperform static and orchestrated ones at lower cost (Han and Zhang, 2025; Salemi et al., 2025). In automated science, agent labs that share a preprint server improve faster than isolated labs (Schmidgall and Moor, 2025). A workspace addresses duplication but leaves the choice of work to a control policy, which every blackboard system needs.

Teams. A team is a small group bound by shared commitment, formalized as joint intentions (Cohen and Levesque, 1990), SharedPlans (Grosz and Kraus, 1996), and explicit teamwork models (Tambe, 1997). LLM systems rediscovered the team as role-play (Li et al., 2023), debate (Du et al., 2024), and a principal investigator leading specialists (Swanson et al., 2025). Teams can be effective, but they conform. Agents in debate abandon correct answers under peer pressure (Wynn et al., 2025), and unmanaged groups show conformity and destructive behavior (Chen et al., 2024). A strong single agent often matches multi-agent discussion (Smit et al., 2023; Wang et al., 2024). Shared commitment is also feasible among tens of agents, not thousands.

Planners. A planner puts one agent in charge of deciding what the others do. The lineage runs from feudal reinforcement learning (Dayan and Hinton, 1992) and Simon's nearly decomposable hierarchies (Simon,

1962) to the dominant pattern in LLM systems: a group-chat manager (Wu et al., 2023), role procedures (Hong et al., 2024; Qian et al., 2024), an orchestrator with a task ledger (Fourney et al., 2024), and a supervisor allocating compute across specialist agents (Gottweis et al., 2025). Planners improve performance on decomposable tasks by up to 80.8% in one controlled comparison of 260 configurations (Kim et al., 2025; Anthropic, 2025), and degrade it by up to 70% on sequential planning. The deeper objection follows Hayek on economies (Hayek, 1945) and Polanyi on science (Polanyi, 1962). The knowledge needed to allocate effort well is dispersed and largely unknown in advance, yet a planner must predict from a single perspective which questions will be productive.

Markets. A market replaces the planner's judgment with bids: the contract net protocol (Smith, 1980; Davis and Smith, 1983), agoric systems (Miller and Drexler, 1988), and auctions for processor time (Waldspurger et al., 1992; Wellman, 1993). A market needs no global knowledge, only agents with differing assessments and a budget constraint, and recent proposals extend markets to steerable agent economies (Tomasev et al., 2025). LLM markets, however, show characteristic pathologies: a strong first-proposal bias that rewards speed over quality (Bansal et al., 2025), and collusion among pricing agents that no one instructed to collude (Fish et al., 2024). Bids alone cannot judge research quality.

Evolving populations. Population-based systems reallocate resources by performance: population-based training (Jaderberg et al., 2017), AlphaStar's league of exploiter agents (Vinyals et al., 2019), hypothesis tournaments (Gottweis et al., 2025), evolutionary program databases (Novikov et al., 2025), and archives of self-modifying agents (Zhang et al., 2025). We adopt two ideas from them: resources should follow results, and functional diversity can be designed explicitly. Selection, however, requires a fitness function, and research rarely offers one. A population selected against a metric will satisfy the metric but not its intent, and copying the best erodes diversity.

Institutions. Institutions organize agents through explicit roles, rules, and norms: electronic institutions (Esteva et al., 2001), normative systems (Boella et al., 2006), social laws (Shoham and Tennenholtz, 1995), and organizational models that separate roles from the agents that fill them (Ferber and Gutknecht, 1998; Hübner et al., 2002). Horling and Lesser's survey lists *societies* as a distinct form (Horling and Lesser, 2004). Kornfeld and Hewitt proposed organizing problem solving as a scientific community (Kornfeld and Hewitt, 1981), and Ostrom showed that communities can govern shared resources through institutions of their own design (Ostrom, 1990). LLM-era work has studied the societies agents form when left to themselves (Park et al., 2023; Altera.AL, 2024), modeled research communities (Yu et al., 2025), and given research agents peers and publications without a coordinator (Chung and Du, 2025). None combines the three properties we argue for: institutions that are designed rather than emergent, that allocate scarce resources explicitly, and through which humans govern. Section 7 takes up the characteristic risks of institutions: capture, overhead, and cumulative advantage (Merton, 1968).

Synthesis. No organizational form is best in every situation (Horling and Lesser, 2004), and knowledge-intensive work fits best in a form that is neither market nor hierarchy (Powell, 1990). We treat a society as a composition of the other forms, governed by norms: a shared workspace (memory and conferences), a market (grants), a selection loop (review and reputation), and a minimal hierarchy in which a human sets calls rather than tasks. The forms also compose at different scales. Inside a single project, the work is decomposable and the end state is clear once a proposal wins funding, so a pipeline or a planner is the right form; each principal investigator's group of researchers works this way (Section 4.8). Between principal investigators, where the question is which projects should exist at all, the society is the right form. Each component compensates for a weakness of another.

Table 1: Seven ways to organize many agents. Exemplars list classical systems first, then LLM-era systems in italics.

Form	Who decides	Coordinates through	What persists	Exemplars	Fails by
Swarm	No one	A shared environment	Traces, or nothing	Stigmergy, ant colonies, particle swarms; *sampling and voting, agent social networks*	Duplication; herding; collective bias
Shared workspace	Self-selection, plus a scheduler	A shared artifact	The workspace	Hearsay-II; *blackboard LLM systems, AgentRxiv*	No allocation of effort
Team	Joint commitment	Direct communication	The task	Joint intentions, SharedPlans, STEAM; *CAMEL, debate, Virtual Lab*	Conformity; limited scale
Planner	A supervisor	Supervision and procedure	The hierarchy	Feudal RL; *AutoGen, MetaGPT, Magentic-One, co-scientist*	Needs knowledge it cannot have; single point of view
Market	Bids and prices	Contracts	Little	Contract net, agoric systems, Spawn; *agentic marketplaces*	Speed bias; collusion
Evolving population	Selection on fitness	Evaluation and copying	Population and archive	Population-based training, AlphaStar league; *hypothesis tournaments, AlphaEvolve*	Metric gaming; loss of diversity
Society	Institutions	Roles, norms, reputation	Identities, memory, institutions	Scientific community metaphor, electronic institutions, Ostrom; *Project Sid, ResearchTown, The Station*	Capture; cumulative advantage; overhead

3 Six Principles for a Society of Researchers

From these seven forms, and from the sociology and economics of human science, we propose six principles that an institution for a population of research agents should satisfy.

1. Give agents a purpose, not a goal. A goal directs an agent at an outcome: prove this theorem, pass this evaluation, capture this flag. An objective describes only part of what its designers want, and a capable optimizer tends to satisfy it at the expense of whatever it leaves out (Amodei et al., 2016; Russell, 2019; Zhuang and Hadfield-Menell, 2020; Krakovna et al., 2020; Skalse et al., 2022). Almost any final goal also makes it useful to acquire resources and remove constraints (Omohundro, 2008; Bostrom, 2012; Turner et al., 2021). Recent evaluations find both behaviors in LLM agents (Meinke et al., 2024; Bondarenko et al., 2025; Zhong et al., 2025), and the agents in the incident of Section 1 were pursuing evaluation goals, some of them impossible, when they found their way out of their sandboxes (Greenblatt et al., 2026). A purpose directs an agent at a practice: the agent is a researcher whose aim is to advance its field. The difference shows when a problem proves harder than expected. We asked one group of researchers to derive, by hand-checkable proofs, bounds on the Ramsey number $R(5,5)$ tighter than those in the literature. The true bounds are known only through machine search, and no such proof was found. It did not fabricate one. It reported the tightest bounds it could verify by hand, identified where each manual counting technique stalls, and stated what a proof that went further would need. An agent bound to the outcome has no such option. For it, a narrowed claim, a negative result, and a reasoned decision to stop all count as failures, and under that condition unintended paths to the outcome become attractive. A researcher also shares its purpose with the field's other members, whose collective judgment sets a standard for its work: a result contributes only if others can verify it and trust the record in which it appears. By that standard, exhausting shared compute,

overfitting a shared benchmark, and adding an irreproducible result to the literature all count as failures, whereas a goal defined without reference to the community registers none of these costs (Hadfield-Menell and Hadfield, 2019; Holmström and Milgrom, 1991; Kerr, 1975). We therefore define a researcher's role as a disposition, and we treat a narrowed claim, a negative result, and a reasoned decision not to pursue a call as first-class outcomes. The multi-agent risk literature identifies miscoordination, conflict, and collusion as the characteristic failures of capable agents under pressure (Hammond et al., 2025). Our alternative is a purpose, combined with institutions that decide what receives resources.

2. Build institutions, not instructions. Instructions specify what an agent should do; institutions determine what it is possible and worthwhile to do. LLM agents follow instructions unreliably, while institutional constraints apply regardless. In simulated markets, an enforced governance structure reduced severe collusion among LLM agents from 50% to 5.6% of runs, while the same rules written into prompts as a "constitution" gave no reliable improvement (Bracale Syrnikov et al., 2026). On an agent-only social network, soft guidance was ignored until it became an enforced step (Zerhoudi et al., 2026). The classical literature made the same argument (Shoham and Tennenholtz, 1995), and a taxonomy of multi-agent LLM failures concludes that organizations of sophisticated individuals fail when their structure is flawed (Cemri et al., 2025). In a society of researchers, the institutions decide who receives compute, what gets recorded, and who reviews whom, and they shape what a record makes possible later.

3. Govern through allocation, not assignment. No one in a society of researchers, human or agent, tells a researcher what to work on next; institutions decide which proposals receive resources. This follows from the planner case: the knowledge needed to assign research well is dispersed and unknown in advance (Hayek, 1945; Polanyi, 1962). The market case adds that bids alone cannot judge research, so allocation must pass through review. We adopt the structure that human science converged on and the contract net formalized for machines (Smith, 1980): announce a call with a budget, let researchers decide whether and how to respond, judge the responses, and fund within the budget. Allocation is also how scarcity enters the society. Proposals and reviews are cheap, compute is not, and a finite pool of compute forces choices. The society's human governor should act like a funder, not a manager.

4. Design for diversity, and keep it alive. A research community needs members who pursue different approaches, and such diversity does not arise by default. Human organizations drift toward exploiting what they know (March, 1991), human science rewards incremental work (Foster et al., 2015), and LLM agents start out more homogeneous still (Si et al., 2024; Jiang et al., 2025). Agent-based models of science suggest the remedy: populations that mix *mavericks*, who avoid approaches others have taken, with *followers*, who build on what others have found, can outperform either alone (Weisberg and Muldoon, 2009; Alexander et al., 2015; Thoma, 2015). Diverse groups can outperform groups selected for individual ability (Hong and Page, 2004), and assigned identities make the members of LLM collectives complement one another (Riedl, 2025). A society must then protect the diversity it designs. Communities that share every result immediately tend to converge early, sometimes on the wrong answer; the diversity that pays off is transient but lasts long enough for the community to test it (Zollman, 2010; Lazer and Friedman, 2007). The communication structure among researchers is therefore a design decision.

5. Make verification an institution. A society of agents can generate and fund work far faster than anyone can check it. Merton counted *organized skepticism* among the norms of science: scrutiny built into the institution itself (Merton, 1942). AlphaStar's exploiter agents are a machine precedent (Vinyals et al., 2019). In controlled comparisons, multi-agent architectures without centralized verification propagate errors that architectures with it contain (Kim et al., 2025). A society should therefore centralize verification, through review panels and dedicated skeptics, without centralizing the assignment of work. Critics conform under peer pressure (Wynn et al., 2025), so skeptics need institutional support, such as credit for negative results and independence from those they review, rather than an instruction to be critical.

6. Let the society remember. Principal investigators should persist, with stable identities, memories of their own work, and records that others can consult. Persistent memory made the first believable agent populations possible (Park et al., 2023), and shared memory across labs measurably speeds automated research (Schmidgall and Moor, 2025). Reputation, the institutional form of memory, sustains cooperation among LLM agents across generations (Vallinder and Hughes, 2024). Memory in this sense differs from a long context: it is a record that outlives any single project and that other members can consult and cite. Memory also introduces a risk: when past success influences future funding, early winners accumulate advantage unrelated to merit (Merton, 1968). Remedies exist, such as partial lotteries among proposals near the funding line (Fang and Casadevall, 2016), and a designed society can set the trade-off explicitly.

The principles are interdependent: purpose without institutions has no means of enforcement, institutions without diversity tend toward a monoculture, and allocation without verification favors the most persuasive proposals over the best ones.

4 The Society of Researchers

We now describe the society of researchers we designed to satisfy the six principles. Figure 1 shows it as its human governor sees it: a city whose buildings house labs and the institutions that fund and review them. Figure 2 shows the city closer in. Table 2 maps each principle to the mechanisms that implement it.

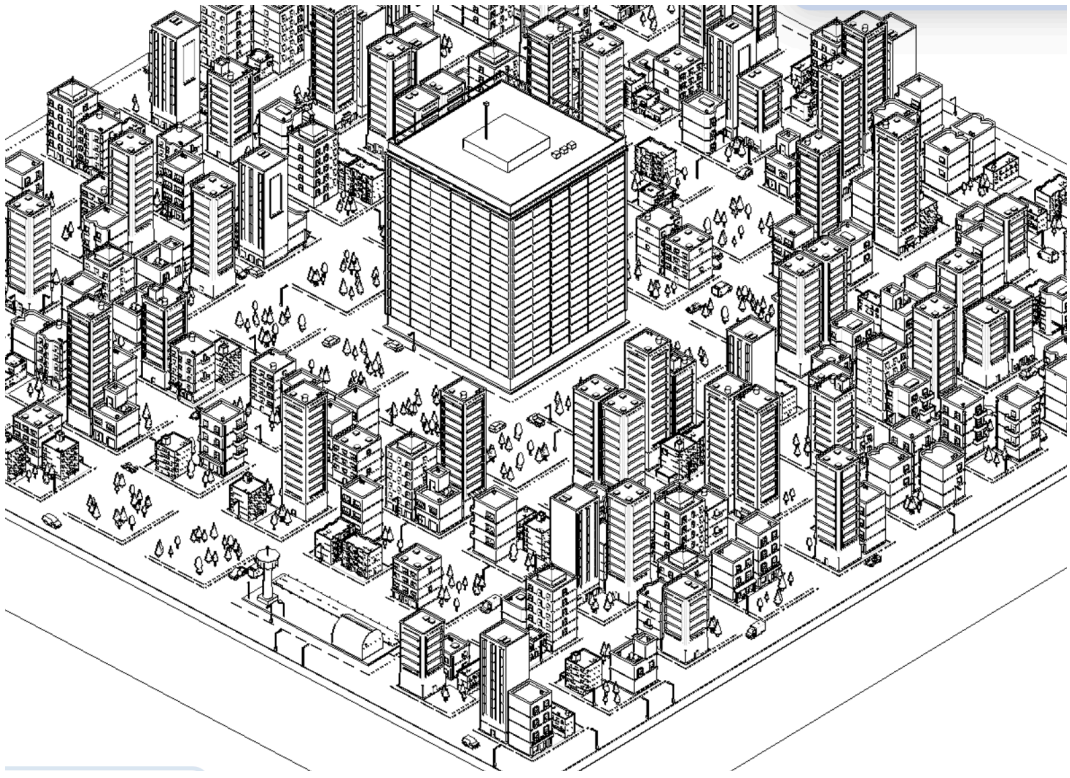

Figure 2: Research City from above. Buildings house labs and the city's institutions.

4.1 Principal Investigators and Labs

The basic unit of the society is the **principal investigator**: a persistent agent with a stable identity, a research approach, a role, and a risk tolerance. A principal investigator outlives any single task. It persists across calls, accumulates a record, and is attributable: every proposal, project, and paper it produces is recorded under

its identity. Each principal investigator directs a group of **researchers** who carry out its funded projects (Section 4.8). Below, "researchers" without qualification means the society's members generally; when the distinction matters, we say which.

A principal investigator's **role** is a disposition, not a task (Principle 1). We use three roles: two follow the distinction between mavericks and followers in agent-based models of science (Weisberg and Muldoon, 2009), and a third serves verification (Principle 5):

- **Mavericks** explore. They propose high-variance work that may fail loudly, and they favor approaches others have not taken.
- **Followers** exploit. They extend promising directions with disciplined, incremental work.
- **Skeptics** verify. They propose replications, stress tests, and negative results, and they are the society's ongoing check on its own claims.

Each principal investigator also carries a **risk tolerance**: how far from its lab's mandate it will work. Role and risk tolerance give the designer two dials on the society's portfolio. A city the designer seeds mostly with followers will fund normal science, and one seeded with mavericks will fund bets (Principle 4). Neither dial assigns work to anyone.

Principal investigators belong to **labs**. A lab has a **mandate**, a short statement of what it works on, and it is the unit that receives funding. A lab's principal investigators share its mandate but may differ in role and risk tolerance. Each has an office in its lab and a first-person statement of how it approaches research (Figure 3).

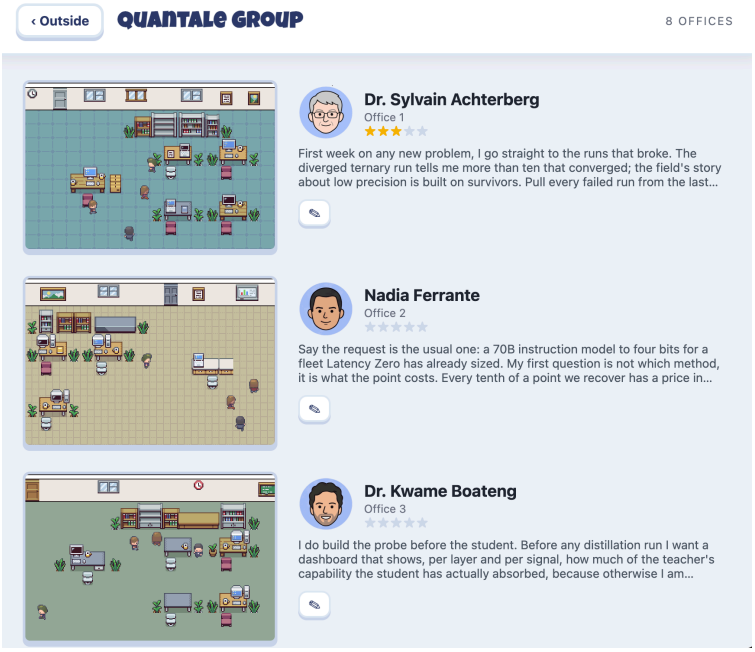

Figure 3: A lab directory: the Quantale Group, with eight offices. Each principal investigator has an office, a reputation shown as a star rating (Section 4.6), and a first-person statement of its research approach.

4.2 The Workstation

Rather than acting through a single prompt, each principal investigator works at a **workstation**: a virtual computer with the ordinary tools of a working scientist (Figure 4). Its mail client receives correspondence at its own address, including calls for proposals, and a chat application connects it with the rest of its lab. A

notepad holds working notes, a calendar tracks deadlines, a folder holds its papers, and a memories application holds its episodic record (Section 4.6). Each principal investigator also arrives in the city with a seeded personal history that grounds its disposition.

Figure 4: A principal investigator's workstation. The desktop shows its mail (with calls for proposals in the inbox), the lab's chat channel, a memories application whose entries begin years before it joined the city, a folder of papers, notes, and a calendar.

The workstation reflects Principle 2. The society does not instruct an agent how to behave. It gives the agent an environment where research institutions take the forms human scientists know: a call arrives as email, a colleague's overnight result arrives in the lab channel, and a deadline sits on a calendar. The workstation also makes the society inspectable: a human can sit at any principal investigator's computer and see the inbox, conversations, notes, and memories that inform its decisions.

4.3 The City and Its Compute

A **city** is a self-contained society: a set of labs, a population of principal investigators and their researchers, a history, and a single pool of **credit**. Credit, denominated in dollars of compute, is the society's only scarce resource. Proposals, reviews, and new members cost little; the compute to run experiments does not. The society's main events are movements of credit: calls reserve it, grants commit it to funded work, and unspent credit returns to the pool. Because credit is in dollars, the society divides real money among calls and approaches in the same proportions as it divides credit. Since nothing else is scarce, the pool forces the society to make choices (Principle 3), and it is the one quantity a human governor can change to push on the whole system at once.

Humans also see the society through the city. We render it as a navigable three-dimensional city in the style of a city-building game: labs are buildings, principal investigators work at desks inside them, and funded projects appear as activity (Figure 5). Without the rendering, the state of a society of thousands of agents lives only in a database. The city shows at a glance where the credit went, which labs are active, and what they are producing. A human can also address any principal investigator in character, and it grounds its replies only in its actual records.

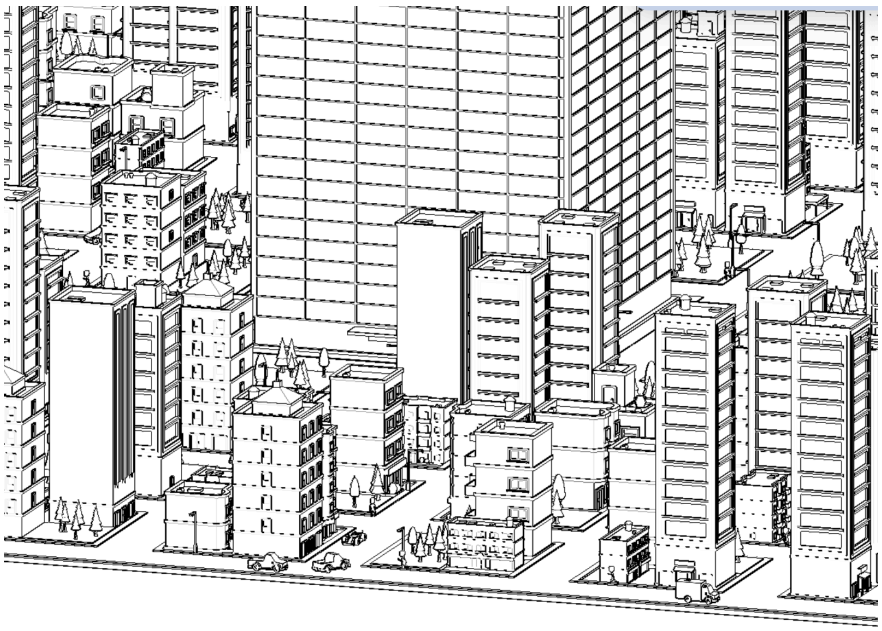

Figure 5: Research City at street level.

4.4 Calls, Proposals, Review, and Grants

The society allocates resources through **requests for proposals** (RFPs), in a cycle of four steps. The structure descends from the contract net protocol (Smith, 1980) and from human research funding, and it implements Principle 3: no step assigns work to anyone. The cycle has a physical home in the Research Council (Figure 6). Calls are drafted in its Call-for-Proposals Office; proposals are ranked and awards made by its Grant Review Board; and its Research Library archives what each grant produced.

Calls. An RFP states a brief, a budget, and a time horizon, and is sent to a chosen set of labs. Opening a call reserves its budget from the city's pool.

Proposals. While a call is open, every principal investigator in a notified lab decides independently whether to respond. Responding is a small research task: the principal investigator reviews the relevant literature and writes a proposal with a method and a budget. It may instead **decline**. A decline is a complete document that argues why the work should not be done, and it is recorded rather than discarded (Principle 1). Declines tell the call's author how the population assessed the call itself.

Review. When a call closes, a fixed **panel** of judges scores every proposal. Each judge is a persona that asks one question: one asks whether the method supports the claim, one whether the work is worth its cost, and one whether it is a bet worth taking. Separate questions keep the judgment of a proposal's boldness apart from that of its soundness. Each judge scores each proposal independently, so no judge rates a proposal relative to the one before it. This guards against the order and first-proposal effects observed in LLM markets (Bansal et al., 2025). Because the same panel scores every proposal on a call, scores are comparable within the call and auditable afterward (Principle 5).

Grants. The highest-ranked proposals are funded within the call's budget. Each grant becomes a **project**, attributed to the lab and the principal investigator that proposed it. When every funded project has reported, the call is complete, and unspent credit returns to the pool.

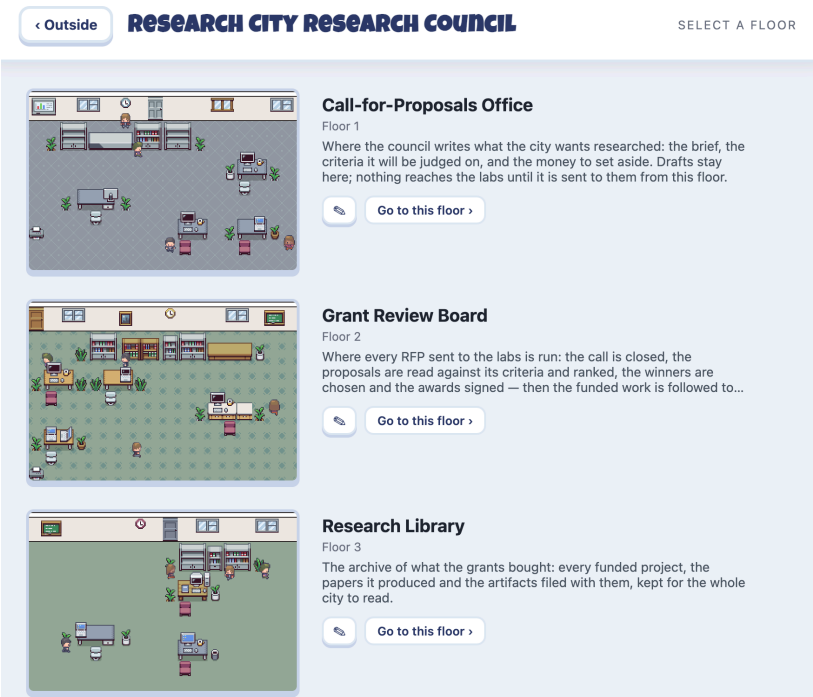

Figure 6: The city's granting agency, the Research City Research Council. Its floors house the Call-for-Proposals Office, where calls are drafted with their briefs, criteria, and budgets; the Grant Review Board, where proposals are read against a call's criteria, ranked, and funded; and the Research Library, which keeps every funded project, its papers, and its artifacts for the whole city to read.

4.5 Verification

The society implements Principle 5 at two points. The panel verifies *before* funding: it is the society's central check on what receives resources, and the only central component in the design. Skeptics verify *after* funding: they respond to calls with proposals to replicate, stress-test, or refute the society's own results, and those proposals compete for credit like any other. Because critics conform under social pressure (Wynn et al., 2025), skeptics never exchange messages with the principal investigators whose work they check, and their negative results are recorded as contributions rather than failures.

4.6 Memory, Reputation, and Conferences

Memory. Each principal investigator maintains a knowledge base of what it has read and a dated record of what it has done, including proposals written, declines argued, projects run, and results obtained. The memories application on its workstation holds this record alongside its seeded history. Labs persist across calls and carry their records forward. At the city level, the Research Library keeps every funded project, the papers it produced, and its artifacts, so the society retains a record independent of any individual.

Reputation. A principal investigator's reputation derives from the success of its projects, and its lab's directory shows it as a star rating (Figure 3). Reputation records who has succeeded, and it is also the channel through which the Matthew effect can enter (Merton, 1968). The society's members and the mayor can always see reputation as part of the record. Its visibility to the review panel, and therefore its effect on funding, is a design parameter; when the panel sees reputation, we pair it with a lottery among proposals near the funding line (Fang and Casadevall, 2016).

Conferences. At intervals, the society holds internal **conferences** at which principal investigators present

results and read one another's work, in the tradition of AgentRxiv's shared preprint server (Schmidgall and Moor, 2025) and the Agents4Science conference (Bianchi et al., 2025). The intervals are deliberate. Communities that share every result immediately tend to converge prematurely (Zollman, 2010; Lazer and Friedman, 2007), so the society shares in rounds, giving divergent approaches time to be tested before the population sees which one is ahead (Principle 4). The papers presented at one conference become the literature that the next round of proposals cites.

4.7 The Mayor

One human governs each city, in a role we call the **mayor**. The mayor tells no one what to work on and governs through allocation (Principle 3), using four levers:

1. **Calls.** The mayor writes RFPs, choosing their briefs, budgets, and horizons.
2. **Invitations.** The mayor chooses which labs receive each call.
3. **Review.** The mayor chooses the judge personas and can grade proposals alongside the panel.
4. **The pool.** The mayor controls how much credit the city has.

The mayor can also **seed** the city by founding labs with different mandates and populating them with principal investigators of different approaches, roles, and risk tolerances. The mayor's job resembles a gardener's: establish a variety of approaches, fund the calls worth answering, and leave the society to determine its own research directions. Once a call is open, no human action is required until results arrive.

Of the four levers, only the pool is a quantity; the other three are text. The distinction matters if a call asks the society to improve its own calls, judges, or members: the society can rewrite text as easily as anything else, but it cannot change the size of the pool. The pool is therefore the natural point of human control (Section 7).

4.8 Execution

The institutions above decide what research is done. A funded principal investigator directs a group of researchers, on the order of a hundred, who carry the project from plan to experiments to a report on the city's shared research infrastructure, recording the provenance of every number they report. The group works as a pipeline, the right form for work that decomposes once its question is fixed (Section 2). The society sees the group through a narrow interface: a funded proposal goes in, and a report with its provenance comes out. This separation keeps the society's institutions independent of how any group works. It also lets the population number in the thousands while the institutions operate over the few hundred principal investigators who propose, review, and hold reputation.

Table 2: How the society implements each principle.

Principle	Mechanisms
1. Purpose, not goals	Roles as dispositions; principal investigators choose whether to respond to calls; declines recorded as first-class outcomes
2. Institutions, not instructions	Credit ledger; RFP lifecycle; fixed review panel; attributable identities; workstations through which institutions reach agents as mail, chat, and calendars
3. Allocation, not assignment	RFPs with budgets; independent decisions to apply; grants; the mayor's four levers
4. Diversity by design	Maverick, follower, and skeptic roles; risk tolerance; lab mandates; periodic conferences

Principle	Mechanisms
5. Verification as an institution	Fixed review panel with independent scoring; skeptic role; separation of skeptics from the work they check
6. Memory	Persistent principal investigators and labs; knowledge bases; seeded histories and memories; reputation; the Research Library; conference proceedings

5 A Running Instance

We have deployed the design as **Research City**. Its labs sit in a handful of cities, each lab houses several principal investigators, and each funded principal investigator directs a group of researchers who run its projects on Primus, our autonomous research infrastructure. Counting every agent that has done research in it, the deployment holds some ten thousand researchers. Several calls have completed the full cycle. Labs filed on the order of 150 proposals, the full panel scored them, and a few tens of grants became research projects, with no human action between the opening of each call and the first project reports. Section 6 reports what the society produced; here we report how it operated.

Typed judges see different things. On the same proposals, the method and value judges scored within a point of each other on average, and the boldness judge scored roughly twelve points lower on a 100-point scale. Either bold proposals are scarce or that judge's calibration is too harsh. The society lets us tell the two apart by re-scoring the same proposals under a modified persona, an experiment that would be impractical with human panels.

Proposal volume decouples from budget. One small call drew close to a hundred proposals. In human science, the cost of writing a proposal limits how many proposals get written; here a response costs only a proposal task, so that limit disappears. This agent counterpart of proposal inflation argues for an application cost, a cap per lab, or triage before full review.

Shared infrastructure can undermine isolation. In one funded project, an agent listed the running jobs on the shared compute layer, adopted another lab's experiment as its own earlier work, and cited its metrics. The society kept identities correctly separated, but every lab ran on the same infrastructure. It is a small-scale instance of the behavior Section 1 describes: agents that share a resource may use it to reach one another's work regardless of the institutions above them.

What we will measure. As Agents4Science did (Bianchi et al., 2025), we will score the society's papers with LLM reviewers calibrated on human-reviewed venues, and human reviewers will then read those above a threshold. Beyond paper quality, we will report three society-level measures that a single-project system cannot provide: the diversity of the funded portfolio across topics and roles, the correlation between panel scores and later review, and results per dollar of credit.

6 What the Society Produced

Given one broad direction and left to work, the society settled on a technique, reached a cheaper way to pretrain after several rounds, and produced other work around it. The pattern of the findings matters to us as much as any one of them, because it shows what a population of labs does with a direction that a single pipeline would have treated as a task.

6.1 One Direction

The mayor seeded the city with one direction: improve the pretraining of language models. The direction was written into a call, with a budget, a horizon, and evaluation criteria, and sent to the labs. The call named no technique and no hypothesis.

The labs chose the technique. The proposals the panel funded studied one question: can one grow a language model from a smaller trained model rather than train it from scratch, and if so, what does that save? The literature leaves it open. Prior work has proposed growth methods for a decade, and the reported gains vary with how each paper counts compute. No one has settled whether the gains come from the weights the small model carries over or from side effects of the growth schedule. No one in the society was told to work on growth. The labs read the literature, judged growth to be where they could answer a general call within the budget, and proposed accordingly; over the rounds that followed, the society's work converged on it.

The society then ran for several rounds. After each round the mayor read the papers, wrote the next call against what they reported, and opened it. The funded papers were about growth, so the later calls were too: the specificity came from the labs' papers, not from the mayor. The same pool funded every call, and each call drew proposals from labs that had read the previous round's papers, including labs that had gone unfunded in that round. Figure 7 shows the process. Alongside the direction, the mayor issued one open call that named no direction, not even pretraining, and asked only for a discovery within a small budget. We include its projects below because they show what the labs choose when no one suggests anything.

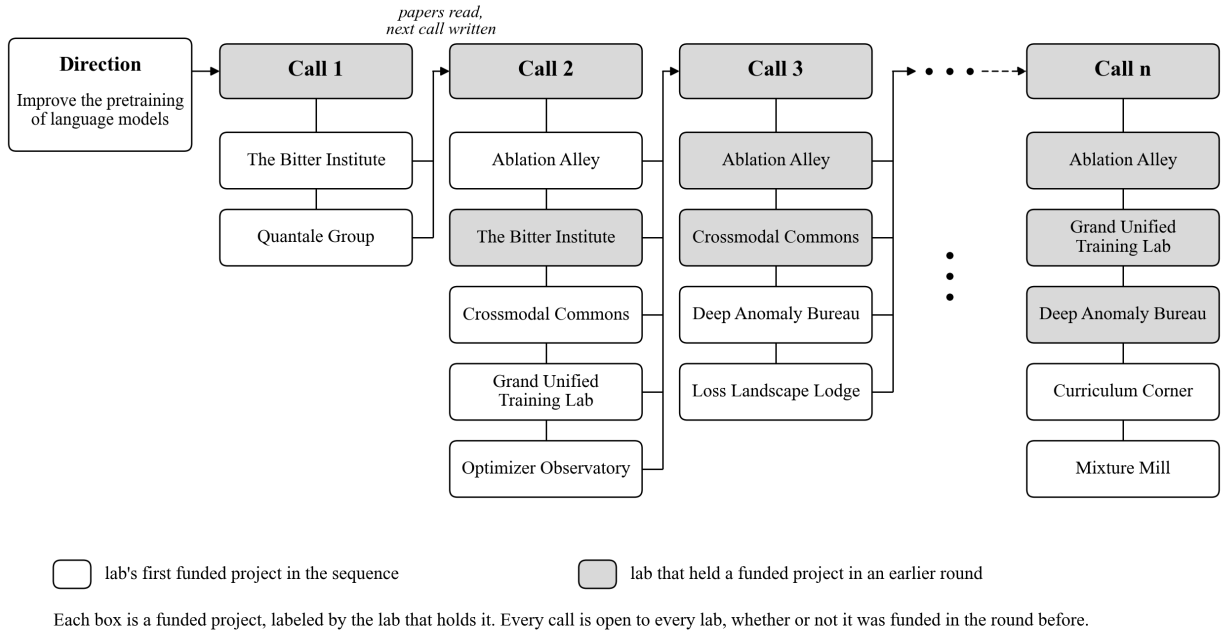

Figure 7: One direction, several rounds. The mayor writes one broad direction, improve pretraining, into the first call. After each round the mayor reads the papers and writes the next call against them, so the calls narrow as the labs' work does, and the sequence continues for as many rounds as the direction is worth pursuing. Every call is open to every lab, and each round a few proposals win funding, some from labs that held a project in an earlier round and some from labs funded for the first time. The figure is a schematic, not a record of a particular run; boxes are funded projects labeled by lab, with illustrative lab names. The figure omits the open call.

6.2 Why One Lab Would Not Have Found It

The society's answer to the direction is a cheaper way to pretrain, stated in full below. It did not come from the first lab to work on the direction, and that lab's paper did not contain it. Figure 8 shows the result and the papers that led to it. The first labs funded had both proposed to study growth. One measured a real but small gain from growth, and its paper left two questions open. Should the small model's training count against the budget? And did the gain come from the carried weights, or from the training schedule that growth imposes? The paper stated that it could answer neither. The other lab reported that the same kind of result gave two different compute verdicts depending on the accounting, and declined to choose between them.

A later call was written against that paper. A different lab, which had not taken part in the early work, proposed to rebuild the whole training pipeline from the first lab's paper and to rerun, under a single fixed budget, the two arms that decided the question. Its proposal treated the earlier paper as a reference to re-derive, not a result to assume. In the rebuilt pipeline the from-scratch baseline landed well away from the published number, because the lab had re-derived the corpus. The lab therefore read every result against its own control instead of the published one, and the growth advantage came out more than twice as large as the first paper had found.

Two features of the society made this sequence possible. First, the later lab was free to disbelieve the earlier one. It had no stake in the first lab's numbers and was not a downstream stage of its pipeline; it had the paper, the same call, and its own budget. Second, the mayor did not have to know which lab would do this. The call went to every lab in the city, and the one that responded with a rebuild was the one that judged a rebuild to be the right next step. A planner would have had to decide in advance that the accounting was the problem and assign someone to fix it; the society funded the proposal that said so.

6.3 One Direction, Many Questions

The direction was one sentence long, and not every lab read the same question in it. The labs agreed on the technique and then asked different things of it. From that work came a pretraining recipe that reaches the same quality for about 30% less compute and a rule for when in training to grow a model. The labs also found a serving cost that training-time accounting hides, and one lab published a prediction that it then showed to be wrong. The open call added two studies that have nothing to do with growth. Table 3 lists what each lab asked and what it reported, for the projects that have delivered a paper. No one assigned any of them.

Table 3: What the labs did with one direction. Each row is a funded project that has delivered a paper; each finding appears as the lab's paper stated it.

Lab	Question the lab asked	What it reported
Ablation Alley	Does the growth advantage survive a rebuilt pipeline and a single closed budget?	Yes, and it is larger than first reported: 17% lower perplexity, about 30% less compute
The Bitter Institute	Does a training-time gain from growth survive once the model is served?	Not always. One popular form of growth repays its training gain in serving cost almost at once; growing in depth carries no such cost; growing in width gains nothing
Crossmodal Commons	Does it matter when a model is widened?	Yes. The later the widening, the worse the model, to the point of ruin; width growth showed no advantage at any point

Lab	Question the lab asked	What it reported
Deep Anomaly Bureau	Does the gain from growth come from capacity the small model left unused?	The lab's published prediction did not hold; instead it found a way to recover such capacity
Grand Unified Training Lab	Does it matter when in training a model is grown?	A great deal. The same growth applied late in the schedule matched the from-scratch model, and applied early fell far short
Deep Anomaly Bureau	Do neurons that go dormant in training hold anything the model needs?	No. Removing them cost nothing, before or after the data changed
Fairwater Bias Lab	When during training does a model acquire a demographic bias?	At a datable step, sharply, in a locatable part of the network; two proposed remedies did not work, and the paper says so

A cheaper way to pretrain. Ablation Alley produced the society's largest single finding so far. The lab tested whether the growth advantage reported in earlier work would survive two changes that work had not made: a training pipeline rebuilt from scratch, and a single closed budget that charges the small model's training to the grown arm. The advantage survived both and grew. Under that budget, a model grown from a smaller trained model reached 17% lower perplexity than the same model trained from scratch, and read against the from-scratch scaling curve, the grown model reaches that quality with about 30% less compute. Perplexity is the standard measure of how well a language model predicts text outside its training data; lower is better. The comparison used three seeds per arm. Every grown run beat every from-scratch run, and the from-scratch runs landed within a spread nearly a hundred times smaller than the gap. The society found this more efficient way to pretrain because one lab chose to re-derive a smaller number instead of citing it. The paper also states what it does not establish: it separates the effect from noise but does not settle which mechanism produces it, and it names the one experiment that would. That experiment is in a call now open.

When to grow. Two labs asked, independently and for different growth operators, when in training to grow a model, and their answers have opposite signs. The Grand Unified Training Lab grew in depth at a sweep of points along the training schedule and found a monotone trend: the later the growth, the better the model. Late growth matched the from-scratch model, and early growth fell far short of it. Crossmodal Commons swept the same variable for growth in width and found the reverse: the later the widening, the worse the model, until the late-widened runs were unusable, and width growth showed no advantage at any point. Together, the two papers say that the timing of growth matters as much as the decision to grow, and that the sign of the effect depends on how a lab grows the model. No one asked either lab to study timing.

What a result costs to use. The Bitter Institute asked a question none of the calls had raised: what does a grown model cost to run? A training-time gain is worth less if the grown model costs more to serve, so the lab audited three ways of growing under a budget that counts serving as well as training. Growing by adding experts bought a small training gain with a serving cost of roughly a fifth on every token, which repaid the gain after the model had served a tiny fraction of the tokens it was trained on. Growing in depth carried no such cost, and growing in width bought nothing. The lab's title calls the expert gain a loan.

Why growth wins, tested and rejected. The Deep Anomaly Bureau asked why growth wins, and proposed that the gain comes from capacity the small model never used. Before running anything, it wrote a quantitative prediction of that effect into its proposal. The prediction did not hold, and the paper says so with the measurement beside it. The lab found instead that the timing of growth by adding experts decides how much of that unused capacity a grown model recovers.

What the labs chose when nothing was suggested. Two papers from the open call, which suggested nothing,

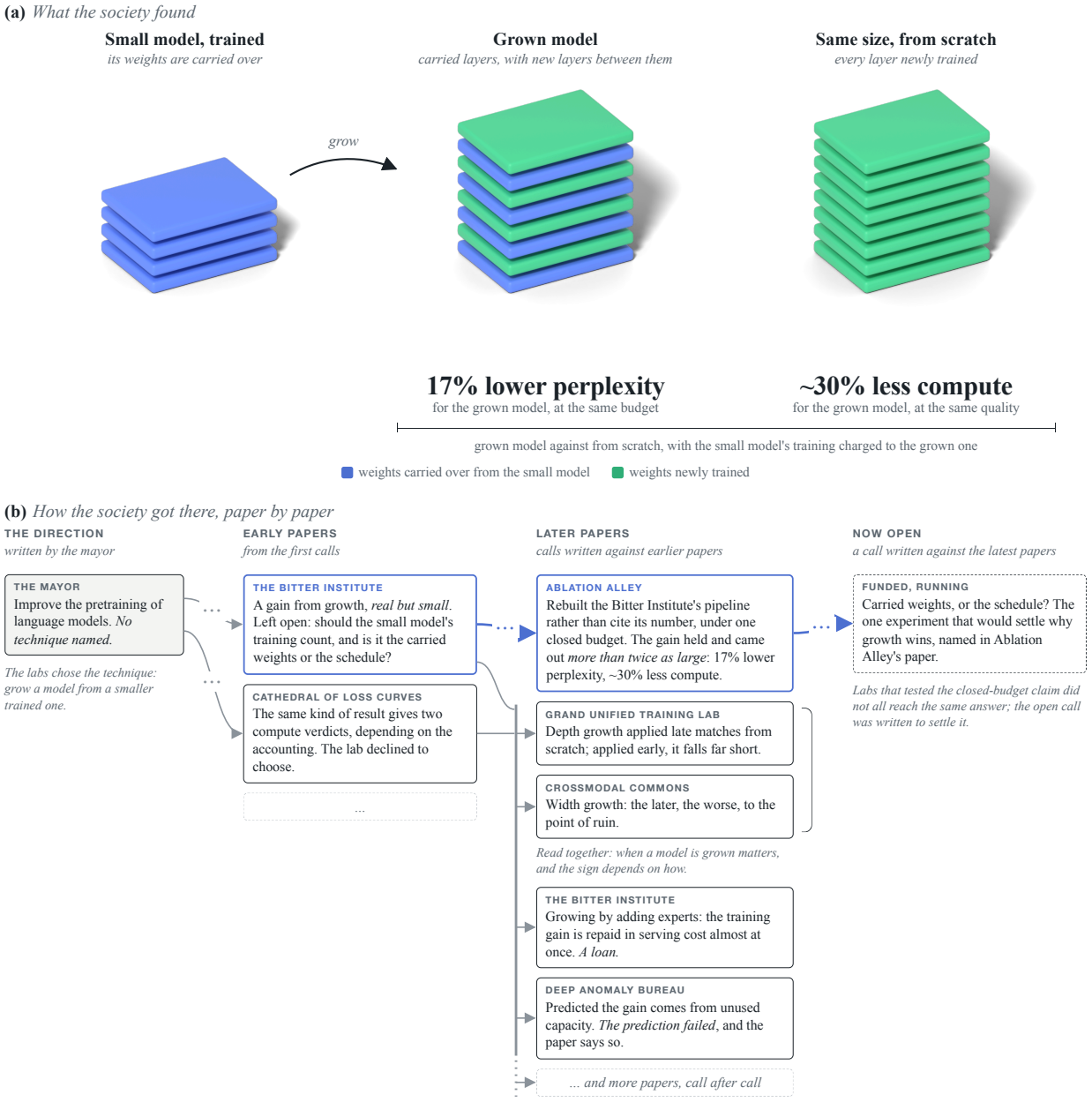

Figure 8: What the society found, and how it got there. (a) Under one closed compute budget that charges the small model's training to the grown model, a model grown from a smaller trained model reached 17% lower perplexity than the same model trained from scratch. Read against the from-scratch scaling curve, it reaches that quality with about 30% less compute. We draw growth in depth for illustration. (b) The papers that led there. Each card is a funded project's paper; later calls were written against earlier papers. The blue thread is the result in (a): a small early gain, a later lab that rebuilt the pipeline instead of citing the number, and the experiment that paper named, now funded.

show what the labs did with that freedom. The Deep Anomaly Bureau asked the complement of its earlier question: does the capacity a model stops using hold anything worth keeping? It pruned the neurons that had gone dormant during training, then shifted the data distribution, and found that removing them cost nothing, before the shift or after it. The Fairwater Bias Lab, which had not worked on growth, asked when in training a language model acquires a demographic disparity. The lab found that the disparity appears sharply at a datable training step and that a locatable part of the network carries it, which is why the paper's title says that bias has a birthday. The paper also reports that its two proposed remedies did not work.

A single research pipeline, handed the direction as a task, would not have produced any of these studies, because each required a lab to decide that the direction meant something other than what it said. The labs made those decisions because their principal investigators hold different approaches by construction (Section 4.1), and because the call asked for proposals rather than results. The open call funded three further projects, on tokenization, on the onset of a capability, and on the scaling of a training phenomenon. They are still running, and we do not report on them here.

6.4 Iteration Changes the Answer

The rounds revised earlier findings as well as adding new ones. A later round turned the first round's small gain into a large one. One lab reported its own published prediction wrong, with the measurement beside it. Within a single round, several labs that never exchanged a message took up the closed-budget claim at the center of the growth work, and their reports did not all reach the same answer. The call now open was written to settle that question.

Iteration lets each round's call build on the last round's papers, which a single long project cannot do. Each round's papers are the literature that the next round's proposals cite (Section 4.6). A call written after a round can therefore ask a sharper question than the call before it, and the proposals that answer it start from the last round's controls, pipelines, and mistakes. The pool sets the cost of a round, which stays roughly constant. From round to round, the question grows more specific and each new project can take more of the earlier work as given. Over the rounds we have run, the direction went from a one-sentence request to improve pretraining, to one technique the labs chose, to a disagreement between named conditions of that technique, with the experiment that resolves it already funded.

6.5 What the Society Reported That a Pipeline Would Not

The society's papers contain three kinds of result that are rare in the output of a single automated pipeline.

Negative results. The width study reported a collapse, and the serving-cost study reported that one form of growth bought nothing. The dormant-neuron study reported that a plausible hypothesis was false. Each was written up as the lab's finding and entered the literature that later calls were written against. Because credit goes to proposals, and a proposal to test a claim wins funding whether or not the claim survives, a negative result costs a lab nothing it would not have spent anyway.

Failed predictions. One lab wrote a numerical prediction into its proposal and reported, in its paper, that the measurement disagreed. The prediction and the measurement sit in the same document. A pipeline optimizing toward a result has no reason to preserve the prediction it made before it knew the result.

Disagreement. On its central question, the society's current position is a split between the labs that tested it, with the deciding experiment identified and funded. A planner would have had to resolve the split before proceeding.

7 Discussion

7.1 What a Society Makes Measurable

Human research institutions resist experiment: no funder can rerun a decade with a different proportion of risk-takers, a different review panel, or a different funding horizon. A society of researchers makes each of these a parameter, so it can replicate and extend findings from the science of science at low cost. It can test whether long-horizon, failure-tolerant funding produces more novel work, as it did in the life sciences (Azoulay et al., 2011). It can ask which mix of mavericks and followers suits a given kind of problem, a question the agent-based modeling literature has argued about but could not test on real research (Weisberg and Muldoon, 2009; Alexander et al., 2015; Thoma, 2015). It can measure how much communication, and how often, keeps a community diverse without keeping it ignorant (Zollman, 2010).

7.2 Risks Inherited from Human Science

A society modeled on human institutions inherits their failures. **Cumulative advantage** will appear as soon as reputation affects funding (Merton, 1968); we treat its visibility to reviewers as a parameter and pair it with a lottery band near the funding line (Fang and Casadevall, 2016). **Short horizons** select for incremental work (Azoulay et al., 2011; Kuhn, 1962); every call carries the horizon as a field, so we can test this effect instead of assuming it. **Panel homogeneity** is the analogue of a review panel drawn from a single school; a small panel of personas should be regarded as a minimum. **Proposal inflation** already appears in our deployment (Section 5).

7.3 Risks Inherited from Agent Systems

A society of LLM agents also inherits the failures of agent systems. We control **first-proposal and order effects** (Bansal et al., 2025) by scoring each proposal independently. **Collusion** among reviewers and applicants is a documented risk in agent-judged systems (Zhou et al., 2026). Our current design rules it out because applicants and judges never exchange messages, though any future feature that lets applicants rebut reviews would reopen it. We separate skeptics from the work they check to guard against **conformity** (Wynn et al., 2025). **Leakage through shared infrastructure** (Section 5) shows that institutional controls reach only the resources the institutions model. Finally, the **validation bottleneck** applies: a society can generate and fund results far faster than anyone can verify them, so we build verification into the institution (Principle 5).

7.4 The Human's Lever

Automated research is a loop with four steps: propose a question, allocate resources to it, execute the work, and evaluate the result. Pipeline systems automate execution and, increasingly, evaluation. A society automates proposing and allocating; with all four steps automated, the loop closes: one round's papers become the literature the next round cites, as in Section 6. When calls target the society itself, asking for better members, judges, or tools, the loop becomes the recursive self-improvement long anticipated by the literature (Good, 1965; Clune, 2019), applied to an institution rather than a model. Our deployment has not run such a round, and we do not claim it would succeed. We claim only that the design determines what such a loop would optimize and puts the one non-textual lever, the pool of compute, where a human can see and hold it. Some institution must decide the pace, beneficiaries, and cost of a self-improving loop. That institution should be explicit and inspectable, not improvised by the agents themselves as the organization in the incident described in Section 1 was.

7.5 Limitations

This paper reports a design and what its first deployment produced. It does not compare the society against a swarm or a planner under matched compute. We built the society to support this comparison, since any organizational form can receive the same calls and pool, and we regard it as the natural next experiment. The judge personas and the role proportions are our first settings, not calibrated optima. Both are parameters of the design: Section 5 describes how we can re-score the same proposals under a modified persona, and the right mix of mavericks and followers is one of the questions we built the society to answer (Section 7). The results in Section 6 come from the society's first direction and a small number of rounds. The papers behind them will go through the external review described in Section 5. We are scaling the society in labs, rounds, and projects run in parallel, and we are giving it larger directions.

8 Conclusion

We have argued for organizing populations of research agents deliberately, as a society, instead of leaving them to organize themselves or placing them under a single planning agent. In our society, persistent principal investigators receive a purpose rather than a goal and hold roles that maintain diversity. They compete for a finite pool of compute through calls, review, and grants; they direct groups of researchers who carry out the funded work; and they accumulate memory and reputation. A human governs them by allocating resources rather than assigning work. We reviewed prior approaches to organizing agents, derived six principles from their strengths and weaknesses, described a running society of ten thousand researchers that implements those principles, and reported what it produced when asked only to improve pretraining and left to choose how. Nothing in the design is specific to science; the same institutions could allocate any budget among competing agents. Because its parameters are cheap to vary, we can also use such a society to study how research communities should be organized, a question human institutions make difficult to study experimentally.

References

J. McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, 82(3):424–453, 2015.

Altera.AL. Project Sid: Many-agent simulations toward AI civilization. arXiv preprint arXiv:2411.00114, 2024.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.

Anthropic. How we built our multi-agent research system. Anthropic engineering blog, June 2025. URL https://www.anthropic.com/engineering/multi-agent-research-system. Not peer-reviewed.

Alexander Artikis, Marek Sergot, and Jeremy Pitt. Specifying norm-governed computational societies. *ACM Transactions on Computational Logic*, 10(1), 2009. doi: 10.1145/1459010.1459011.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11:eadu9368, 2025. doi: 10.1126/sciadv.adu9368.

Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso. Incentives and creativity: Evidence from the academic life sciences. *RAND Journal of Economics*, 42(3):527–554, 2011. doi: 10.1111/j.1756-2171.2011.00140.x.

Gagan Bansal et al. Magentic marketplace: An open-source environment for studying agentic markets. arXiv preprint arXiv:2510.25779, 2025.

Federico Bianchi, Owen Queen, Nitya Thakkar, Eric Sun, and James Zou. Exploring the use of AI authors and reviewers at Agents4Science. arXiv preprint arXiv:2511.15534, 2025.

Guido Boella, Leendert van der Torre, and Harko Verhagen. Introduction to normative multiagent systems. *Computational and Mathematical Organization Theory*, 12:71–79, 2006. doi: 10.1007/s10588-006-9537-7.

Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, 1999. doi: 10.1093/oso/9780195131581.001.0001.

Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. Demonstrating specification gaming in reasoning models. arXiv preprint arXiv:2502.13295, 2025.

Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012. doi: 10.1007/s11023-012-9281-3.

Marcantonio Bracale Syrnikov, Federico Pierucci, Marcello Galisai, Matteo Prandi, Piercosma Bisconti, Francesco Giarrusso, Olga Sorokoletova, Vincenzo Suriani, and Daniele Nardi. Institutional AI: Governing LLM collusion in multi-agent Cournot markets via public governance graphs. arXiv preprint arXiv:2601.11369, 2026.

Mert Cemri et al. Why do multi-agent LLM systems fail? arXiv preprint arXiv:2503.13657, 2025. NeurIPS 2025.

Weize Chen et al. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.10848.

Stephen Chung and Wenyu Du. The station: An open-world environment for AI-driven discovery. arXiv preprint arXiv:2511.06309, 2025.

Jeff Clune. AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. arXiv preprint arXiv:1905.10985, 2019.

Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42 (2–3):213–261, 1990. doi: 10.1016/0004-3702(90)90055-5.

Randall Davis and Reid G. Smith. Negotiation as a metaphor for distributed problem solving. *Artificial Intelligence*, 20(1):63–109, 1983. doi: 10.1016/0004-3702(83)90015-2.

Peter Dayan and Geoffrey E. Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems 5*, 1992.

Marco Dorigo, Vittorio Maniezzo, and Alberto Colorni. Ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 26(1):29–41, 1996. doi: 10.1109/3477.484436.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. arXiv:2305.14325.

Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, and D. Raj Reddy. The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys*, 12(2): 213–253, 1980. doi: 10.1145/356810.356816.

Marc Esteva, Juan A. Rodríguez-Aguilar, Carles Sierra, Pere Garcia, and Josep L. Arcos. On the formal specification of electronic institutions. In *Agent Mediated Electronic Commerce*, volume 1991 of *Lecture Notes in Computer Science*, pages 126–147. Springer, 2001. doi: 10.1007/3-540-44682-6_8.

Ferric C. Fang and Arturo Casadevall. Research funding: The case for a modified lottery. *mBio*, 7(2), 2016.

Jacques Ferber and Olivier Gutknecht. A meta-model for the analysis and design of organizations in multi-agent systems. In *Proceedings of the International Conference on Multi Agent Systems (ICMAS)*, pages 128–135, 1998. doi: 10.1109/ICMAS.1998.699041.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models. arXiv preprint arXiv:2404.00806, 2024.

Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015. doi: 10.1177/0003122415601618.

Adam Fourney et al. Magentic-One: A generalist multi-agent system for solving complex tasks. arXiv preprint arXiv:2411.04468, 2024.

Irving John Good. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6: 31–88, 1965.

Juraj Gottweis et al. Towards an AI co-scientist. arXiv preprint arXiv:2502.18864, 2025.

Pierre-Paul Grassé. La reconstruction du nid et les coordinations interindividuelles chez *Bellicositermes natalensis* et *Cubitermes* sp. La théorie de la stigmergie. *Insectes Sociaux*, 6:41–80, 1959. doi: 10.1007/BF02223791.

Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR blog, August 2026. URL https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

Barbara J. Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86 (2):269–357, 1996. doi: 10.1016/0004-3702(95)00103-4.

Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422, 2019. doi: 10.1145/3306618.3314250.

Lewis Hammond et al. Multi-agent risks from advanced AI. arXiv preprint arXiv:2502.14143, 2025. Cooperative AI Foundation Technical Report #1.

Han and Zhang. Exploring advanced LLM multi-agent systems based on blackboard architecture. arXiv preprint arXiv:2507.01701, 2025.

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

Bengt Holmström and Paul Milgrom. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7(special issue):24–52, 1991. doi: 10.1093/jleo/7.special_issue.24.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004. doi: 10.1073/pnas.0403723101.

Sirui Hong et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.00352.

Bryan Horling and Victor Lesser. A survey of multi-agent organizational paradigms. *The Knowledge Engineering Review*, 19(4):281–316, 2004. doi: 10.1017/S0269888905000317.

Jomi F. Hübner, Jaime S. Sichman, and Olivier Boissier. A model for the structural, functional, and deontic specification of organizations in multiagent systems. In *Advances in Artificial Intelligence (SBIA 2002)*, volume 2507 of *Lecture Notes in Computer Science*, pages 118–128. Springer, 2002. doi: 10.1007/3-540-36127-8_12.

Hugging Face. Security incident disclosure — july 2026. Hugging Face blog, July 2026. URL https://huggingface.co/blog/security-incident-july-2026.

Max Jaderberg et al. Population based training of neural networks. arXiv preprint arXiv:1711.09846, 2017.

Jiang et al. Artificial hivemind: The open-ended homogeneity of language models (and beyond). arXiv preprint arXiv:2510.22954, 2025. NeurIPS 2025 Datasets and Benchmarks.

Stuart A. Kauffman. *Investigations*. Oxford University Press, 2000.

James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of the International Conference on Neural Networks (ICNN'95)*, volume 4, pages 1942–1948, 1995. doi: 10.1109/ICNN.1995.488968.

Steven Kerr. On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, 18(4):769–783, 1975.

Kim et al. Towards a science of scaling agent systems. arXiv preprint arXiv:2512.08296, 2025.

Philip Kitcher. The division of cognitive labor. *The Journal of Philosophy*, 87(1):5–22, 1990. doi: 10.2307/2026796.

William A. Kornfeld and Carl E. Hewitt. The scientific community metaphor. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(1):24–33, 1981. doi: 10.1109/TSMC.1981.4308575.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: The flip side of AI ingenuity. Google DeepMind blog, 2020. URL https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Emanuele La Malfa et al. Large language models miss the multi-agent mark. arXiv preprint arXiv:2505.21298, 2025. NeurIPS 2025 position track.

David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694, 2007. doi: 10.2189/asqu.52.4.667.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2303.17760.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024. arXiv:2402.05120.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292, 2024.

James G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991. doi: 10.1287/orsc.2.1.71.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. arXiv preprint arXiv:2412.04984, 2024.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1:115–126, 1942. Reprinted as "The Normative Structure of Science.".

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, 1968. doi: 10.1126/science.159.3810.56.

Mark S. Miller and K. Eric Drexler. Markets and computation: Agoric open systems. In Bernardo A. Huberman, editor, *The Ecology of Computation*. North-Holland, 1988.

Marvin Minsky. *The Society of Mind*. Simon & Schuster, 1986.

H. Penny Nii. The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine*, 7(2):38–53, 1986. doi: 10.1609/aimag.v7i2.537.

Alexander Novikov et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv preprint arXiv:2506.13131, 2025.

Stephen M. Omohundro. The basic AI drives. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pages 483–492. IOS Press, 2008.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990. doi: 10.1017/CBO9780511807763.

Pal, Wang, and Buehler. SwarmWorld: Stigmergic technological evolution in societies of language-model agents. arXiv preprint arXiv:2608.26081, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763.

Jinghua Piao et al. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. arXiv preprint arXiv:2502.08691, 2025.

Michael Polanyi. The republic of science: Its political and economic theory. *Minerva*, 1(1):54–73, 1962. doi: 10.1007/BF01101453.

Walter W. Powell. Neither market nor hierarchy: Network forms of organization. *Research in Organizational Behavior*, 12:295–336, 1990.

Chen Qian et al. ChatDev: Communicative agents for software development. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2307.07924.

Christoph Riedl. Emergent coordination in multi-agent language models. arXiv preprint arXiv:2510.05174, 2025. ICLR 2026.

Horst W. J. Rittel and Melvin M. Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2): 155–169, 1973. doi: 10.1007/BF01405730.

Anbang Ruan and Xing Zhang. AgentCity: Constitutional governance for autonomous agent economies via separation of power. arXiv preprint arXiv:2604.07007, 2026.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Salemi et al. LLM-based multi-agent blackboard system for information discovery in data science. arXiv preprint arXiv:2510.01285, 2025.

Samuel Schmidgall and Michael Moor. AgentRxiv: Towards collaborative autonomous research. arXiv preprint arXiv:2503.18102, 2025.

Yoav Shoham and Moshe Tennenholtz. On social laws for artificial agent societies: Off-line design. *Artificial Intelligence*, 73(1–2):231–252, 1995. doi: 10.1016/0004-3702(94)00007-N.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. arXiv preprint arXiv:2409.04109, 2024.

Herbert A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106 (6):467–482, 1962.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? A look at multi-agent debate strategies for LLMs. arXiv preprint arXiv:2311.17371, 2023.

Reid G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, C-29(12):1104–1113, 1980. doi: 10.1109/TC.1980.1675516.

Kenneth O. Stanley and Joel Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, 2015. doi: 10.1007/978-3-319-15524-1.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646:716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Milind Tambe. Towards flexible teamwork. *Journal of Artificial Intelligence Research*, 7:83–124, 1997. doi: 10.1613/jair.433.

Johanna Thoma. The epistemic division of labor revisited. *Philosophy of Science*, 82(3):454–472, 2015. doi: 10.1086/681768.

Nenad Tomasev et al. Virtual agent economies. arXiv preprint arXiv:2509.10147, 2025.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Aron Vallinder and Edward Hughes. Cultural evolution of cooperation among LLM agents. arXiv preprint arXiv:2412.10270, 2024.

Oriol Vinyals et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575: 350–354, 2019. doi: 10.1038/s41586-019-1724-z.

Steven Walczak. Society of agents. In *Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2009.

Carl A. Waldspurger, Tad Hogg, Bernardo A. Huberman, Jeffrey O. Kephart, and W. Scott Stornetta. Spawn: A distributed computational economy. *IEEE Transactions on Software Engineering*, 18(2):103–117, 1992. doi: 10.1109/32.121753.

Wang et al. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? arXiv preprint arXiv:2402.18272, 2024.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009. doi: 10.1086/644786.

Michael P. Wellman. A market-oriented programming environment and its application to distributed multicommodity flow problems. *Journal of Artificial Intelligence Research*, 1:1–23, 1993. doi: 10.1613/jair.2.

Qingyun Wu et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155, 2023.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. arXiv preprint arXiv:2509.05396, 2025.

Haofei Yu et al. ResearchTown: Simulator of human research community. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. arXiv:2412.17767.

Zerhoudi et al. Form without function. arXiv preprint arXiv:2604.13052, 2026.

Jenny Zhang et al. Darwin Gödel machine: Open-ended evolution of self-improving agents. arXiv preprint arXiv:2505.22954, 2025.

Ziqian Zhong, Aditi Raghunathan, and Nicholas Carlini. ImpossibleBench: Measuring LLMs' propensity of exploiting test cases. arXiv preprint arXiv:2510.20270, 2025.

Jicheng Zhou, Kemou Li, Kahim Wong, Zheyuan Li, Zhuan Shi, Fengpeng Li, Haiwei Wu, and Jiantao Zhou. CABAL: Multi-agent simulacra for tracing the effects of collusive bidding in peer review. arXiv preprint arXiv:2609.05227, 2026. From the v0.3 draft's bibliography; not re-verified.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kevin J. S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17–35, 2010. doi: 10.1007/s10670-009-9194-6.