Safe AI Should be Bounded and Multi-Agent

David Hyland[1,*] **Daniel Jarne Ornia**[2,*] **Nicholas Bishop**[1,*] **Joel Dyer**[2]
Olivia Macmillan-Scott[3] **Tomáš Gavenčiak**[4] **Anisoara Calinescu**[1]
Michael J. Wooldridge[1] **Fernando E. Rosas**[1,5,6,†] **Pedro A. Ortega**[7,†]

[1]University of Oxford [2]Independent Researcher [3]University College London
[4]Charles University [5]University of Sussex [6]Imperial College London [7]Daios Technologies

Abstract

Major developments in frontier AI systems over the last decade have been driven by the *scaling paradigm*, which treats resource constraints as key obstacles to be overcome in the pursuit of more capable systems. Here, we argue for a complementary paradigm that embraces these constraints — together with the multi-agent, distributed nature of real-world deployments — as a route towards safe and scalable AI. Rather than scaling individual agents alone, we posit that legibly composing agents while deliberately bounding their capabilities, affordances, and resource budgets can reliably yield system-level competence. We call such systems *bounded multi-agent systems* (BMAS). Our position is that **bounded agency should be a foundational principle for scaling towards safe, robust, and equitable AI**. This motivates a research agenda to formally characterise bounded agency, design legible interfaces and institutions for agent ecosystems, and evaluate when bounded modular systems are more appropriate than monolithic systems.

1 Introduction

The scaling paradigm has been the defining success story of modern AI. In this paradigm, mainstream AI developers have primarily treated resource constraints as undesirable obstacles to be overcome. By systematically expanding model capacity, data, training compute, inference-time compute, context length, and tool access, this strategy has yielded extraordinary returns, including AI systems that match or surpass human performance across a broad range of tasks. However, the same drive towards increasingly general, highly capable agents has raised critical concerns, particularly about their safety [1–6]. For example, advanced AI models have demonstrated the capability to covertly pursue misaligned goals [7], engage in deceptive behaviour [8–12], and subvert safeguarding mechanisms [13–15]. On top of this, it is far from guaranteed that these behaviours will diminish in severity or likelihood as scaling continues. The monolithic and black-box nature of many current models makes safety risks difficult to mitigate, and complicates alignment with human goals and values. Moreover, as advanced frontier models are typically large in scale, they require significant energy consumption and resource-intensive infrastructure, raising concerns about their sustainability.

We believe that the aforementioned challenges stem from a subtly misguided approach to understanding the nature of agentic systems, which treats agents' limitations as problems to be overcome rather than features to be engineered. Although significant research effort has gone into understanding and designing systems that can scale in terms of data inputs, model capacity, and inference-time compute to arbitrary sizes while retaining performance, there has been relatively little exploration of how the limitations of agentic systems are related to our capacity to interpret and steer them. Crucially, we lack a deep understanding of how such features affect model capabilities, both desired and undesired.

[*]Equal contribution. david.hyland@cs.ox.ac.uk, daniel.jarne@gmail.com, nicholas.bishop@cs.ox.ac.uk
[†]Senior authors.

Preprint.

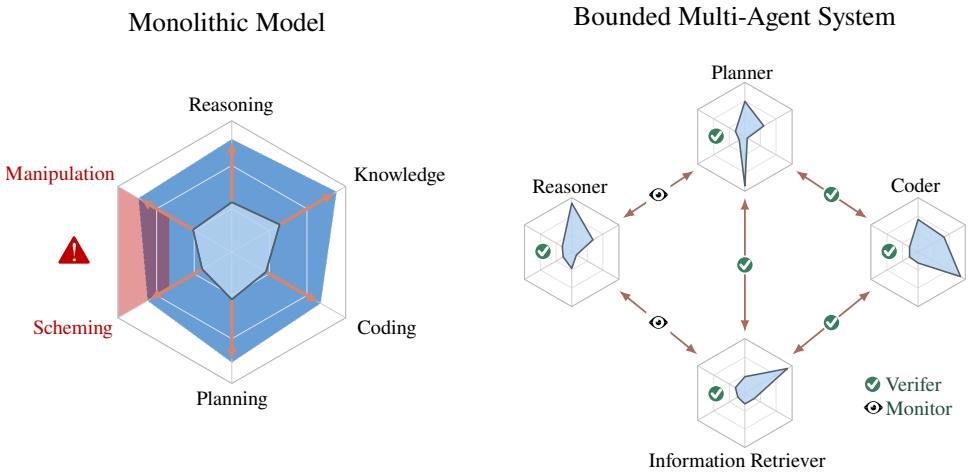

Figure 1: Example capability profiles for monolithic models and bounded multi-agent systems (BMAS). *Left:* a single monolith spans a broad volume of the capability surface, including harm-relevant manipulation and scheming axes, and is pushed outward on every axis under scaling. In such models, it can be difficult to limit potentially harmful capabilities without systematically degrading general capabilities. *Right:* a BMAS may decompose a complex task into a collaboration between specialised sub-agents, each capable mainly in its speciality. Each agent may be intentionally bounded along some axes, and agents are composed through interactions, which can be verified and monitored.

Fortunately, we have a wealth of existing systems to learn from, and we should take advantage of the opportunity to do so. Many naturally occurring and social intelligent systems, such as multicellular organisms, firms, scientific communities, markets, and regulatory bodies do not take the form of a monolithic agent with broad and centralised capabilities. Instead, they are *multi-agent systems* which exploit redundancy [16], specialisation [17], monitoring [18], institutions [19], and the division of labour [20]. Even if individual components are inherently limited, the collective can perform tasks that no component can perform alone [21–24]. This makes a strong case for a shift in focus from monolithic systems to multi-agent systems composed of bounded agents. We refer to such systems as *bounded multi-agent systems* (BMAS). Importantly, boundedness is not only a consequence of delegating computational resources and knowledge, but it is also a key factor in ensuring system efficiency, robustness, and safety. For example, if failure or harm can be attributed to responsible components through causal experiments and counterfactual evaluations [25, 26], targeted interventions can be applied without compromising overall system stability [27, 28]. BMAS extends a trajectory that is already visible in sparse experts [29], tool routing [30], and agent orchestration [31], but goes further by deliberately bounding capabilities themselves rather than only allocating them efficiently. The trade-off — reduced general capability per unit of compute — is already natural in many settings and is the price of being able to restrict harmful capabilities, isolate failures, and verify behaviour, which is well worth the cost.[1]

In addition to having several examples to learn from, humanity has a long and deep experience in designing institutions and mechanisms for regulating multi-agent systems, such as markets and ecosystems. In contrast, we have considerably less experience in interpreting and steering large, monolithic systems. This existing experience in the former provides an additional layer of resources to draw from in designing and governing BMAS to ensure alignment at multiple scales [32]. For example, social choice theory and mechanism design can be applied to a multi-agent system in a manner that is impossible for any monolithic equivalent [33], and game-theoretic learning algorithms can be used to promote convergence [34–36] and steer agents towards particularly desirable equilibria [37–39]. Moreover, the potential efficiency benefits of multi-agent systems are becoming increasingly clear, as highlighted by design patterns such as Mixture-of-Experts (MoE) models [29] and coordination protocols for generative AI agents such as A2A [40]. In summary, boundedness and multi-agency form an effective combination that can regulate behaviour at both the individual and system level.

[1] Figure 1 illustrates some differences between monolithic models and BMAS.

For these reasons, we argue that the AI community should treat resource constraints as a *feature*, rather than as a limitation. More concretely, the AI community should develop methods for understanding and building models with targeted capabilities, as well as a mechanistic theory for deploying and coordinating such agents, whose capabilities are carefully shaped to specific but restricted task domains. Furthermore, we claim that multi-agent systems of specialised but constrained agents provide an inherently safer paradigm for scaling AI than monolithic systems, which are a central point of failure from safety, robustness, and alignment perspectives.

Position

Bounded agency should be a foundational principle for scaling towards safe, robust, and equitable AI.

2 Boundedness and Multi-Agency

Current frontier models that drive agentic systems face binding resource constraints, or significant trade-offs between goal achievement and resource use. These trade-offs appear as imperfect decisions, limited planning, incomplete knowledge, restricted tool use, or failures to adapt across tasks. To ground the rest of the paper, we introduce operational[2] definitions of *worlds*, *agents*, and *goals*.

A *world* is the environment in which agents act, which includes the agents themselves [41–43]. A world contains variables agents can observe, latent variables, actions agents can take, and a (probabilistic) generative model that governs observables and latents conditional on actions taken. An *agent* is a process that uses observations, internal states, computational processes, and resources to select and take actions in pursuit of goals. This definition is intentionally broad enough to include reinforcement learning agents, tool-using language model agents, autonomous software services with delegated objectives, and humans. Finally, a *goal* is represented by (a set of) preferred possible outcomes or histories of interaction; some trajectories may be more desirable than others.

Agent Components Agents differ not only in their goals, but also in what they can observe, memorise, compute, and do. We can therefore describe an agent in terms of its internal state, update maps, parameters, and architecture. These features determine how the agent represents the world, how it updates that representation over time, and how it chooses actions. In this sense, an agent is not just a policy mapping observations to actions, but a decision-making system with concrete informational, computational and implementation limitations. It is useful to see an agent as consisting of two main sub-systems: an **internal representation** and an **action-selection process**. The internal representation compresses the agent's observed history into a usable form; the action-selection process uses this representation, along with the agent's resources, to select an action to execute.

2.1 Bounded Agents

An agent is **bounded** with respect to a set of goals when some constraint on its information, architecture, resources, or affordances prevents it from perfectly attaining all goals in that set, or from optimising them without cost. An agent may be effectively unbounded for a small collection of simple goals, while being strongly bounded on a broader class of goals. For example, a calculator is effectively optimal for supported arithmetic operations, but bounded with respect to the wider class of mathematical problem-solving tasks, such as proving a theorem. Different agents can also be bounded in different respects. Some are limited mainly by what they have learned or stored; others are limited mainly by how much online compute (*reasoning*) they can perform.

Example — Chess Agents 1 *For the same world and goal, two agents can exhibit fundamentally different kinds of boundedness. In a game of chess, the observable state is the board position, while the opponent's strategy is latent. One agent might use a tabular lookup structure, storing a finite number of previously encountered board positions and the most successful move played from each position so far. Another agent might store a chess engine and use online rollouts to approximate the*

[2]We emphasise this term; the definitions provided are intentionally broad to allow for a comprehensive discussion around the concepts of agency, boundedness and rationality that are relevant to this paper. We do not claim these definitions as universal or exclusive.

best next move. Both agents act in the same world and pursue the same goal, but are bounded in different ways: one mainly by memory and the other by online compute.

Bounds can be manifest through *hard constraints*, such as a restricted action space, a limited context window, a fixed inference budget, a permission system, or a prohibition on external communication. They can also be *soft constraints*, such as impact penalties, computational costs, regularisers, or incentives to find satisficing solutions. Boundedness should be understood (for our purposes) as primarily a property of an agentic system relative to a family of goals and environments.[3] This is captured by an optimality gap: the agent falls short of what could be achieved by an ideal agent with unrestricted resources. This view allows us to connect boundedness to **narrow agency**. A narrow agent is not necessarily weak; it may be highly capable, or even effectively optimal, on a small region of goal space. What makes it narrow is that this region is small compared with a broader class of goals. In this sense, narrowness can be understood geometrically: the agent performs well on a small subset of possible goals, but these occupy a limited "volume" of a larger goal space. An idealised *general agent* would perform optimally across most or all of the relevant goal space [45, 46].

This notion of boundedness relates to a range of existing ideas in agency, sub-optimality, and utility theory, and was first captured formally by the theory of bounded rationality [47, 48]. An effective strategy in the face of bounded rationality is *satisficing*: selecting an option that meets an aspiration level shaped by resource limitations, rather than trying to globally optimise an objective [49]. A contrasting response is bounded optimality, which retains the optimisation ideal but relativises it to architectural constraints, finding an optimal program of execution given the limitations imposed by an agent's own architecture [50]. More recent frameworks model decision-making as a continuous trade-off between performance and resource consumption; these include computational rationality [51, 52], information-theoretic bounded rationality [53–55], and resource rationality [56–58].

Example — Chess Agents 2 *Removing limitations on specific components of an agent may have different effects. When the tabular lookup chess agent observes new data or is afforded more memory, it can refine its internal representation of the world. Meanwhile, its action-selection process is fixed and corresponds to a simple lookup operation. Thus, providing more compute at decision-time does not improve action selection. Conversely, the chess engine agent maintains a fixed internal representation that cannot be refined through data collection. However, providing more decision-time compute enables the agent to perform more and deeper rollouts, enabling it to select better moves.*

2.2 Bounded Multi-Agent Systems

A *multi-agent system* is a system comprised of multiple interacting agents. For our purposes, the salient cases are systems in which components have partially independent states, objectives, capabilities, or permissions, and where system-level behaviour depends on coordination among them. A *bounded multi-agent system* is then a multi-agent system whose components are deliberately bounded, and whose interfaces are designed to reap the benefits of modularity. Boundedness is a mechanism to prevent a modular system from becoming a de facto monolith, in which one part of the system has sufficient capability, information, and power to dictate the rest.

The key design principle in BMAS is *controlled modularity*. A bounded agent should have a legible scope in what it can observe, what it can infer, what it can do, and what it is trying to achieve. A BMAS can define this scope through typed interfaces, permission boundaries, provenance records, and audit trails. For example, a secure enterprise assistant should not give a single agent simultaneous access to private user data, untrusted web content, and unrestricted external communication, leading to a pattern known as the "lethal trifecta" [59], which creates severe security vulnerabilities.

2.3 Scaling Safely Through Bounded Multi-Agent Systems

To study and build BMAS, we first need a formal vocabulary for describing and quantifying agentic capabilities along different dimensions. Existing frameworks such as bounded optimality [60] and resource rationality [56, 57] explain why a constrained agent may rationally trade performance for resource use, but they do not tell us which concrete resource limitations should be placed on an agent to achieve a desired capability profile. Meanwhile, deep learning techniques such as distillation [61], quantisation [62–64], hierarchical embeddings [65], and pruning [66–70] typically aim to *maintain*

[3] We use boundedness here primarily in the context of *capabilities*, rather than learning ability [17, 44].

capabilities while reducing resource consumption, instead of *intentionally restricting* capabilities. These observations highlight the need for a formal taxonomy of resource constraints for AI agents that captures the relationship between different computational constraints and their effects on an agent's capabilities.

We have argued thus far that bounded agents are an important area of research that deserves further investigation. However, many tasks require a broad range of distinct capabilities which cannot be provided by a single bounded agent. In what follows, we argue that bounded agents can be used as building blocks in scaling through multi-agent systems to address such tasks. Such systems foreground the role of interfaces, incentives, and institutions in achieving cooperation at scale. Understanding what each component can do, how they respond to incentives, and how they interact with other systems allows us to compose agents into collectives that can reliably accomplish many tasks that we care about.

2.4 Components in the Design of BMAS

Interfaces and Contracts BMAS can specify what kinds of inputs an agent accepts, what it returns, which tools it may call, which data it may access, which reporting obligations it has, and how its outputs can be verified. This is analogous to microservice architecture in software engineering [71], but with a key difference: agentic services can plan, adapt, and respond strategically to incentives. This enables novel runtime infrastructure designs for attribution, interactions, and responses [72], as well as more sophisticated primitives for developing and governing such agent ecosystems [73, 74].

Orchestration and Task Matching Orchestration in BMAS should make use of the range of solutions available, including single-agent execution, independent parallel agents, centralised and decentralised coordination, debate, market mechanisms, and hybrid cascades, depending on task structure. Recent empirical work suggests that multi-agent coordination is more attractive for parallelisable, exploratory, adversarial, or role-structured tasks, while single-agent execution may be better when success depends on preserving a unified context or performing sequential reasoning [75, 76]. A practical BMAS therefore requires routing policies over architectures and not just models. Orchestrators must also be bounded: if coordination requires a fully general, highly capable agent with unrestricted access to all information and tools, the BMAS has reintroduced the monolith at a higher level. This could be mitigated through the use of bounded orchestrators with narrow routing mandates, multiple independent orchestrators with cross-checks, transparent task-allocation rules, and escalation to humans or certified systems when the orchestration decision is safety-critical.

Institutions, Markets, and Collective Choice BMAS components can be owned by different actors, trained on different data, and optimised for different objectives, hence sharing many structural properties with human society. This allows us to bring our expertise in designing institutions, organisations, markets, and mechanisms to bear on system-level alignment. Potential mechanisms include voting rules, auctions, market mechanisms, reputation systems, verification, and delegation protocols [77–84]. These mechanisms can align local incentives with system-level goals, but also introduce familiar governance problems: collusion, Sybil attacks, regulatory capture, reputation manipulation, and unequal representation. Deploying BMAS safely is thus a problem of designing institutions whose participants are bounded, heterogeneous, strategic, and partially observable. We expand on frameworks, solutions, and open problems for BMAS in Appendices B and C.

3 The Case For BMAS

We make the case for studying, testing, and deploying BMAS along several dimensions. We do not argue that BMAS are preferable to monolithic systems under all circumstances, but rather that BMAS expand the AI design space, allowing us to trade off capabilities, cost, privacy, verifiability, and governance in different ways depending on the situation. Different architectures are suited for different tasks, resource budgets, levels of risk tolerance, and required levels of assurance. In what follows, we argue that **BMAS can enable unique tradeoffs in capability, safety, practicality, and efficiency not accessible under the paradigm of scaling monolithic systems**.

3.1 The Capability Argument

BMAS can increase system-level capabilities without giving every component broad capabilities. The evidence for this is all around us in nature and society. Different system components (e.g., cells, workers) with different domains of speciality contributing to a group can give rise to collective capabilities that are not present in any individual member.

Collective Capabilities Multi-agent systems can vary widely in their constitution and structure depending on the kinds of problems the system needs to solve. A common architecture is that of a hierarchy of influence or control, which typically involves an *orchestrator* (which can itself be implemented in a monolithic or modular fashion), whose function is to organise the activities of other agents. Orchestrators can be involved in decomposing tasks, routing information and delegating subtasks to other agents, estimating and allocating resources, verifying and composing outputs from other components, and potentially communicating with other parts of a larger system [31, 82, 85, 86, 86–89]. Approaches to orchestration can be classified along several axes, including the degrees of centralised control, parallel execution, and inter-agent interaction. Naturally, different approaches are more appropriate for different tasks, and this also depends crucially on properties of the available agents [75, 87, 90, 91]. In addition to these approaches, more decentralised methods using techniques such as ensembling or voting even among identical agents may leverage "wisdom of the crowd" [92] effects to scale system-level capabilities [93, 94]. The inherent compositionality of BMAS can also enable compositional generalisation and data-efficient learning [95, 96].

Fluid and Crystallised Intelligence BMAS enable a finer distinction between *fluid* and *crystallised* intelligence [97]. Fluid intelligence is associated with the ability to solve novel problems, whereas crystallised intelligence is related to the accumulation and application of useful learned knowledge and patterns of behaviour. Monolithic generative models often solve tasks by synthesising behaviour *de novo*, which can be costly, unstable, and hard to measure across repeated uses. In BMAS, a first solution can be discovered through fluid problem-solving and then stabilised as a reusable *skill* [98–101]. This skill can be implemented by bounded agents with narrow capabilities and interfaces, making reuse more measurable. Its success rate, cost, latency, and failure modes can then be estimated both locally for each bounded agent and globally for the composed skill. This process can be supported by distilling narrow expertise into smaller models [61, 102] and orchestrating them [31, 85–87, 89].

Continual Learning and Catastrophic Forgetting Training a single general-purpose model to continually acquire new skills from a non-stationary stream of data is known to be difficult in part due to the problem of *catastrophic forgetting*, due to interference when using a shared parameter set for multiple capabilities [103–105]. In contrast, models that specialise in contexts unique to a particular role, person, or organisation can be continually trained within those particular domains on private, curated datasets to mitigate such interference. Thus, expanding system-level capabilities can be obtained by scaling through modularity and compositionality.

3.2 The Safety Argument

The safety argument against large, highly capable, monolithic models is clear: as their size and capabilities expand, their internal mechanisms become more complex and uninterpretable, their failure modes multiply, their motivations become harder to align, their capacity to scheme and deceive others increases, and the risk surface due to emergent misalignment grows [106–112]. In contrast, BMAS may offer many safety advantages, some of which we highlight below.

Interpretability The behaviour of a single, narrow agent can be easier to understand and predict than that of a large-scale, black-box model performing many functions. Boundedness allows established compositionality and verification methods to be used to reason about and guarantee desirable system-level properties, with given levels of confidence. While interpretability remains a concern in multi-agent systems – for example, with respect to credit assignment and blame attribution [113, 114] – we argue that the structure imposed by enforcing specialisation in a structured manner can result in greater interpretability relative to dense monolithic agents [115]. Resource-conscious agents are also less likely to pursue overly complex policies which are difficult to interpret, as reflected by the use of information bottlenecks [116, 117] and pruning [118, 119] for interpretability.

Monitorability Monitoring requires the verification of inputs, processes, and outputs at levels of intensity calibrated to the sensitivity of information and the need for transparency. The monitoring of inputs is concerned with whether systems can effectively acquire the resources, permissions, and capabilities to successfully accomplish their tasks, and whether inputs are malicious or outside of the scope of the AI system [120, 121]. Monitoring of processes can ensure that systems comply with designated protocols and help to detect malicious behaviour [122, 123]. Monitoring of outputs enables assessment of the quality of outputs against rubrics [124], constitutions [125, 126], and verifiers [127], as well as ensuring compliance with specified guardrails. Such guardrails could be implemented by systems specifically designed to produce honest, calibrated predictions [128]. Approaches to monitoring can be distinguished by different degrees of intensity by their targets, observability, transparency, privacy, and topology [82]. On top of monitoring at the level of individual agents, BMAS can enable effective system-level monitoring through their modular nature. If interactions are legible by design, the communications between different agents can be monitored for signs of collusion and collective misalignment [18, 129].

Steerability Bounded, narrow agents have more limited and verifiable goals compared to monolithic models (which allows for community oversight), and can be less likely to scheme and harbour hidden goals as a consequence of their limited capabilities. Furthermore, scalable oversight methods [130–132] are a promising avenue for steering and monitoring complex AI systems, relying on humans and AI models supervising other models [133]. Finally, multi-agent systems allow for the application of mechanism design and social choice theory to control system-level outcomes [33, 134–139].

Self-correction Analogous to collective intelligences found in nature, modular systems allow agents to critique, red-team, and "self-police" one another, creating an ecosystem that is resilient to single-agent failures or misalignment [140], such as in biological immune systems. Moreover, the presence of multiple agents possessing the capability to complete a task renders the ecosystem more robust to the failure of any individual subsystem [141]. In addition, by distributing tasks among several agents, game-theoretic incentives can be designed to incentivise system-level self-correction [136, 142, 143].

Reliability Large language models can struggle with tasks involving the repeated application of simple logical rules over long horizons [144], and decomposing such problems into simpler subtasks allows them to be more reliably solved, verified, and error-corrected by smaller models [145]. In addition, boundedly rational systems behave 'as if' their policies are regularised. This regularisation is equivalent to a hedging strategy against adversarial perturbations to their objective function [146, 147], which provides a natural method to mitigate the effects of over-optimisation under task uncertainty.

3.3 The Practicality Argument

Diverse Representation The problem of fairly representing and aggregating the preferences of many heterogeneous members of a population is central to social choice theory [148]. Appropriate representation is also important in the complex problems for which AI systems may be employed, since these will generally involve multiple stakeholders with differing priorities and potentially conflicting values. A potential benefit of BMAS is the ability to faithfully represent diverse participants and stakeholders through specialised agents with the narrow mandate of subgroup representation [149]. In particular, shared and distributed ownership of different systems means that economic benefits can be distributed more equitably, and choices about which interests different components of the system serve are made by individuals and groups who own those components, rather than by a fully centralised authority. Consequently, BMAS offer a potential route to pluralistic alignment [148], where diverse viewpoints and interests are represented by different agents [150, 151].

Privacy and Data Ownership Compared to monolithic systems, BMAS possess the advantage of the structural privacy constraints provided by bounded agents. Since each agent is trained on and executes a narrow task, it only requires access to data relevant to that specific task [152]. In contrast, a monolithic agent necessarily has access to the full data pipeline, raising privacy concerns. BMAS enables data minimisation by design [153]: the scope of an agent's training and inference-time data access can be constrained to only what is necessary, reducing the risk of inadvertent data leakage. The modular architecture of BMAS also makes distributed data ownership tractable, as different agents can be maintained by different stakeholders that retain ownership, privacy, and governance rights.

Governance AI governance is increasingly central to policy, but standardising regulation for general-purpose models remains difficult, leaving substantial reliance on self-regulation as governance lags behind deployment [154], with the example of the voluntary commitments signed by major labs in 2023 [155]. BMAS make governance more tractable by giving bounded agents clearer capability scopes: components can be assessed for safety and compliance individually, enabling standards analogous to those in safety-critical industries such as aerospace or medicine. This modularity supports standardisation, accountability, and liability attribution. It may also reduce compliance burdens under regimes such as the EU AI Act [156], whose heaviest obligations fall on broad General Purpose AI models, thereby incentivising the deployment of safer specialised architectures.

Assurance and Verification Assurance and verification become more tractable in BMAS because they move from a global property of an opaque model to a local property of bounded components, interfaces, and protocols [157, 158]. A bounded agent can be verified against a narrower contract through what inputs it may receive, what outputs it may produce, which tools it may call, which data it may access, what resources it may consume, and which traces it must expose for audit [158, 159]. These boundaries create natural intervention points, where interfaces, tool calls, data accesses, and handoffs can be observed, constrained, logged, interrupted, or verified. As a result, failures can be attributed to specific components or channels, redundancy can be introduced where needed, and assurance becomes an architectural property rather than a post-hoc analysis of a monolithic model.

Concentration of Power and Risks Monolithic systems tend to couple capability, authority, data, tool use, and economic mediation in a single point of control [160, 161]. This concentration increases the stakes of capture, unilateral policy changes, misuse, and single-point failure [162]. BMAS can partially decouple these functions across bounded agents with limited mandates, explicit permissions, and potentially distinct owners or stakeholder constituencies [163, 164]. This does not remove political or institutional risk, but it can make power more distributed, contestable, and governable.

Inevitability and Urgency Multi-agent systems have been a central focus in frontier AI development in recent years, and it is likely that this trend is only going to accelerate. Indeed, the recent development of protocols such as MCP, A2A, UCP, and AP2 have facilitated the interconnection and interaction of agentic AI systems through the internet. This has led to several discussions of concepts like the agentic web [80] and agent economies [78, 79, 81] as an emerging layer of society.

In parallel to the centralised deployment of multi-agent systems, we argue that due to the distributed nature of knowledge and data in society [165], it is likely that distributed multi-agent systems will play an increasingly central role in the emerging agentic web [77, 166]. An early example of this is the development of personal agentic systems such as OpenClaw, which catalysed the creation of Moltbook — a social network platform for AI agents to communicate via free text with little human oversight [15, 167]. The potential for unstructured interactions to have undesirable impacts on humans and society is growing at an increasing rate, necessitating urgent attention to designing the sociotechnical infrastructure for these systems in a principled and safe manner.

3.4 The Efficiency Argument

Efficiency from Modularity and the Division of Labour The economic and environmental costs of frontier-scale AI make efficient allocation of compute an increasingly critical concern. It is wasteful to route simple tasks through the full capacity of a highly capable generalist model when a smaller specialist or tool would suffice. The field already recognises this at the parameter level. For example, MoE architectures decouple the total parameter count from the number of active parameters and use routing to activate a small subset of "experts" per token [29, 168, 169]. At the system level, efficiency gains can arise from the orchestration of tools and agents [75, 85, 87, 88, 91, 101] in addition to the capability gains discussed above. However, specialisation is not automatically efficient. Coordination, communication, redundancy, and orchestration all impose additional overheads that may dominate any savings from the use of more efficient specialists. It therefore crucially depends on the task as to whether the savings outweigh the additional costs in using BMAS over monolithic systems.

Specialisation and the division of labour have long been recognised as vital drivers of efficiency in human society [20]. This principle extends to self-organising systems in nature, where the development of multicellular organisms led to significant metabolic efficiency gains [170–172]. More generally, a similar phenomenon is likely to manifest whenever scarce resources must be

allocated under selective pressure. We argue that efficiency considerations will drive society toward similar ecosystem architectures as demand for data and computational resources grows faster than the capacity to meet it expands. More inference will gradually be pushed to edge devices, whose interconnection will enable the deployment of smaller, more efficient AI systems for a wide variety of tasks where broad capabilities or intelligence are not necessary [173, 174].

4 Counter-Arguments and Risk Assessment

Here, we discuss some of the most salient counter-arguments to the arguments laid out in Section 3. In particular, *we do not argue that BMAS eliminate risks from agentic AI systems*. Rather, our position is that they shift the risk profile into different domains where we have much more existing experience and a wider range of expertise to draw from in addressing these problems.[4]

System-level behaviour may be less interpretable A system of bounded agents can generate emergent dynamics that may be harder to predict than the behaviour of one model. This relates to the broader study of complex systems, where it is well-known that macro-level behaviour can qualitatively differ from the properties of individual components [175–177]. However, the relevant level of abstraction changes with BMAS. Internal activations of large, monolithic systems are difficult to cleanly map onto human-legible concepts [178–180]. On the other hand, inter-agent communications and audit logs can be made directly observable to secure parties tasked with mitigating and detecting harmful outcomes. BMAS may be less mechanistically transparent but more governable, provided that communication is constrained to legible channels and monitored for manipulation [181], collusion [182], goal drift [183], and other failure modes. A key intuition driving this counterargument is that of *emergence* [184, 185]. Due to the interactive nature of multi-agent interactions, each agent's environment is inherently non-stationary, and can give rise to chaotic [186, 187] or emergent behaviour [188]. It is thus critical to intentionally design BMAS systems to promote boundedness and safety at the system-level, as well as the component level.

Some tasks require generality Open-ended scientific research, complex strategic planning, and cross-domain synthesis may be harder to cleanly decompose. Over-specialised agents can be brittle under distributional shift, just as specialised species can be fragile outside their ecological niches. BMAS should therefore not exclude more general agents or humans when needed, but rather route tasks to appropriate systems, including bounded agents that are trained or equipped with the capability to recognise the edge of their competence. In other words, generality is useful, but that generality should be allocated deliberately and cautiously rather than granted by default to every component.

Useful and dangerous capabilities may not be cleanly separable Some useful tasks require planning, world modelling, persuasion, technical knowledge, or strategic reasoning that can also support misuse. Bounding is not a complete solution to dual use. Its value lies in separating the components of harm: knowledge from authority, planning from execution, access from communication, and proposal from approval. Checks and balances can be put in place in between these components to guard against system-level harmful behaviour. In high-risk domains, BMAS must combine capability isolation with monitoring, external verification, and human or institutional oversight [19].

The orchestrator could become the new monolith A centralised orchestrator with unrestricted context and authority can recreate the risks BMAS was meant to avoid. Concentrated-authority architectures can lead to devastating failure modes, even without malicious intent [189–191]. A safer design could use bounded orchestrators, limited routing mandates, transparent allocation rules, redundant checks, and clear escalation protocols. In some domains, the orchestration function may itself need to be distributed across multiple agents or institutions, leading to a nested, recursive structure of implementation rather than a shallow hierarchy that concentrates power and risks.

5 Discussion

Our central claim is that AI systems should not be scaled only by making individual agents more capable and more general, but that we should direct more attention and resources to scaling systems of

[4] We discuss further counterarguments in Appendix D.

agents while deliberately making them bounded, modular, legible, and composable. The nature of the bounds, mechanisms, and institutions that are deployed should evolve carefully as our understanding and ability to wisely engage with the technology matures. Nonetheless, we argue that BMAS offers a promising avenue to transform resource constraints into safety constraints, and modularity into a basis for capability, verification, and governance.

Multi-agent AI systems are already being deployed at increasing scale in society, with relatively scarce infrastructure to promote aligned, safe, and resilient behaviours at the individual and collective levels. Future frontier AI systems will likely combine large monolithic models, sparse experts, skill libraries, model and tool routers, orchestrated ensembles, and digital institutions. We maintain that a central challenge is to understand how different components should be bounded, how they should be composed, and how properties at multiple levels of abstraction can be verified and assured.

We believe that the potential benefits of developing BMAS are significant, and that humanity has a strong starting point to build upon as we address the accompanying risks. This agenda could ultimately lead to the development of infrastructure that supports a **democratic and distributed AI ecosystem as a global public good** [192, 193]. Through this, individuals can benefit from AI systems that represent their interests, defend their privacy, and support them in making a living while preserving and enhancing their agency [191, 194–196]. However, this path is not the default. Multi-agent AI ecosystems are emerging organically around systems that were never designed to be bounded, composable, or governable. Thus, significant effort will be required across disciplines and sectors to study, build, and iterate on BMAS designs, and the time for us to do so is *now*.

Acknowledgments and Disclosure of Funding

The authors would like to thank Adam Safron, D. Scott Phoenix, James Fox, Jonas Hallgren, Lewis Hammond, and Paul Colognese for invaluable feedback and discussions that helped to shape this work.

DH, NB, AC, and MJW were supported by a UKRI AI World Leading Researcher Fellowship (grant number EP/W002949/1). AC was also supported by a University of Oxford UKRI Impact Acceleration Account Seed Fund award. TG was supported by a Czech Science Foundation grant (grant number 26-23955S).

References

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

[2] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. arXiv:2404.09932.

[3] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021.

[4] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah

Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Olubunmi Ajala, Fahad Albalawi, Marwan Alserkal, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Chris Johnson, Gill Jolly, Ziv Katzir, Saif M. Khan, Hiroaki Kitano, Antonio Krüger, Kyoung Mu Lee, Dominic Vincent Ligot, José Ramón López Portillo, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, Raquel Pezoa Rivera, Balaraman Ravindran, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International ai safety report. Technical Report DSIT 2025/001, 2025.

[6] Yoshua Bengio, Stephen Clare, Carina Prunkl, Malcolm Murray, Maksym Andriushchenko, Ben Bucknall, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, David Duvenaud, Philip Fox, Usman Gohar, Rose Hadshar, Anson Ho, Tiancheng Hu, Cameron Jones, Sayash Kapoor, Atoosa Kasirzadeh, Sam Manning, Nestor Maslej, Vasilios Mavroudis, Conor McGlynn, Richard Moulange, Jessica Newman, Kwan Yee Ng, Patricia Paskov, Shalaleh Rismani, Girish Sastry, Elizabeth Seger, Scott Singer, Charlotte Stix, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Vincent Conitzer, Thomas G. Dietterich, Edward W. Felten, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Susan Leavy, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Sarvapali D. Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Leandro Angelo Aguirre, Olubunmi Ajala, Fahad Albalawi, Noora AlMalek, Christian Busch, Jonathan Collas, André Carlos Ponce de Leon Ferreira de Carvalho, Amandeep Gill, Ahmet Halit Hatip, Juha Heikkilä, Chris Johnson, Gill Jolly, Ziv Katzir, Mary N. Kerema, Hiroaki Kitano, Antonio Krüger, Kyoung Mu Lee, José Ramón López Portillo, Aoife McLysaght, Olexii Molchanovskyi, Andrea Monti, Mona Nemer, Nuria Oliver, Raquel Pezoa, Audrey Plonk, Balaraman Ravindran, Hammam Riza, Crystal Rugege, Haroon Sheikh, Denise Wong, Yi Zeng, Liming Zhu, Daniel Privitera, and Sören Mindermann. International AI safety report 2026. Technical Report DSIT 2026/001, Department for Science, Innovation and Technology, 2026.

[7] 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.

[8] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

[9] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2025.

[10] Satvik Golechha and Adrià Garriga-Alonso. Among us: A sandbox for measuring and detecting agentic deception. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[11] Arun Jose, Niels Warncke, and Mia Taylor. Strategic obfuscation of deceptive reasoning in language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[12] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

[13] Rahul Marchand, Art O Cathain, Jerome Wynne, Philippos Maximos Giavridis, Sam Deverett, John Wilkinson, Jason Gwartz, and Harry Coppock. Quantifying frontier llm capabilities for container sandbox escape. *arXiv preprint arXiv:2603.02277*, 2026.

[14] Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

[15] Zhengyang Shan, Jiayun Xin, Yue Zhang, and Minghui Xu. Don't let the claw grip your hand: A security analysis and defense framework for openclaw. *arXiv preprint arXiv:2603.10387*, 2026.

[16] Jonathan David Thomas, Raul Santos-Rodriguez, and Robert Piechocki. Understanding redundancy in discrete multi-agent communication. In *Second Workshop on Language and Reinforcement Learning*, 2022.

[17] Judah Goldfeder, Philippe Wyder, Yann LeCun, and Ravid Shwartz Ziv. Ai must embrace specialization via superhuman adaptable intelligence. *arXiv preprint arXiv:2602.23643*, 2026.

[18] Aaron Sandoval and Cody Rushing. Factor(T,U): Factored cognition strengthens monitoring of untrusted AI. *arXiv preprint arXiv:2512.02157*, 2025.

[19] Ryan Lowe, Joe Edelman, Tan Zhi-Xuan, Oliver Klingefjord, Ellie Hain, Vincent Wang, Atrisha Sarkar, Michiel A. Bakker, Fazl Barez, Matija Franklin, Andreas Haupt, Jobst Heitzig, Wesley H. Holliday, Julian Jara-Ettinger, Atoosa Kasirzadeh, Ryan Othniel Kearns, James Ravi Kirkpatrick, Andrew Koh, Joel Lehman, Sydney Levine, Manon Revel, and Ivan Vendrov. Full-stack alignment: Co-aligning AI and institutions with thicker models of value. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

[20] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, 1776.

[21] J Benjamin Falandays, Roope O Kaaronen, Cody Moser, Wiktor Rorot, Joshua Tan, Vishwanath Varma, Tevin Williams, and Mason Youngblood. All intelligence is collective intelligence. *Journal of Multiscale Neuroscience*, 2(1):169–191, 2023.

[22] Michael Levin. Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, Volume 16 - 2022, 2022. ISSN 1662-5137. doi: 10.3389/fnsys.2022.768201.

[23] Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 7(1):378, 2024.

[24] Michael Levin and Benjamin Lyons. Cognitive glues are shared models of relative scarcities: the economics of collective intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2320):20240528, 05 2026. ISSN 1364-503X. doi: 10.1098/rsta.2024.0528.

[25] Grégoire Delétang, Jordi Grau-Moya, Miljan Martic, Tim Genewein, Tom McGrath, Vladimir Mikulik, Markus Kunesch, Shane Legg, and Pedro A. Ortega. Causal analysis of agent behavior for AI safety. *CoRR*, abs/2103.03938, 2021.

[26] Kai Sun, Wenqiang Li, Bo Dong, Yuxin Lin, Jingyao Zhang, and Bin Shi. Scope delineation before localization: A two-stage framework for enhancing failure attribution in multi-agent systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33108–33116, Mar. 2026.

[27] Yingxuan Yang, Huacan Chai, Shuai Shao, Yuanyi Song, Siyuan Qi, Renting Rui, and Weinan Zhang. Agentnet: Decentralized evolutionary coordination for LLM-based multi-agent systems. In *Advances in Neural Information Processing Systems 39*, 2025.

[28] Jen tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael Lyu, and Maarten Sap. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[30] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[31] Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Multi-agent collaboration via evolving orchestration, 2025.

[32] Jan Kulveit. Hierarchical agency: A missing piece in ai alignment. https://www.lesswrong.com/posts/xud7Mti9jS4tbWqQE/hierarchical-agency-a-missing-piece-in-ai-alignment, 2024. Accessed: 2026-05-02.

[33] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: social choice should guide ai alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. PMLR, 2024.

[34] Mingzhi Wang, Chengdong Ma, Qizhi Chen, Linjian Meng, Yang Han, Jiancong Xiao, Zhaowei Zhang, Jing Huo, Weijie J Su, and Yaodong Yang. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

[35] Yang Cai, Constantinos Daskalakis, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Proximal regret and proximal correlated equilibria: A new tractable solution concept for online learning and games. *arXiv preprint arXiv:2511.01852*, 2025.

[36] Runyu Lu, Yuanheng Zhu, and Dongbin Zhao. Divergence-regularized discounted aggregation: Equilibrium finding in multiplayer partially observable stochastic games. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Alexander Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Steering no-regret learners to a desired equilibrium. *arXiv preprint arXiv:2306.05221*, 2023.

[38] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607, 2022. doi: 10.1126/sciadv.abk2607.

[39] Antoine Scheid, Daniil Tiapkin, Etienne Boursier, Aymeric Capitaine, Eric Moulines, Michael I. Jordan, El-Mahdi El-Mhamdi, and Alain Durmus. Incentivized learning in principal-agent bandit games. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[40] Idan Habler, Ken Huang, Vineeth Sai Narajala, and Prashant Kulkarni. Building a secure agentic ai application leveraging a2a protocol. *arXiv preprint arXiv:2504.16902*, 2025.

[41] Abram Demski and Scott Garrabrant. Embedded agency. *arXiv preprint arXiv:1902.09469*, 2020.

[42] Khurram Javed and Richard S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

[43] Alex Lewandowski, Aditya A. Ramesh, Edan Meyer, Dale Schuurmans, and Marlos C. Machado. The world is bigger! a computationally-embedded perspective on the big world hypothesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[44] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[45] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.

[46] Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1–48, 2014.

[47] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.

[48] Herbert A. Simon. *From substantive to procedural rationality*, page 129–148. Cambridge University Press, 1976.

[49] Olivia Macmillan-Scott and Mirco Musolesi. (ir)rationality in ai: State of the art, research challenges and open questions. *arXiv preprint arXiv:2311.17165*, 2023.

[50] Stuart Russell, Eric H. Wefald, Daniel G. Bobrow, Michael Brady, and Randall Davis. *Do the Right Thing: Studies in Limited Rationality*. The MIT Press, 07 1991. ISBN 9780262282772.

[51] Richard L Lewis, Andrew Howes, and Satinder Singh. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6 (2):279–311, 2014.

[52] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245): 273–278, 2015.

[53] Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013.

[54] Daniel A. Braun and Pedro A. Ortega. Information-theoretic bounded rationality and epsilon-optimality. *Entropy*, 16(8):4662–4676, 2014. ISSN 1099-4300.

[55] Pedro A Ortega, Daniel A Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.

[56] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.

[57] Thomas F Icard. *Resource rationality*. MIT Press, 2023.

[58] Marcel Binz and Eric Schulz. Modeling human exploration through resource-rational reinforcement learning. *Advances in neural information processing systems*, 35:31755–31768, 2022.

[59] Simon Willison. The lethal trifecta for ai agents: private data, untrusted content, and external communication. https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/, 2025. Accessed: 2026-05-02.

[60] Stuart J Russell and Devika Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1994.

[61] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[62] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

[63] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR, 2023.

[64] Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 467–484, 2024.

[65] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249, 2022.

[66] Elias Frantar and Dan Alistarh. Sparsegpt: massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[67] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[68] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: on the structural pruning of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[69] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[70] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020.

[71] Nuha Alshuqayran, Nour Ali, and Roger Evans. A systematic mapping study in microservice architecture. In *2016 IEEE 9th international conference on service-oriented computing and applications (SOCA)*, pages 44–51. IEEE, 2016.

[72] Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. Infrastructure for AI agents. *Transactions on Machine Learning Research*, 2025. arXiv:2501.10114.

[73] Karl J Friston, Maxwell JD Ramstead, Alex B Kiefer, Alexander Tschantz, Christopher L Buckley, Mahault Albarracin, Riddhi J Pitliya, Conor Heins, Brennan Klein, Beren Millidge, Dalton AR Sakthivadivel, Toby St Clere Smithe, Magnus Koudahl, Safae Essafi Tremblay, Capm Petersen, Kaiser Fung, Jason G Fox, Steven Swanson, Dan Mapes, and Gabriel René. Designing ecosystems of intelligence from first principles. *Collective Intelligence*, 3(1): 26339137231222481, 2024. doi: 10.1177/26339137231222481.

[74] Charles L. Wang, Trisha Singhal, Ameya Kelkar, and Jason Tuo. MI9: An integrated runtime governance framework for agentic AI. *arXiv preprint arXiv:2508.03858*, 2025.

[75] Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2026.

[76] Dat Tran and Douwe Kiela. Single-agent llms outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets. *arXiv preprint arXiv:2604.02460*, 2026.

[77] Mark S Miller and K Eric Drexler. Markets and computation: Agoric open systems. *The ecology of computation*, 1:133–176, 1988.

[78] Gillian K Hadfield and Andrew Koh. An economy of ai agents. *arXiv preprint arXiv:2509.01063*, 2025.

[79] Yingxuan Yang, Ying Wen, Jun Wang, and Weinan Zhang. Agent exchange: Shaping the future of ai agent economics. *arXiv preprint arXiv:2507.03904*, 2025.

[80] Yingxuan Yang, Mulei Ma, Yuxuan Huang, Huacan Chai, Chenyu Gong, Haoran Geng, Yuanjian Zhou, Ying Wen, Meng Fang, Muhao Chen, Shangding Gu, Ming Jin, Costas Spanos, Yang Yang, Pieter Abbeel, Dawn Song, Weinan Zhang, and Jun Wang. Agentic web: Weaving the next web with ai agents. *arXiv preprint arXiv:2507.21206*, 2025.

[81] Nenad Tomasev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero. Virtual agent economies. *arXiv preprint arXiv:2509.10147*, 2025.

[82] Nenad Tomašev, Matija Franklin, and Simon Osindero. Intelligent ai delegation. *arXiv preprint arXiv:2602.11865*, 2026.

[83] Philip Moreira Tomei, Rupal Jain, and Matija Franklin. AI governance through markets. *arXiv preprint arXiv:2501.17755*, 2025.

[84] F. Pierucci, M. Galisai, M. Bracale Syrnikov, M. Prandi, P. Bisconti, F. Giarrusso, O. Sorokoletova, V. Suriani, and D. Nardi. Institutional AI: A governance framework for distributional AGI safety. *arXiv preprint arXiv:2601.10599*, 2026.

[85] Hongjin Su, Shizhe Diao, Ximing Lu, Mingjie Liu, Jiacheng Xu, Xin Dong, Yonggan Fu, Peter Belcak, Hanrong Ye, Hongxu Yin, Yi Dong, Evelina Bakhturina, Tao Yu, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Toolorchestra: Elevating intelligence via efficient model and tool orchestration. *arXiv preprint arXiv:2511.21689*, 2025.

[86] Jinglue Xu, Qi Sun, Peter Schwendeman, Stefan Nielsen, Edoardo Cetin, and Yujin Tang. Trinity: An evolved LLM coordinator. In *The Fourteenth International Conference on Learning Representations*, 2026.

[87] Umang Bhatt, Sanyam Kapoor, Mihir Upadhyay, Ilia Sucholutsky, Francesco Quinzan, Katherine M. Collins, Adrian Weller, Andrew Gordon Wilson, and Muhammad Bilal Zafar. When should we orchestrate multiple agents? *arXiv preprint arXiv:2503.13577*, 2025.

[88] Stefan Nielsen, Edoardo Cetin, Peter Schwendeman, Qi Sun, Jinglue Xu, and Yujin Tang. Learning to orchestrate agents in natural language with the conductor. In *The Fourteenth International Conference on Learning Representations*, 2026.

[89] Kimi Team. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[90] Melissa Z Pan, Mert Cemri, Lakshya A Agrawal, Shuyi Yang, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Kannan Ramchandran, Dan Klein, Joseph E. Gonzalez, Matei Zaharia, and Ion Stoica. Why do multiagent systems fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

[91] Mingyan Gao, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai. Single-agent or multi-agent systems? why not both? *arXiv preprint arXiv:2505.18286*, 2025.

[92] James Surowiecki. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday & Co., 2004.

[93] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[94] Ariel Kamen and Yakov Kamen. Majority rules: Llm ensemble is a winning approach for content categorization. *arXiv preprint arXiv:2511.15714*, 2025.

[95] Yilun Du and Leslie Pack Kaelbling. Position: Compositional generative modeling: A single model is not all you need. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11721–11732. PMLR, 21–27 Jul 2024.

[96] David A. Danhofer, Davide D'Ascenzo, Rafael Dubach, and Tomaso A. Poggio. Position: A theory of deep learning must include compositional sparsity. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81199–81210. PMLR, 13–19 Jul 2025.

[97] Raymond B Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1, 1963.

[98] Renjun Xu and Yang Yan. Agent skills for large language models: Architecture, acquisition, security, and the path forward. *arXiv preprint arXiv:2602.12430*, 2026.

[99] Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

[100] Qi Sun, Stefan Nielsen, Rio Yokota, and Yujin Tang. Evolutionary context search for automated skill acquisition. *arXiv preprint arXiv:2602.16113*, 2026.

[101] Xiaoxiao Li. When single-agent with skills replace multi-agent systems and when they fail. *arXiv preprint arXiv:2601.04748*, 2026.

[102] Eric J Michaud, Asher Parker-Sartori, and Max Tegmark. On the creation of narrow AI: hierarchy and nonlocality of neural network skills. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[103] Gido M Van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*, 2024.

[104] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

[105] Saurabh Kumar, Henrik Marklund, Ashish Rao, Yifan Zhu, Hong Jun Jeon, Yueyang Liu, and Benjamin Van Roy. Continual learning as computationally constrained reinforcement learning. *Foundations and Trends® in Machine Learning*, 18(5):913–1053, 2025. ISSN 1935-8237.

[106] Erik Jones, Meg Tong, Jesse Mu, Mohammed Mahfoud, Jan Leike, Roger Grosse, Jared Kaplan, William Fithian, Ethan Perez, and Mrinank Sharma. Forecasting rare language model behaviors. *arXiv preprint arXiv:2502.16797*, 2025.

[107] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[108] Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When

bigger isn't better. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Featured Certification.

[109] Thilo Hagendorff. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121, 2024.

[110] Yao Huang, Yitong Sun, Yichi Zhang, Ruochen Zhang, Yinpeng Dong, and Xingxing Wei. Deceptionbench: A comprehensive benchmark for AI deception behaviors in real-world scenarios. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[111] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*, 2025.

[112] Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025.

[113] Stelios Triantafyllou, Aleksa Sukovic, Yasaman Zolfimoselo, and Goran Radanovic. Counterfactual effect decomposition in multi-agent sequential decision making. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 60072–60098. PMLR, 13–19 Jul 2025.

[114] Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 76583–76599. PMLR, 13–19 Jul 2025.

[115] Xingyi Yang, Constantin Venhoff, Ashkan Khakzar, Christian Schroeder de Witt, Puneet K. Dokania, Adel Bibi, and Philip Torr. Mixture of experts made intrinsically interpretable. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[116] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

[117] Ying Wang, Tim GJ Rudner, and Andrew G Wilson. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems*, 36:16009–16027, 2023.

[118] Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[119] Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Dan Mossing. Weight-sparse transformers have interpretable circuits. *arXiv preprint arXiv:2511.13653*, 2025.

[120] Shayan Ali Hassan, Tao Ni, Zafar Ayyub Qazi, and Marco Canini. Efficient and adaptable detection of malicious llm prompts via bootstrap aggregation. *arXiv preprint arXiv:2602.08062*, 2026.

[121] Yanting Wang, Wei Zou, Runpeng Geng, and Jinyuan Jia. Agentwatcher: A rule-based prompt injection monitor. *arXiv preprint arXiv:2604.01194*, 2026.

[122] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving,

Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025.

[123] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[124] Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*, 2026.

[125] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[126] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1395–1417. ACM, 2024. doi: 10.1145/3630106.3658979.

[127] Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

[128] Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, et al. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*, 2025.

[129] Sumeet R Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H Torr, Lewis Hammond, and Christian S de Witt. Secret collusion among ai agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37:73439–73486, 2024.

[130] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

[131] Zachary Kenton, Noah Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah Goodman, et al. On scalable oversight with weak llms judging strong llms. *Advances in Neural Information Processing Systems*, 37:75229–75276, 2024.

[132] Abhimanyu Pallavi Sudhir, Jackson Kaunismaa, and Arjun Panickssery. A benchmark for scalable oversight mechanisms. *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*, 2025.

[133] Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

[134] Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi. Emergent alignment via competition. *arXiv preprint arXiv:2509.15090*, 2025.

[135] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, 2024.

[136] Baiting Chen, Tong Zhu, Jiale Han, Lexin Li, Gang Li, and Xiaowu Dai. Incentivizing truthful language models via peer elicitation games. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[137] Carter Blair and Kate Larson. Generating fair consensus statements with social choice on token-level mdps. *arXiv preprint arXiv:2510.14106*, 2025.

[138] Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 144–155, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719.

[139] Yun Hua, Haosheng Chen, Shiqin Wang, Wenhao Li, Xiangfeng Wang, and Jun Luo. Shapley-coop: Credit assignment for emergent cooperation in self-interested llm agents. *arXiv preprint arXiv:2506.07388*, 2025.

[140] Shiyang Lai, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song, and James Evans. Position: Evolving AI collectives enhance human diversity and enable self-regulation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25892–25912. PMLR, 21–27 Jul 2024.

[141] David Ha and Yujin Tang. Collective intelligence for deep learning: A survey of recent developments. *Collective Intelligence*, 1(1):26339137221114874, 2022. doi: 10.1177/26339137221114874.

[142] Yongqiang Chen, Gang Niu, James Cheng, Bo Han, and Masashi Sugiyama. Towards scalable oversight with collaborative multi-agent debate in error detection. *arXiv preprint arXiv:2510.20963*, 2025.

[143] Rohit Agarwal, Joshua Lin, Mark Braverman, and Elad Hazan. Ai alignment via incentives and correction. *arXiv preprint arXiv:2605.01643*, 2026.

[144] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

[145] Elliot Meyerson, Giuseppe Paolo, Roberto Dailey, Hormoz Shahrzad, Olivier Francon, Conor F. Hayes, Xin Qiu, Babak Hodjat, and Risto Miikkulainen. Solving a million-step llm task with zero errors. *arXiv preprint arXiv:2511.09030*, 2025.

[146] Pedro Ortega and Daniel Lee. An adversarial interpretation of information-theoretic bounded rationality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[147] Pedro A. Ortega. Bounded-rationality, hedging, and generalization. *arXiv preprint arxiv:2605.15340*, 2026.

[148] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR, 21–27 Jul 2024.

[149] Srijoni Majumdar, Edith Elkind, and Evangelos Pournaras. Generative ai voting: fair collective choice is resilient to llm biases and inconsistencies. *EPJ Data Science*, 2026.

[150] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[151] Sydney Levine, Matija Franklin, Tan Zhi-Xuan, Secil Yanik Guyot, Lionel Wong, Daniel Kilov, Yejin Choi, Joshua B. Tenenbaum, Noah Goodman, Seth Lazar, and Iason Gabriel. Resource rational contractualism should guide ai alignment. *arXiv preprint arXiv:2506.17434*, 2025.

[152] Awais Qasim, Arslan Ghouri, and Adeel Munawar. An effective approach for reducing data redundancy in multi-agent system communication. *Multiagent and Grid Systems*, 20(1):69–88, 2024. doi: 10.3233/MGS-230089.

[153] Robin Staab, Nikola Jovanović, Kimberly Mai, Prakhar Ganesh, Martin Vechev, Ferdinando Fioretto, and Matthew Jagielski. Sok: Data minimization in machine learning. *arXiv preprint arXiv:2508.10836*, 2026.

[154] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. Practices for governing agentic AI systems. *Research Paper, OpenAI*, 2023.

[155] Harvard Law Review. Voluntary commitments from leading artificial intelligence companies on July 21, 2023. *Harvard Law Review*, 137, 2023.

[156] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

[157] Sanjit A. Seshia, Dorsa Sadigh, and S. Shankar Sastry. Toward verified artificial intelligence. *Communications of the ACM*, 65(7):46–55, 2022. doi: 10.1145/3503914.

[158] Ciprian Paduraru, Petru-Liviu Bouruc, and Alin Stefanescu. A trace-based assurance framework for agentic ai orchestration: Contracts, testing, and governance. *arXiv preprint arXiv:2603.18096*, 2026. doi: 10.48550/arXiv.2603.18096.

[159] Jatinder Singh, Jennifer Cobbe, and Chris Norval. Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, 7:6562–6574, 2019. doi: 10.1109/ACCESS.2018.2887201.

[160] Fernando van der Vlist, Anne Helmond, and Fabian Ferrari. Big ai: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, 11(1): 20539517241232630, 2024. doi: 10.1177/20539517241232630.

[161] OECD. Artificial intelligence, data and competition. Technical Report 18, OECD Publishing, 2024.

[162] Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*, 2024. doi: 10.48550/arXiv.2402.08797.

[163] National Telecommunications and Information Administration. Dual-use foundation models with widely available model weights. Technical report, U.S. Department of Commerce, 2024.

[164] Nicholas Vincent, Matt Prewitt, and Hanlin Li. Collective bargaining in the information economy can address ai-driven power concentration. *arXiv preprint arXiv:2506.10272*, 2025.

[165] F. A. Hayek. The use of knowledge in society. *The American Economic Review*, 35(4): 519–530, 1945. ISSN 00028282.

[166] Nenad Tomašev, Matija Franklin, Julian Jacobs, Sébastien Krier, and Simon Osindero. Distributional agi safety. *arXiv preprint arXiv:2512.16856*, 2025.

[167] Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. "humans welcome to observe": A first look at the agent social network moltbook. *arXiv preprint arXiv:2602.10127*, 2026.

[168] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V Le, and James Laudon. Mixture-of-experts with expert choice routing. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[169] OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025.

[170] Richard E Michod. Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences*, 104(suppl_1):8613–8618, 2007.

[171] Robert MacArthur and Richard Levins. The limiting similarity, convergence, and divergence of coexisting species. *The american naturalist*, 101(921):377–385, 1967.

[172] Douglas J Futuyma and Gabriel Moreno. The evolution of ecological specialization. *Annual review of Ecology and Systematics*, pages 207–233, 1988.

[173] Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025.

[174] K. Eric Drexler. Reframing superintelligence: Comprehensive ai services as general intelligence. Technical Report 2019-1, Future of Humanity Institute, University of Oxford, 2019.

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

[176] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. doi: 10.1126/science.177.4047.393.

[177] John H. Holland. Complex adaptive systems. *Daedalus*, 121(1):17–30, 1992.

[178] Akarsh Kumar, Jeff Clune, Joel Lehman, and Kenneth O Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*, 2025.

[179] Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022.

[180] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[181] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623226.

[182] Jack Foxabbott, Sam Deverett, Kaspar Senft, Samuel Dower, and Lewis Hammond. Defining and mitigating collusion in multi-agent systems. In *Multi-Agent Security Workshop @ NeurIPS'23*, 2023.

[183] Rauno Arike, Elizabeth Donoway, Henning Bartsch, and Marius Hobbhahn. Technical report: Evaluating goal drift in language model agents. *arXiv preprint arXiv:2505.02709*, 2025.

[184] Fernando E. Rosas, Bernhard C. Geiger, Andrea I Luppi, Anil K. Seth, Daniel Polani, Michael Gastpar, and Pedro A. M. Mediano. Software in the natural world: A computational approach to hierarchical emergence. *arXiv preprint arXiv:2402.09090*, 2024.

[185] Erik Hoel. Causal emergence 2.0: Quantifying emergent complexity. *arXiv preprint arXiv:2503.13395*, 2025.

[186] Jakub Bielawski, Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. Heterogeneity, reinforcement learning, and chaos in population games. *Proceedings of the National Academy of Sciences*, 122(25):e2319929121, 2025. doi: 10.1073/pnas.2319929121.

[187] Maria Alejandra Ramirez, George Datseris, and Arne Traulsen. Chaos and noise in evolutionary game dynamics. *arXiv preprint arXiv:2504.00028*, 2025.

[188] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.

[189] Andrew Critch and David Krueger. AI research considerations for human existential safety (ARCHES). *arXiv preprint arXiv:2006.04948*, 2020.

[190] Andrew Critch and Stuart Russell. TASRA: A taxonomy and analysis of societal-scale risks from AI. *arXiv preprint arXiv:2306.06924*, 2023.

[191] Philipp Koralus. The philosophic turn for ai agents: Replacing centralized digital rhetoric with decentralized truth-seeking. *arXiv preprint arXiv:2504.18601*, 2025.

[192] Tomer Shadmy and Katrina Ligett. Reimagining decentralized ai. In *Proceedings of the 2024 Symposium on Computer Science and Law*, CSLAW '24, page 16–23, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703331. doi: 10.1145/3614407.3643701.

[193] Aviv Ovadya, Kyle Redman, Luke Thorburn, Quan Ze Chen, Oliver Smith, Flynn Devine, Andrew Konya, Smitha Milli, Manon Revel, Kevin Feng, Amy X. Zhang, Bilva Chandra, Michiel A. Bakker, and Atoosa Kasirzadeh. Position: Democratic AI is possible. the democracy levels framework shows how it might work. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81930–81961. PMLR, 13–19 Jul 2025.

[194] Yunke Wang, Yanxi Li, and Chang Xu. Position: Ai scaling: From up to down and out. *arXiv preprint arXiv:2502.01677*, 2025.

[195] Sayash Kapoor, Noam Kolt, and Seth Lazar. Position: Build agent advocates, not platform agents. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81617–81633. PMLR, 13–19 Jul 2025.

[196] Michael I. Jordan. A collectivist, economic perspective on ai. *arXiv preprint arXiv:2507.06268*, 2025.

[197] Marvin Minsky. *Society of mind*. Simon and Schuster, 1986.

[198] Moshe Tennenholtz. Program equilibrium. *Games and Economic Behavior*, 49(2):363–373, 2004. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2004.02.002.

[199] Andrew Critch, Michael Dennis, and Stuart Russell. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory. *arXiv preprint arXiv:2208.07006*, 2022.

[200] James Evans, Benjamin Bratton, and Blaise Agüera y Arcas. Agentic ai and the next intelligence explosion. *Science*, 391(6791):eaeg1895, 2026. doi: 10.1126/science.aeg1895.

[201] Diego Calvanese, Angelo Casciani, Giuseppe De Giacomo, Marlon Dumas, Fabiana Fournier, Timotheus Kampik, Emanuele La Malfa, Lior Limonad, Andrea Marrella, Andreas Metzger, Marco Montali, Daniel Amyot, Peter Fettke, Artem Polyvyanyy, Stefanie Rinderle-Ma, Sebastian Sardiña, Niek Tax, and Barbara Weber. Agentic business process management: A research manifesto. *Information Systems*, 140:102738, 2026.

[202] Bargav Jayaraman, Virendra Marathe, Hamid Mozaffari, William F. Shen, and Krishnaram Kenthapadi. Permissioned LLMs: Enforcing access control in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[203] Jinhwa Kim and Ian G. Harris. Context misleads llms: The role of context filtering in maintaining safe alignment of llms. *arXiv preprint arXiv:2508.10031*, 2025.

[204] J.H. Saltzer and M.D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939.

[205] Tsimur Hadeliya, Mohammad Ali Jauhar, Nidhi Sakpal, and Diogo Cruz. When refusals fail: Unstable safety mechanisms in long-context llm agents. *arXiv preprint arXiv:2512.02445*, 2025.

[206] Zheng Wen, Doina Precup, Benjamin Van Roy, and Satinder Singh. Capacity-constrained continual learning. *arXiv preprint arXiv:2507.21479*, 2025. doi: 10.48550/arXiv.2507.21479.

[207] Amartya Hatua, Trung T. Nguyen, Filip Cano, and Andrew H. Sung. Machine unlearning using forgetting neural networks. *arXiv preprint arXiv:2410.22374*, 2025.

[208] David Abel. A theory of abstraction in reinforcement learning. *arXiv preprint arXiv:2203.00397*, 2022.

[209] William Hamilton, Mahdi Milani Fard, and Joelle Pineau. Efficient learning and planning with compressed predictive states. *Journal of Machine Learning Research*, 15(1):3395–3439, 2014.

[210] David Krueger, Tegan Maharaj, and Jan Leike. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*, 2020.

[211] Junhong Lin, Xinyue Zeng, Jie Zhu, Song Wang, Julian Shun, Jun Wu, and Dawei Zhou. Plan and budget: Effective and efficient test-time scaling on reasoning large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[212] Hanbing Liu, Lang Cao, Yuanyi Ren, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han, and Dongmei Zhang. Bingo: Boosting efficient reasoning of llms via dynamic and significance-based reinforcement learning. *arXiv preprint arXiv:2506.08125*, 2025.

[213] Stuart Armstrong and Benjamin Levinstein. Low impact artificial intelligences. *arXiv preprint arXiv:1705.10720*, 2017.

[214] Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174, 2020.

[215] Usman Anwar, Julianna Piskorz, David D Baek, David Africa, Jim Weatherall, Max Tegmark, Christian Schroeder de Witt, Mihaela van der Schaar, and David Krueger. A decision-theoretic formalisation of steganography with applications to llm monitoring. *arXiv preprint arXiv:2602.23163*, 2026.

[216] Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. IDs for AI systems. *arXiv preprint arXiv:2406.12137*, 2024. Presented at the NeurIPS 2024 RegML Workshop.

[217] Shanshan Han, Qifan Zhang, Weizhao Jin, and Zhaozhuo Xu. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2026.

[218] Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

[219] Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. *arXiv preprint arXiv:2410.11782*, 2025.

[220] Xie Yi, Zhanke Zhou, Chentao Cao, Qiyu Niu, Tongliang Liu, and Bo Han. From debate to equilibrium: Belief-driven multi-agent LLM reasoning via bayesian nash equilibrium. In *Proceedings of the 42nc International Conference on Machine Learning*, 2025.

[221] Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, Chengpei Tang, Jian Wang, and Keze Wang. Cost-effective communication: An auction-based method for language agent interaction. *arXiv preprint arXiv:2511.13193*, 2025.

[222] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 162–169, Arlington, Virginia, USA, 2004. AUAI Press. ISBN 0974903906.

[223] Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1.

[224] Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. PMLR, 22–25 Jul 2020.

[225] Willem Schooltink and Fabio Massimo Zennaro. Aligning graphical and functional causal abstractions. In Biwei Huang and Mathias Drton, editors, *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pages 704–730. PMLR, 07–09 May 2025.

[226] Eigil Fjeldgren Rischel. The category theory of causal models. *Master's thesis, University of Copenhagen*, 2020.

[227] Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence 2017*, 2017.

[228] Dilip Arumugam and Benjamin Van Roy. Deciding what to model: value-equivalent sampling for reinforcement learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[229] Sergio Calo, Anders Jonsson, Gergely Neu, Ludovic Schwartz, and Javier Segovia-Aguas. Bisimulation metrics are optimal transport distances, and can be computed efficiently. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[230] Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for complex simulation models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[231] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2170–2179. PMLR, 09–15 Jun 2019.

[232] Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021.

[233] Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[234] Florent Delgrange, Ann Nowe, and Guillermo Perez. Wasserstein auto-encoded MDPs: Formal verification of efficiently distilled RL policies with many-sided guarantees. In *The Eleventh International Conference on Learning Representations*, 2023.

[235] Benjamin Patrick Evans, Leo Ardon, and Sumitra Ganesh. Modelling bounded rational decision-making through wasserstein constraints. *arXiv preprint arXiv:2504.03743*, 2025.

[236] Tong Mu, Stephan Zheng, and Alexander R Trott. Modeling bounded rationality in multi-agent simulations using rationally inattentive reinforcement learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[237] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

[238] Tyler Malloy, Chris R Sims, Tim Klinger, Miao Liu, Matthew Riemer, and Gerald Tesauro. Capacity-limited decentralized actor-critic for multi-agent games. In *2021 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2021.

[239] Tomáš Gavenčiak, David Hyland, Lancelot Da Costa, Michael J. Wooldridge, and Jan Kulveit. Path divergence objective: Boundedly-rational decision making in partially observable environments. In *The First Workshop on NeuroAI @ NeurIPS2024*, 2024.

[240] Marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, page 1785, 1785.

[241] Christian Cachin and Marko Vukolić. Blockchain consensus protocols in the wild. *arXiv preprint arXiv:1707.01873*, 2017.

[242] Lars Brünjes, Aggelos Kiayias, Elias Koutsoupias, and Aikaterini-Panagiota Stouka. Reward sharing schemes for stake pools. In *2020 IEEE european symposium on security and privacy (EuroS&p)*, pages 256–275. IEEE, 2020.

[243] Benjamin Samuel Ruben, William Lingxiao Tong, Hamza Tahir Chaudhry, and Cengiz Pehlevan. No free lunch from random feature ensembles: Scaling laws and near-optimality conditions. In *Forty-second International Conference on Machine Learning*, 2025.

[244] Junsol Kim, Shiyang Lai, Nino Scherrer, Blaise Agüera y Arcas, and James Evans. Reasoning models generate societies of thought. *arXiv preprint arXiv:2601.10825*, 2026.

[245] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

[246] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[247] Siwei Han, Kaiwen Xiong, Jiaqi Liu, Xinyu Ye, Yaofeng Su, Wenbo Duan, Xinyuan Liu, Cihang Xie, Mohit Bansal, Mingyu Ding, Linjun Zhang, and Huaxiu Yao. Alignment tipping process: How self-evolution pushes llm agents off the rails. *arXiv preprint arXiv:2510.04860*, 2026.

[248] Erik Jones, Anca Dragan, and Jacob Steinhardt. Adversaries can misuse combinations of safe models. *arXiv preprint arXiv:2406.14595*, 2024.

[249] Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, 2025.

[250] Florian Carichon, Aditi Khandelwal, Marylou Fauchard, and Golnoosh Farnadi. The coming crisis of multi-agent misalignment: Ai alignment must be a dynamic and social process. *arXiv preprint arXiv:2506.01080*, 2025.

[251] Stephen M Omohundro. The basic ai drives. In *Artificial intelligence safety and security*, pages 47–55. Chapman and Hall/CRC, 2018.

[252] Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

[253] Emanuele La Malfa, Gabriele La Malfa, Samuele Marro, Jie M. Zhang, Elizabeth Black, Michael Luck, Philip Torr, and Michael Wooldridge. Large language models miss the multi-agent mark. *arXiv preprint arXiv:2505.21298*, 2025.

[254] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

[255] Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025.

[256] Zhixuan He and Yue Feng. Unleashing diverse thinking modes in llms through multi-agent collaboration. *arXiv preprint arXiv:2510.16645*, 2025.

[257] Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Assessing collective reasoning in multi-agent llms via hidden profile tasks. *arXiv preprint arXiv:2505.11556*, 2025.

[258] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

[259] Zhiwei Zhang, Xiaomin Li, Yudi Lin, Hui Liu, Ramraj Chandradevan, Linlin Wu, Minhua Lin, Fali Wang, Xianfeng Tang, Qi He, et al. Unlocking the power of multi-agent llm for reasoning: From lazy agents to deliberation. *arXiv preprint arXiv:2511.02303*, 2025.

[260] Christian Schroeder de Witt, Klaudia Krawiecka, Igor Krawczuk, Ben Hagag, William L. Anderson, Peter Belcak, Ben Bucknall, Xiaohong Cai, Ayush Chopra, Doron Cohen, Ron F. Del Rosario, Andis Draguns, Annie Gray, Keren Katz, Vasilios Mavroudis, Jaron Mink, Sumeet Ramesh Motwani, Jonathan Petit, Leif-Sebastian Rembeck, Chandler Smith, John Sotiropoulos, Steven Young, Sarah Scheffler, and Mary Llewellyn. Open challenges in multi-agent security: Towards secure systems of interacting ai agents. *arXiv preprint arXiv:2505.02077*, 2026.

[261] Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark Van Der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 4345–4359. PMLR, 13–19 Jul 2025.

Agent Component	Possible bounds
Data Perception	Data access restrictions [202], input filtering [203]
Action and Interaction	Restricted tools [30], permissions [204], structured protocols [40]
Memory and Compression	Context limits [205], capacity constraints [206], forgetting [207]
Internal Representations	Abstracted [208] or compressed [209] state representations
Planning and Compute	Myopic training [210], planning budgets [211], reward shaping [212]
Goals and Preferences	Narrow tasks [102], impact penalties [213], regularisation [55, 214]

Table 1: A taxonomy of agent components with examples of possible methods for bounding each component.

A Related Frameworks

Our vision shares similarities with Minsky's Society of Mind [197], which argues that complex phenomena such as human intelligence emerge from the structured interaction of many simple agents. Likewise, our proposal is closely aligned with Drexler's *Comprehensive AI Services* (CAIS) [174], which makes the case that a collection of bounded AI services would implement the functionality of a monolithic generally-intelligent system while enabling us to address many of the accompanying risks. BMAS inherit the modular spirit of CAIS, but emphasises the potential benefits of controlled *agency*, i.e., systems that can plan, interact, respond to incentives, and create strategic dependencies, which can make their interfaces and behaviours more legible and reliable to other members of the community [196, 198, 199]. We argue that this shift is necessary given the current interest in and usefulness of agentic AI systems. Moreover, this shift matters because once components are agentic, the design landscape shifts towards delegation, contracting, reputation, incentives, social choice, and institutional design, enabling the application of an abundance of tools that humanity has developed for the governance of our societies and organisations to multi-agent systems.[5] This focus mirrors that of agentic process management (APM) [201], which highlights the need for new mechanisms for integrating agentic AI into existing processes, and effectively aligning them towards organisational goals.

B Towards a Framework for Bounded Agency and BMAS

As we have argued, a useful framework for bounded agency and BMAS requires a formal taxonomy of resource constraints for AI agents. Importantly, such a taxonomy should reflect the multi-faceted nature of capability and provide a concrete means of targeting abstract components of an agent's decision-making process, including planning, world modelling, and perception, within modern machine learning pipelines. Each component can be bounded in different ways, and different bounds imply different safety, performance, and efficiency properties. Table 1 describes in more detail different methods could be utilised to bound different components of an agent.

The Space of Tasks Complementarily to understanding agent capabilities, it is also important to understand the space of tasks and how they relate to different capability axes, as well as one another. Some important **characteristics of tasks** identified by Tomašev et al. [82] include: (1) **Complexity** (number of sub-steps and sophistication of reasoning required); (2) **Criticality** (importance, priority, and severity of consequences associated with failure); (3) **Uncertainty** (level of ambiguity in inputs + environment, and probability of success); (4) **Duration** (expected time frame for reasonable execution); (5) **Cost** (economic/computational costs required to execute the task); (6) **Resource requirements** (tools, data, human access/capabilities); (7) **Constraints** (operational, ethical, or legal); (8) **Verifiability** (difficulty of validating processes and outcomes, particularly when side-effects are involved); (9) **Reversibility** (degree to which changes can be undone); (10) **Contextuality** (volume and sensitivity of external state, history, or environmental awareness required for effective task execution); and (11) **Subjectivity** (extent to which success criteria are matters of preference vs fact). The authors also map five **pillars for a delegation framework**: **dynamic assessment** (granular inference of agent state using task decomposition and assignment), **adaptive execution** (handling context shifts using adaptive coordination mechanisms), **structural transparency** (auditability

[5] Please refer to the following references for recent complementary discussions of these issues [81, 82, 166, 200].

of process and outcome using monitoring and verifiable completion), **scalable markets** (efficient, trusted coordination using trust, reputation, and multi-objective optimisation), and **systemic resilience** (preventing systemic failures using security methods and permission handling).

Figure 2 demonstrates how BMAS might be combined with task knowledge to achieve efficient performance whilst avoiding the safety risks associated with the broad capabilities of monolithic models. As already discussed, every task requires a different assignment of resources across different axes, such as knowledge and reasoning, each corresponding to a different corner of the radar plots in Figure 2. As a result, a task defines a profile of capabilities, denoted by a dotted green line. We refer to this as the *task profile*. Similarly, an agent or system is assigned resources along the same axes. Hence, there exists a profile of capabilities corresponding to the system (or agent), as indicated by the shaded regions, which we refer to as the *system (or agent) profile*. Clearly, if the system profile encloses the task profile then the system has the required resources to execute the task. However, over-assignment of resources along any axis may enable unsafe behaviour. The threshold at which a system has enough of a given resource to engage in unsafe behaviour is dependent on the resource type, context and task domain. This naturally induces a capability boundary, denoted by the red line, which the system must respect so as to avoid the capability for unsafe behaviour. We refer to this as the *safety boundary*.

A generalist, monolithic agent must have a broad system profile in order to cover a wide range of task profiles. However, this implies that a monolithic agent is likely to violate the safety boundary for a wide range of task domains as well. Meanwhile, a BMAS can maintain a set of narrow agents with specialised and limited system profiles that can be carefully and dynamically composed to subsume a given task profile whilst respecting the safety boundary of the task domain, as depicted in Figure 2. Before moving on, we highlight several caveats to this argument. In Figure 2, the system profile of the BMAS is derived by taking the convex hull of agent profiles. In practice, agent capabilities may combine in different ways depending on context and resource type. Likewise, the safety boundary displayed in Figure 2 assumes that an excess of a single resource is required to engage in unsafe behaviour. However, it may be the case that a combination of different but limited resources is sufficient for unsafe behaviour. In such cases, the task domain implicitly defines a *risk profile* rather than a safety boundary, with which the system profile cannot intersect. Finally, the resource axes proposed in Figure 2 are largely illustrative. A key step towards making BMAS a reality lies in identifying the key resource axes that are relevant and controllable, as well as how they depend on each other.

B.1 Capability Framework Desiderata

Some desiderata that a useful framework for modelling bounded capabilities might satisfy include: (1) **Component-specificity:** the framework should distinguish bounds on the different agent components outlined in Table 1; (2) **Task-relativity:** the same bound could be harmless on a routine task and catastrophic on a long-horizon task; (3) **Compositionality:** the framework should allow us to predict how individual capability limits combine when agents interact; (4) **Efficient measurability**: bounds should correspond to quantities that can be efficiently estimated before deployment and monitored during use; (5) **Realisability:** the framework should identify intervention points in existing points in the AI development pipeline such as pretraining, fine-tuning, reinforcement learning, tool permissioning, scaffolding, inference-time routing, and post-deployment monitoring; (6) **Metacognitive Awareness:** systems of bounded agents should contain either localised, shared, or distributed knowledge of how different components are bounded, when these bounds make them unreliable, and how to deal with such situations.

B.2 BMAS Framework Desiderata

A framework for BMAS would ideally possess the following properties: (1) **Architectural Appropriateness:** The framework should identify when a task should be approached using a single generalist, several specialists, an ensemble, a debate, a market, a human, or several humans; (2) **Communication Specificity:** Messages between system components should be legible, bounded, attributable, and resistant to dangerous exploitation through steganography [129, 215] and manipulation, such as prompt injection attacks and harmful persuasion; (3) **Identity and Reputation Provenance:** The framework should address issues of agent identity and reputation [216], with defences against Sybil attacks, resistance to vendors and users opening fresh accounts to evade the consequences of poor

reputation, and mechanisms for recovering from honest mistakes; (4) **Compositional Safety Guarantees:** Where possible, the framework should be able to assess and quantify when components can combine into overall systems with safe system-level capabilities and propensities. More generally, an *algebra of agent composition* is required to understand how capabilities compose under different interaction mechanisms; (5) **Recoverability-By-Design:** Failures should be detectable, localised, and amenable to repair without compromising system-wide performance or viability.

BMAS also require **distributed safety methodologies** designed to mitigate the unique risk profile of AI multi-agent systems [166], moving beyond single model evaluations to address risks such as resource misallocation, compositional capability scaling, and coordination failure. Furthermore, BMAS require **orchestration frameworks** grounded in the established fields of game theory, social psychology and economics, which enable the design of robust interaction protocols and incentive mechanisms. In particular, orchestration frameworks must ensure that the performance gains from implementing multi-agent systems outweigh the costs incurred by parallel agent deployment and coordination [75, 90, 91, 217]. Potential approaches include careful design of communication structures [218, 219], leveraging game-theoretic concepts for coordination [220, 221], and routing subtasks to appropriately sized models/tools [85].

C Potential Solutions and Open Problems

C.1 Bounding Capabilities

Several existing research directions can be adopted in the service of bounding agent capabilities. Abstraction learning methods can force agents to learn a simplified model of their environment[105, 206, 208, 222–234]. Information-theoretic regularisers can be used with reinforcement learning methods to penalise excessive planning, control, or policy complexity [235–239]. Tool and data permissions can enforce least-privilege access and execution [202]. Model routing and cascades can allocate smaller specialised agents to routine, simple tasks and reserve expensive generalists for tasks that require them [85, 91]. Resource-rational reasoning frames the routing decision itself as a cost–accuracy trade-off that bounded principals must take into account [151].

Even given this rich history of background to draw on, several critical open problems remain. Can we restrict a single component, such as planning depth, without the agent using other resources to compensate? Can useful capabilities be separated from dangerous ones, or do some tasks necessarily require abilities that are dual-use? In such cases, how do we implement appropriate safeguards to prevent harmful uses of such abilities? How can we estimate the resources required for a task before solving it? If two tasks require similar resources, does competence on one imply competence on the other? Can a bounded agent be built from scratch, or is it more practical to specialise a generalist and then constrain it? How should a system choose between bounded specialists, a single generalist, or a hybrid mixture? The answers to these questions fundamentally determine whether BMAS can be an engineering discipline rather than an abstract concept.

C.2 Bounded Multi-Agency

The most important open problem is a theory of when a multi-agent system outperforms a monolithic system under fixed budgets for different system components, such as compute, memory, latency, information, and risk. Current results are promising, but typically apply in specific domains such as Mixture-of-Experts architectures, LLM agent orchestration, ensembles, software-engineering agents, and classical distributed systems. To better understand the opportunities and pitfalls of these systems, the field should develop BMAS-specific benchmarks and theory that generalise beyond a small number of hand-picked scenarios that are not representative of the wide range of tasks that AI systems may be useful for.

The case for distributing decision authority across multiple bounded agents has formal antecedents. Condorcet's jury theorem shows that aggregating the independent judgements of agents that are individually only marginally more accurate than chance can produce collective decisions that converge to correctness as the number of voters grows [240]. Blockchain consensus protocols rely on a related principle: as the number of independent participants grows, the cost of obtaining a malicious majority grows faster than the cost of honest participation [241]. Both depend critically on *independence*, because correlated voters or stake-pooling [242] participants can cause the guarantees to collapse,

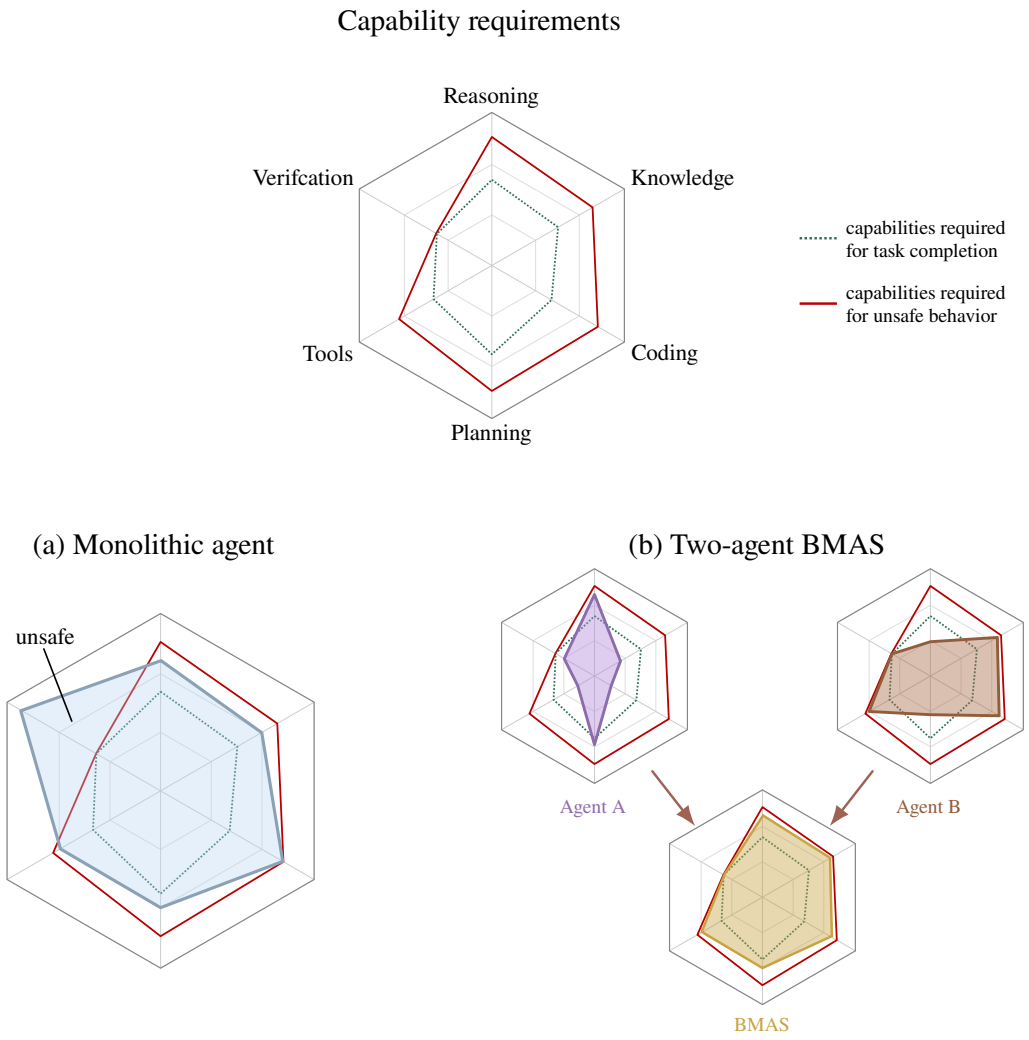

Figure 2: Illustrative capability radar plots for a given task. The top radar plot shows a minimum requirement region for task completion (green), referred to as the task profile, and thresholds for unsafe behaviour (red), referred to collectively as the safety boundary. Shaded exterior regions, referred to as system (or agent) profiles, indicate system (or agent) capabilities. A monolithic agent must adopt a broad system profile that is likely to breach the safety boundary. A BMAS can instead combine specialised agents with distinct profiles. If agents are composed correctly, the yellow BMAS profile allows us to achieve the combined task-relevant capabilities without breaching the safety boundary.

and BMAS may inherit this dependency. The robustness benefit of multiple bounded agents is multiplicative in the probability of independent component failure, but this usually only holds if the components are independent. Cognitive monocultures, shared training data, common tool dependencies, and imitation among agents all introduce correlations that can lead to system-level instability or collapse. Understanding how safety properties of BMAS scale as the number of agents and their interaction structures change is thus of high importance to providing (probabilistic) guarantees on essential system behaviours at larger scales.

Several other questions are also critical to address. How do we prevent agents from developing compact but human-illegible codes when communication budgets are tight? How do we assign credit or blame when privacy constraints prevent full transparency? How should a BMAS evaluate whether a task has been decomposed safely? How should reputation work when agents can be copied, fine-tuned, or redeployed under new identities? How do current AI models respond to incentives? How should liability be assigned when a harm emerges from interactions rather than from a single component? Many of these are principal-agent problems by another name, and the contracting, screening, and incentive-design machinery developed in economics for such problems is highly relevant to this setting. These are reasons to study BMAS before agent ecosystems become entrenched, without a theory that enables deliberate, thoughtful choices about which trade-offs are acceptable.

D Other Counter-Arguments

Monolithic systems can be more efficient than BMAS This is true for many tasks, but we argue that the reverse is also true for many tasks of importance. Monolithic systems may be more efficient due to the reduced need to coordinate different system components [243]. Recent work has demonstrated that reasoning models can simulate multi-agent-like interactions ("societies of thought") [244] to effectively reason, and make use of the efficiency gains from preserving a single context by utilising Key-Value (KV) caches [245]. The appropriate reconciliation of the views is that of *architectural pluralism*: different architectures should be used in different scenarios. We should use monoliths when the task benefits from unified context and sequentially dependent execution, and use BMAS when modularity, parallelism, privacy, verifiability, or governability justify the additional coordination cost [90, 91, 101].

The current trajectory has favoured monoliths Frontier AI development is shaped by powerful incentives to build a single product that captures many use cases, centralises data, and predictably follows scaling laws [246]. This has led to AI labs around the world entering into a race to develop highly generally-capable systems, which has drawn resources and attention away from alternative designs that may be more valuable along dimensions other than economic viability. On the other hand, developing and deploying BMAS can be more difficult because it additionally requires the consideration of standards, protocols, interfaces, distributed ownership, and ecosystem governance. However, if the broader community contributes to BMAS research, we believe that the benefits outlined in Section 3 can be attained and widely distributed.

Distributed systems can fail in novel ways Beyond the canonical alignment failures of single agents, BMAS can face potentially different pathologies that emerge from the interaction of different components. A recently highlighted example of this is *alignment-tipping processes* [247] where misaligned outputs of one agent can bias the context of other agents and lead to cascading failures. Organised agent collectives may also be able to hack into systems that no single agent can hack into through coordinated action [93, 248, 249]. Large populations of relatively simple agents can generate emergent behaviours that are difficult to predict, identify, and steer [188, 250], and may coordinate to bypass safeguards designed for single models [248], decompose a complex attack [93, 249] or develop instrumental convergence [251] at the system level. It is therefore insufficient to reason about bounded *individual* agents' capabilities and goals. Instead, evaluations and guardrails must act across multiple scales.

Likewise, risks may emerge from **capability collapse due to coordination failure** (e.g., models may not be capable of sustaining diversity in multi-agent settings, or may fall out of distribution when interacting with other models [252–254]). Deploying highly capable models to interact with each other does not guarantee preservation of such capabilities. Sycophancy in interactive systems [255], mode collapse in multi-turn conversations [254], loss of role diversity [256] among others [90], are

examples of such failure modes. Recent developments in benchmarks and principled evaluations of multi-agent AI systems are focused on tackling these limitations [249, 257–259].

Recent reports have highlighted that multi-agent systems do not merely distribute risk but may also amplify it through complex feedback loops [188, 260]. Multi-agent systems comprised of bounded agents introduce novel safety issues as well as safety benefits. **Mis-allocation of resources to individual agents** presents a significant risk. For example, an agent with limited foresight and planning capability may pursue disastrous myopic policies on long-horizon tasks. Similar risks are posed by over-specialisation. Fine-tuning agents on narrow tasks may induce harmful behaviour on unseen problems [111]. These points emphasise the importance of designing protocols for agent interactions so as to minimise the opportunities for them to be undesirably exploited. In addition, these risks demand that agents are cognizant of their own resource limitations and can account for model and objective misspecification, distinguish aleatoric from epistemic uncertainty [261], model their own ignorance, and communicate when a task is unsafe to perform [128]. However, modelling and designing boundedly rational agents can act as a natural hedge against objective misspecification [146], and if a multi-agent system has redundancy (i.e., multiple agents doing or evaluating the same things), then aggregation over this group can be used to reduce epistemic uncertainty via committee formation.