Progression from Sparsity-Based Modularity to Emergent Multi-Agent Roles

Tong Zhang

Cognitive Science and Psychology: Archetypes, Roles, and Internalization

Cognitive theories offer rich metaphors for thinking about modular AI systems and emergent roles. Carl Jung introduced the idea of a *collective unconscious* populated by universal archetypes – recurring characters or patterns (the Hero, the Mother, the Trickster, etc.) shared across human psyches. These archetypes represent timeless roles or ideas that resonate with human experience, though they are not rigidly defined behaviors. The persona, in Jungian terms, is one's social face or role; importantly, one's chosen persona can be influenced by archetypes from the collective unconscious. For example, an individual might consciously or unconsciously adopt a "hero" persona in public life – striving to be courageous and resilient – essentially internalizing the Hero archetype as a guiding role. This reflects how intrinsic role-patterns (archetypes) can shape outward behavior (persona), a concept that intriguingly parallels how AI agents might assume specialized roles based on ingrained patterns or learned biases in their models.

Developmental psychology also emphasizes role internalization as key to cognition. George H. Mead's theory of self posits that children develop a sense of self by internalizing the roles and attitudes of others. In early childhood play, kids often imitate different roles with no overarching structure; but in more organized group games, the child must take into account the perspective of all other players. By "playing" at being each member of the team, the child gradually internalizes the generalized other – essentially an internal model of the group's expectations. Mead argued that through this process, the organized social role of the group (e.g. a baseball team) enters the child's mind as an integrated perspective, forming part of their own identity. In effect, the child's mind becomes a multi-agent system of internalized roles: an internal conversation between "self" and the generalized other that guides behavior. Similarly, Lev Vygotsky highlighted how interpersonal processes (like dialogues with a teacher or parent) are internalized to become intrapersonal cognitive strategies. For instance, children's private speech (talking through a task aloud) eventually becomes inner speech, a silent self-dialogue used for planning and self-reflection. This transition from social, external guidance to internal, self-guided thought demonstrates how learning to play multiple roles (teacher vs. student, parent vs. child) can yield internal cognitive modules or voices that later collaborate within one mind.

The "society of mind" metaphor, famously proposed by Marvin Minsky, directly envisions the mind as a collection of semi-autonomous sub-agents. Each mental agent is specialized (e.g. for vision, language, motor control, emotional monitoring) and works in parallel; intelligence and even consciousness are emergent properties of their interactions rather than the function of any single monolithic process. According to Minsky, no single agent is conscious or intelligent on its own, but when enough specialized agents collaborate in the right ways, they give rise to the unified capabilities we associate with a mind. This view aligns remarkably well with modern AI trends: instead of a single all-knowing model, we might achieve more robust intelligence via multiple specialized models (agents) that communicate and cooperate. In cognitive terms, disorders like dissociative identity disorder have even been interpreted through the society-of-mind lens, as cases where the coordination between internal agents breaks down, leading to distinct "personalities" that don't integrate. The key takeaway is that human cognition appears to leverage modularity and specialization, with our brains and minds containing many task-specific processes (perception, language, planning, social reasoning) that can act semi-independently and then bind together through communication (e.g. via language or conscious attention). This provides conceptual support for AI architectures composed of multiple modules

or agents that develop emergent roles – analogous to archetypes or internalized social roles – which together produce complex, flexible intelligence.

Theoretical Neuroscience: Sparsity, Specialization, and Predictive Coding

Neuroscience further reinforces the value of sparsity and functional specialization, both as explanations of brain function and as inspirations for AI. The cerebral cortex is highly modular: distinct regions and circuits specialize in particular functions (visual cortex for sight, auditory cortex for sound, motor cortex for actions, etc.). Even within a single modality, hierarchical specialization is evident – for example, primary visual cortex (V1) contains neurons that fire for simple features like oriented edges, while higher visual areas have neurons tuned to complex stimuli like faces or specific objects. This suggests the brain organizes knowledge into reusable modules, each expert in a facet of the world. Importantly, the activity in the brain is also extremely sparse at any moment: out of billions of neurons, only a small fraction fire in response to a given stimulus. Such sparse coding is metabolically efficient and carries more information per spike by avoiding redundancy. In fact, theoretical work shows that sparse neural codes naturally arise if one assumes the environment has many latent causes but only a few are active at once – under an efficient coding principle, neurons develop representations where only a few neurons respond to any given input. This matches experimental findings: visual neurons tend to have selective firing (a given neuron might respond only to a specific angle or motion direction) and most neurons stay silent for any particular image. Sparse distributed representations are thus a biological instance of "mixture-of-experts"-like processing – the brain has many potential feature detectors, but only a sparse subset "gates in" for each stimulus, reminiscent of a neural mixture-of-experts where only a few experts are activated per input.

Beyond static specialization, the brain also exhibits dynamic, context-dependent routing of information, which parallels adaptive modularity in AI. Cognitive neuroscientists Baars and Dehaene have proposed the Global Workspace Theory, where many unconscious specialist processes feed information into a limited-capacity "global workspace" (working memory) that broadcast selectively to the rest of the system. Attention is the mechanism that selects a small subset of content (perhaps the most relevant or novel) to enter this global workspace, and then that content is made globally available to influence other modules. This theory implies a sparse communication bottleneck: only a few items or signals are globally broadcast at a time. The global neuronal workspace model (Dehaene, 2017) specifically suggests that conscious thought corresponds to a sparse activation of higher-level cortical circuits forming a broadcast message, while countless other processors remain active but non-conscious in the background. In essence, the brain may be leveraging a sparsely-activated modular architecture: numerous experts work in parallel (largely unconscious), and a controlled sparsity (attention/consciousness) lets certain expert results be shared and combined for flexible, novel problem-solving. This biological design principle motivates AI architectures that incorporate attention and gating to regulate communication among modules, just as transformers use attention to sparsely connect information or mixture-of-experts layers use gating to activate only a few network components per input.

Another influential neuroscience theory is predictive coding, which casts the brain as a hierarchical prediction machine. In predictive coding models, each cortical area doesn't simply "react" to stimuli – instead it actively predicts its inputs using an internal model, and only the prediction errors (the differences between expected and actual signals) are propagated forward (or backward) in the circuit. Neurons in this framework often have specialized roles: some ("generative" neurons) convey predictions from higher levels downwards, while others ("error units") encode the mismatch and send it upward. This naturally yields sparse, efficient communication – if a prediction is good, very little error signal is transmitted (most of the activity cancels out). Only when something unexpected happens do neurons fire strongly to communicate that news. Thus, sparsity emerges as an implicit differentiation mechanism: by transmitting only error signals, the network isolates what is novel or informative, effectively differentiating signal from redundancy. Theoretical work unifying efficient coding and predictive coding shows that a network optimizing for both past compression and future prediction can end up with neural responses that decorrelate inputs (as efficient coding demands) while also emphasizing features with predictive power. Interestingly, there can be trade-offs: encoding to best compress past inputs vs. best predict future inputs can lead to different optimal codes. The brain likely employs a mix of objectives, which could explain why we see diverse cell types and response

properties (some neurons tightly encode predictable stimuli for fidelity, others highlight unpredictability). For instance, some visual neurons adapt to remove redundant background and highlight surprise elements, aligning with predictive coding, while others robustly track important known features (like an eye tracking a moving object smoothly, effectively predicting its motion). The takeaway for AI is that predictive models with sparse error-driven learning (e.g. variational autoencoders, world models with prediction-error loss) might capture the brain's learning strategy. Indeed, modern deep networks like capsule networks and recurrent independent mechanisms (RIMs) explicitly enforce modules that communicate only through limited interfaces when necessary. Goyal et al. (2020) demonstrated that in a recurrent model with independent modules that update only when triggered by an attention mechanism, the modules self-organize into specialized "experts" handling different factors of the input sequence. This specialization significantly improved generalization, especially on distribution-shift tasks (where certain factors change from training to testing). Such findings echo the brain's approach: modularity + sparse interaction = adaptability. The brain's blend of sparse coding, predictive modeling, and modular specialization provides a blueprint for AI: we expect systems with these properties to be more data-efficient, interpretable, and robust out-of-distribution.

In summary, theoretical neuroscience suggests that an intelligent system benefits from sparsity (only use the resources you need at a time), functional specialization (different parts handle different aspects of the task or input), and predictive modeling (learn a model of the world to drive perception and action). These principles directly motivate the progression from sparse Mixture-of-Experts networks to more complex agent-based architectures with emergent roles and internal world models. The brain's example also underscores the importance of learning the modular structure rather than hard-coding it: human infants are not born with a complete adult role repertoire; they learn and internalize roles through social interaction and self-reflection. Likewise, in AI, we seek mechanisms by which useful modular roles can emerge through learning.

Artificial Intelligence: Multi-Agent Systems, Emergent Roles, and Self-Reflective Agents

In AI research, we see a clear trajectory aligning with these cognitive and neural principles: from sparsely-activated modular networks to multi-agent systems with emergent specialization, and now towards reflective, world-model–driven agent architectures.

Sparsity-based modularity first rose to prominence with techniques like the Mixture-of-Experts (MoE), which enables a large model to consist of many sub-networks ("experts") but activates only a few for each input. For example, Shazeer et al. (2017) introduced a sparsely gated MoE layer in a deep network that could scale to extremely large parameter counts (hundreds of billions) without a proportional increase in computation, by routing each input to only the top-k expert networks. This approach achieved state-of-the-art results in language modeling at unprecedented scales. The success was not just due to scale, but also due to implicit specialization: each expert network can focus on a subset of the input space or certain features, developing niche expertise. Sparsity in the gating ensures that for a given input, only the most relevant experts are updated, which also means each expert's knowledge remains more localized and interpretable. Subsequent research has found that MoEs often learn sensible division of labor among experts (e.g. different experts focusing on different language topics or syntax vs. semantics), although ensuring balanced usage of experts can be challenging (to avoid some experts overpowering others). Nonetheless, the MoE paradigm demonstrated that modularity with learned sparse activation could greatly improve efficiency and even generalization, provided the gating mechanism is well-designed. It effectively performs a kind of implicit differentiation by routing gradients only through the chosen experts, thereby isolating which part of the network should "learn" from each example. We can view this as a first step toward multi-agent systems: each expert is like a simple agent specializing in part of the task, and the router is like a manager assigning the current problem to a small team of experts. However, in standard MoEs the experts do not explicitly communicate or plan; their interaction is limited to the instantaneous routing decision.

Building on modular networks, multi-agent reinforcement learning (MARL) takes modularity into the realm of multiple learning agents interacting in an environment. In cooperative MARL, several agents learn policies that must coordinate to maximize a shared reward, while in competitive or mixed settings agents might have opposing goals. A key research question has been whether roles or specialization emerge among learning agents without being pre-programmed. Recent work shows that indeed, given the right training

setup, agents can spontaneously adopt distinct roles that improve the group's performance. Wang et al. (2020) introduce ROMA (Role-Oriented Multi-Agent RL), a framework where roles are latent variables that condition each agent's policy. Initially, no roles are pre-defined; the agents are identical and freely learn. ROMA then uses regularizers to encourage any emerging roles to be identifiable and useful – essentially pushing agents to specialize on different sub-tasks and to behave similarly if they have the same latent role. Remarkably, under this training, agents did develop meaningful roles in complex team scenarios. For example, in StarCraft II micromanagement tasks, one agent might assume a "tank" role (drawing enemy fire) while another takes a "damage-dealer" role (staying back to deal damage safely). At runtime, each agent's observation feeds into a learned role encoder, and agents with similar role embeddings exhibit similar behavior (sharing experience). Figure 1 of ROMA vividly shows this emergent specialization: an agent with the highest health moved forward to absorb damage (a tank-like behavior), thereby protecting weaker allies, even though no one explicitly told the agents to adopt a tank or support role. Roles in ROMA are dynamic – if that agent's health later drops, some other agent might take over the tank role. The roles also correlated with intuitive metrics like location or unit type, but were not fixed to those – it was an adaptive assignment based on the scenario. Importantly, ROMA achieved state-of-the-art performance on the StarCraft benchmark, showing that emergent division of labor can significantly improve learning efficiency and final reward. The learned roles were identifiable in the sense that one could cluster agents' latent role vectors and see distinct groupings corresponding to specialized behaviors. This provides evidence that multi-agent systems can discover and internalize advantageous role structures (akin to how human teams or societies develop specialized roles), rather than requiring designers to pre-define those roles.

Unsupervised role emergence isn't limited to cooperative teams – even competitive self-play can yield surprising specialization. OpenAI's hide-and-seek experiments (Baker et al., 2020) famously showed that through many rounds of self-play, agents in a simple environment invented tool use and strategy shifts in an open-ended way. In that simulation, hiders and seekers were initially only programmed to chase or evade, but as they co-trained (each learning to exploit the other's weaknesses), they autonomously discovered strategies like using movable boxes to block exits, surfing on ramps to climb over obstacles, etc. Each new strategy by one side created a pressure that led the other side to counter with another emergent strategy – an autocurriculum of increasing complexity. While the roles of "hider" and "seeker" were given by the environment, the specific tactics and sub-roles (e.g. one hider might focus on building barricades while another scouts) were not scripted. This demonstrates that even with very sparse rewards and simple objectives, multi-agent learning can give rise to qualitatively new behaviors and implicit roles as agents co-evolve. Such results support the vision of progressively complex cooperation/competition dynamics leading to open-ended skill acquisition, which some see as essential for generally intelligent agents. In a sense, the multi-agent system as a whole becomes the learner, with each agent's policy shaping the learning environment of the others – a form of co-training where agents mutually bootstrap each other into regimes far beyond the initial conditions.

Beyond pure reinforcement learning, the language-based multi-agent paradigm has gained traction recently, especially with large language models (LLMs) as the agents. Here, multiple LLMs (or multiple instances of a single LLM given different "persona" prompts) interact via natural language to collaborate or compete on tasks. Language is used as the medium for coordination, effectively creating an agent society that communicates in human-like dialogue. A simple example is two LLM agents in a role-play, such as a "developer" and a "assistant", working together to solve a programming task by exchanging messages. These setups have shown benefits like more divergent thinking (different agents can propose different solutions or critiques) and improved factual accuracy (one agent can double-check or question the other). For instance, one study found that having an LLM "debate" or discuss with another LLM can reduce reasoning errors and catch mistakes that a single-pass solution missed. In effect, the agents take on complementary roles (e.g. protagonist vs. devil's advocate, or questioner vs. solver). Such roles are not pre-programmed but can be induced by prompt engineering (providing different background instructions to each agent).

Crucially, researchers are now imbuing these language-agent teams with self-reflective capabilities to enhance coordination and planning. Xiaohe Bo et al. (2024) propose a framework called COPPER for LLM-based multi-agent collaboration, which introduces a learned reflector component. In their setup, multiple LLM agents work on tasks (like multi-hop question answering or collaborative math solving) and periodically generate reflections on their own behavior or the team's strategy. A separate "reflector" model is trained (using reinforcement learning, specifically a counterfactual PPO approach) to take the conversation history and produce insightful feedback or adjustments for the agents. What makes it powerful is that the reflector

is shared and role-aware: it learns to give tailored feedback to each agent according to that agent's role in the system. They achieve this by assigning counterfactual rewards – essentially measuring how much a single agent's reflection improved the final outcome – which helps credit assignment for learning the reflector. The result is a form of meta-learning on top of the multi-agent interaction: the agents not only solve problems, but also learn to critique and refine their approach using an internal feedback loop. Empirically, COPPER showed stronger performance and generalization on tasks like collaborative QA and even chess analysis compared to non-reflective multi-agent baselines. This points to the feasibility and importance of reflective reasoning in agent systems. It's analogous to a team meeting where team members periodically step back to discuss "how are we doing? what should we do differently?" and then adjust their tactics – a level of self-awareness and adaptation beyond reactive policy execution.

Even single-agent LLM systems have demonstrated the value of reflection and world-modeling. The ReAct and Reflexion approaches (2022–2023) showed that an LLM can plan actions in an environment and then analyze its own chain-of-thought for errors or improvements, leading to better problem-solving. For instance, an LLM can be prompted to output not just an answer, but a step-by-step reasoning followed by a self-critique: "Was my reasoning correct? If not, where did I go wrong?" If it finds a flaw, it can attempt the problem again while avoiding the previous mistake. Remarkably, this kind of self-reflection significantly improved accuracy on tasks like math word problems and logic puzzles. One study instructed several popular LLMs to answer multiple-choice questions, then reflect on each wrong answer and try again; across the board, the models improved their performance after self-reflection with high statistical significance. This "introspective learning" is done purely in natural language (the model generates an explanation to itself), hinting that even without weight updates, an agent can learn within its own runtime by examining and adjusting its reasoning trajectory. It's an instance of online, reflective training – the model is effectively training itself using the world model in its head (in this case, its knowledge and the scratchpad of its chain-of-thought).

World-model–driven agents are another frontier carrying these ideas forward. In reinforcement learning, world models refer to an agent's learned model of environment dynamics, which can be used for simulation, planning, and imagination. For example, Hafner et al.'s Dreamer agent learns a compact recurrent world model of the environment's state transitions and reward structure, and then plans actions by "imagining" future trajectories in its latent model. Dreamer v3 showed that a single agent with a learned world model can achieve human-level performance on diverse control tasks (from Atari games to robotic control) by planning in the latent space and training entirely on imagined trajectories rather than expensive real environment steps. The world model enables a form of reflective reasoning: the agent can ask "if I take action A in state S, what might happen?" and use the answer (from its model) to choose the best action, rather than blindly trial-and-error in the real world. Such imagination-based planning is analogous to how humans mentally simulate outcomes ("If I push this chair, will it support the door as a barricade?" – a thought a hider agent might have!). The success of Dreamer and related model-based RL algorithms underscores the importance of endowing agents with an internal model of their environment. These models make the agent more sample-efficient and flexible, especially in non-stationary or combinatorially complex tasks.

When we extend world models to multi-agent systems, we encounter fascinating possibilities: agents can form collective world models and even evolve communication protocols grounded in those models. Nomura et al. (2025) propose a decentralized multi-agent world model where each agent maintains its own representation but shares information with others through a learned communication channel. Their approach uses a form of collective predictive coding: each agent predicts the next state and gets feedback not only from the environment but also via messages from other agents. Through training, the agents develop a shared symbolic language (encoded in the messages) that aligns with actual environment states and events. Notably, they found this approach achieved near-centralized performance in a coordination task even though each agent had only partial observability, and the emergent communication was meaningful – it wasn't just trivial signals, but conveyed information about the environment that the other agent lacked. Because the agents were constrained (they couldn't directly peek at each other's internal state), the learned messages ended up corresponding to useful abstractions about the world (a form of emergent language). This echoes cognitive science ideas about language evolution: a shared world model and the need to cooperate can drive the emergence of shared symbols that represent the world in a convenient way.

In a related theoretical vein, Taniguchi et al. (2024) argue that large language models themselves can be viewed as collective world models. They present a unifying framework called Generative Emergent Commu-

nication which mathematically bridges multi-agent emergent communication and world-model learning. In their view, if multiple agents each have experiences of the world, and they communicate to build a shared understanding, this process can be framed as decentralized Bayesian inference – essentially, each agent's perspective contributes to a global probabilistic model (the "collective world model"). Remarkably, they show that an LLM, which has been trained on vast text (which indirectly encodes many humans' experiences), can be interpreted as encoding a world model that is collective – it has integrated patterns across many data sources, i.e., across "multiple agents' experiences" in the human internet-scale dataset. This provides a theoretical foundation for why LLMs are so powerful: they might function as a distilled representation of the world knowledge shared through language by countless people. Moreover, Taniguchi et al. link this to cognitive development and language evolution. Just as a child learns language (symbols) through interactions (bridging individual cognition with the societal-level language system), multi-agent systems learn communication protocols that become like a language for that domain, and a trained language model captures the statistical structure of those communications.

In short, the boundary between a world model (in the sense of a model-based RL agent) and a language model starts to blur – an LLM can act as a world model for a group of agents, and conversely a group of communicating agents is building a world model through their dialogue. This convergence is particularly exciting for designing agent architectures that use language as a tool for thought (e.g., an agent can simulate a conversation between internal "expert personas" to reason about a problem, effectively using its language ability to model different aspects of the world or different solution paths).

Co-training in an AI context can thus be seen at multiple levels: at the lowest level, mixture-of-experts co-training means different experts train on the portions of data they are most activated for (a form of specialization through sparsity); at the highest level, distinct agents (or agent personas) can train each other by providing diverse experiences (as in self-play yielding an autocurriculum, or an AI assistant being refined by interacting with another AI agent playing the user). In all cases, the system avoids a single monolithic training loop and instead has multiple interacting training loops that induce structure in each other.

A concrete example is OpenAI's InstructGPT/GPT-4 alignment training, where an LLM is improved by engaging in dialogue with humans or AI overseers – effectively a two-agent system (model and human) co-training via feedback. Another example is population-based training in AlphaStar, where many agent instances play StarCraft against each other in a league, resulting in specialist agents (some rush attackers, some economic macro players, etc.) that collectively push the frontier of skill. The league maintained a diverse set of strategies to avoid overfitting to any single opponent, which is a case of explicitly encouraging multiple roles (or "Nash league" equilibrium strategies) to coexist. The AlphaStar endeavor demonstrated that superhuman performance in a very complex, open-ended game could be achieved by a carefully orchestrated community of agents rather than one super-agent – again highlighting that an agent ecosystem with roles may solve problems that a solo agent cannot.

Synthesis: Toward Reflective, Role-Specialized Agent Architectures

Across cognitive science, neuroscience, and AI, a unifying picture emerges: intelligence is most powerful when it is structured as a collection of specialized parts that communicate and adapt. Sparsity is a crucial principle binding these perspectives. In the brain, sparse activation and sparse connectivity yield efficient, disentangled representations; in deep networks, sparse MoE gating allows trillion-parameter models to be tractable. But sparsity is not just about efficiency – it's also an implicit differentiation of concerns. By only activating certain modules or agents for a given context, the system separates "who deals with what," which in turn makes learning credit assignment easier (each module adjusts for the cases it handles) and reduces interference between skills. This forms a basis for modularity, which both cognitive science and AI argue is essential for robustness and generalization.

Building on a sparse modular foundation, we move toward systems of multiple agents or modules with emergent role specialization. The progression mirrors human development: initially we have simple reflex-like experts (analogous to innate modules or early sensorimotor networks), but as the system (child or AI) experiences more complex scenarios – especially social or multi-faceted ones – it benefits from dividing cognitive labor into roles. In multi-agent AI experiments, we literally see roles like "leader/follower,"

"attacker/defender," or "explorer/exploiter" appear without being hardwired, because those roles confer an advantage under the task dynamics. This is analogous to how human groups tend to self-organize roles (even in a playground game one child might naturally take charge as the leader, another as the strategist, etc.). When we allow and encourage agents to adopt roles (through frameworks like ROMA's role encoder or through population diversity maintenance), the system's performance and adaptability improve, indicating that role specialization is not just incidental, but beneficial. Each role can be seen as focusing on a subset of the overall problem – much like Jung's archetypes each reflect a fundamental facet of human experience, AI agents' roles may capture fundamental sub-problems of a task (e.g. offense vs defense, or speed vs accuracy trade-offs).

The final piece is reflective reasoning and world modeling, which introduces a meta-cognitive layer. Having multiple specialized parts is good, but to truly approach higher-order cognition (like planning, theory of mind, and abstract reasoning), a system needs to orchestrate these parts, evaluate outcomes, and adjust strategies on the fly. Humans do this via reflective thinking – we can mentally step back and say "this approach isn't working, maybe I should try a different way or consult a different perspective." In AI, we now see initial glimmers of this: an agent can internally simulate different "experts" debating (as a proxy for different approaches) or use a world model to test action sequences in imagination before committing to them. The COPPER multi-agent reflector is essentially an internal coach that helps a team of agents improve by reflecting on their behavior. Likewise, an LLM agent with a Reflexion loop is doing self-coaching, generating an explanation of what it did wrong and how to avoid it. These are embryonic forms of self-reflection in AI, and they already show measurable gains in performance and generality. Looking ahead, we can envision AI architectures that integrate these elements tightly: imagine an agent that consists of many sub-agent "personas" (each a sparse expert or learned role), all embedded in a shared world model (so they have a common understanding of context), and a global workspace/reflector that enables communication, conflict resolution, and planning among them. Such an architecture would be world-model–driven and role-based. It would operate by, say, having the world model generate hypothetical scenarios, the specialist sub-agents each propose actions or analyses according to their role (one agent might simulate the physical outcome, another checks consistency with mission goals, another evaluates risk/reward), and a reflection mechanism weighs these inputs and decides or learns which course to take. This is somewhat analogous to a human brainstorming with themselves – conjuring multiple viewpoints or "inner voices" before making a decision – which in turn echoes the Jungian idea of archetypes influencing one's behavior or Mead's idea of taking the perspective of the generalized other.

Is such a complex architecture feasible and important? Current research suggests yes on both counts. Feasibility is supported by the pieces already in place: transformers with sparse MoE layers have been deployed at scale in e.g. Google's Switch Transformer and GLaM, proving we can train gigantic modular models; multi-agent role-based training (ROMA, etc.) has solved challenging coordination tasks that were out of reach for monolithic policies; LLM-based agent frameworks show that even today's models can engage in multi-step, multi-agent dialogues to tackle problems like tool use or complex reasoning; and model-based planners like Dreamer show that learning and using world models endows agents with imagination and foresight. The importance of this progression lies in its potential to break through limitations of current AI. By combining sparsity, specialization, and reflection, we address key issues:

- **Generalization:** Sparse modular systems have better out-of-distribution generalization, as independent modules can recombine to handle novel inputs. Emergent roles create a form of causal disentanglement – each role tackles a causal aspect of the task, making the overall behavior more robust to changes in one aspect. A role-based multi-agent team can adapt if one aspect of the environment shifts, by shifting which agent takes the lead, rather than having to relearn the entire task.

- **Scalability:** It's more parameter-efficient and computation-efficient to add experts or agents than to scale a single model. More agents can be trained in parallel, and sparse activation keeps runtime costs manageable. This is akin to how the brain's cortex scaled by adding many columns and areas, rather than one giant neuron – scaling through multiplication of specialists.

- **Interpretability:** Modules and agents with distinct roles are easier to interpret and troubleshoot. In ROMA, one can visualize the role embeddings and see intuitive clusters; in multi-agent dialog, one can

follow each agent's utterances. This modularity by design could help address the black-box nature of deep learning by giving us "handles" (each module/agent can be understood in isolation to a degree).

- **Adaptability and lifelong learning:** A reflective multi-agent system can perform online learning. If the system encounters a new challenge, it might allocate a new role (like recruiting a new expert) on the fly or use its world model to simulate solutions and then incorporate the successful one as a new skill. This is much like human organizations evolving new job roles in response to new challenges, or an individual learning a new problem-solving strategy and integrating it into their repertoire. By having an architecture that inherently supports adding or refining parts without retraining the whole, we move closer to lifelong, curriculum-based learning in AI.

In conclusion, the journey from sparsity-based modular networks to multi-agent systems with emergent roles and reflective reasoning is not just a series of incremental improvements, but a coherent research vision for building more human-like intelligence. Cognitive science gives us the high-level blueprint (a society of mind with internalized roles and self-reflection), neuroscience confirms that this blueprint has efficiency and flexibility benefits (sparse predictive coding and specialized circuits in the brain), and AI research provides early prototypes of each component (from MoEs to ROMA, COPPER, and world-model planners). The synthesis of these disciplines points to agent-based, world-model–driven architectures as a promising path: systems that, like a community of experts, can differentiate tasks, assume roles, communicate in a shared language, and reflect on their performance to continually improve. Such systems could tackle complex problems (which require breaking the problem into parts or roles) far better than a uniform model. They also align with the trend that as AI systems become more embedded in human society, they may need to understand and emulate human-like roles (teacher, assistant, devil's advocate, etc.) and work in teams (either with humans or other AI). In sum, sparsity provides the mechanism (implicit differentiation and efficient specialization), emergent roles provide the structure (modular organization of skills), and reflective reasoning provides the glue and steering (metacognitive oversight and adaptation) for the next generation of intelligent architectures. Early results across domains substantiate the feasibility of this progression, and they highlight that embracing these principles is likely key to achieving both the feasibility and the safety/controllability of AI as it scales. The pieces are coming together for AI systems that are not monolithic black boxes, but rather transparent societies of specialized minds that can reason about the world and about themselves – a long-standing dream in AI now grounded in concrete technical progress.