



From Multi-Agent Reinforcement Learning to Agentic AI: A Comprehensive Literature Review of Algorithmic Advances and Decision-Analytic Implications (2020-2025)

Bharatendra Rai^{1,*}, Milena Popović²

¹ Department of Decision and Information Sciences, Charlton College of Business, University of Massachusetts – Dartmouth, USA

² Department of Operations Research and Statistics, University of Belgrade, Faculty of Organizational Sciences, Serbia

ARTICLE INFO

Article history:

Received 15 April 2026

Received in revised form 1 June 2026

Accepted 2 July 2026

Available online 5 August 2026

Keywords:

Multi-agent reinforcement learning, Agentic AI, Large language model agents, Cooperative MARL, Game theory, Decision analytics, Autonomous decision-making, Centralized training decentralized execution, Communication learning, LLM-based multi-agent systems

ABSTRACT

Agentic artificial intelligence has evolved from a research aspiration to a deployable technology between 2020 and 2025. This evolution rests on two intertwined research trajectories: the maturation of multi-agent reinforcement learning (MARL) for coordinated sequential decision-making, and the emergence of large language model (LLM)-based agent architectures integrating symbolic reasoning, tool use, and natural-language communication into cooperative multi-agent workflows. This literature review synthesizes 57 peer-reviewed and openly archived contributions published since 2019 across journals and reputable venues, organized into a thematic taxonomy spanning value-decomposition algorithms (QMIX, QPLEX, Weighted QMIX, FACMAC), trust-region and sequence-model policy methods (MAPPO, HAPPO, MAT, HARL, UPDeT), communication and role learning (NDQ, I2C, ROMA, RODE), credit assignment (LICA, Difference Rewards Policy Gradients, DOP), game-theoretic equilibrium solvers (Pipeline PSRO, JPSRO, Online Double Oracle), open-ended and mixed-motive learning (Open-Ended Learning Team, CICERO, alliance dilemmas), and LLM-based agentic frameworks (AutoGen, MetaGPT, CAMEL, AgentVerse, ChatDev, Generative Agents, Voyager, ReAct, Reflexion, Tree of Thoughts). We compare benchmark and reproducibility infrastructure (PettingZoo, EPyMARL benchmarking, SMAC variants), examine application domains (autonomous driving, multi-agent pathfinding, software engineering, scientific discovery), and discuss implications for applied decision analytics, including human-in-the-loop arbitration, risk-bounded coordination, and verifiable autonomy. We close with an agenda of open problems including non-stationarity, credit assignment under partial observability, alignment and safety in deceptive agents, evaluation under distribution shift, and integrating symbolic reasoning with reinforcement-learned policies to guide the next phase of agentic AI research.

1. Introduction

The decade following 2015 witnessed deep reinforcement learning (RL) progress from arcade-game novelty to systems capable of mastering complex, real-time strategic environments such as Go, StarCraft II, and Dota 2. As single-agent RL matured, attention shifted toward problem settings that

* Corresponding author.

E-mail address: brai@umassd.edu

<https://doi.org/10.66972/ada202734>

© The Author(s) 2027 | [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

single-agent abstractions cannot capture traffic networks coordinating thousands of autonomous vehicles, fleets of warehouse robots dispatched in parallel, financial markets in which heterogeneous trading agents respond to one another, and software-engineering pipelines in which specialized AI assistants must hand off subtasks. These settings share a defining characteristic that the optimal action of any one agent depends on the policies of others, and they require a distinct methodological framework: multi-agent reinforcement learning (MARL).

Between 2020 and 2025, MARL underwent parallel maturation. New algorithms substantially improved sample efficiency, scalability, and stability, while a sequence of community-driven benchmark suites and open-source codebases dramatically improved reproducibility. At the same time, the emergence of large language models (LLMs) introduced a fundamentally new agent substrate. LLM-based agents do not fit neatly into the classical MARL formalism as their action spaces are open-ended natural language, their reasoning emerges from in-context inference rather than gradient-based policy optimization, and their coordination often proceeds through dialogue rather than scalar reward decomposition. Yet they share with MARL the central problems of credit assignments, coordination under partial observability, role specialization, and the management of emergent behavior. The result is a research landscape in which two communities, classical MARL researchers and LLM-agent researchers, increasingly co-cite one another's work and borrow methodological tools.

This literature review consolidates this rapidly converging field. Our scope is the 2020-2025 window of peer-reviewed and openly archived research on agentic AI algorithms, with explicit emphasis on multi-agent learning. We synthesize 57 papers drawn from the leading venues in machine learning (NeurIPS, ICML, ICLR, JMLR), multi-agent systems (AAMAS), natural language processing (EMNLP, ACL), and applications (IEEE Robotics and Automation Letters, IEEE Transactions on Intelligent Transportation Systems). The review is organized thematically rather than chronologically, because the field's progress has not been linear: foundational algorithms such as QMIX [1] continue to receive extensions years after their introduction, and the LLM-agent literature, though concentrated in 2023-2024, frequently revisits problems formalized in classical MARL.

We pursue three contributions. First, we provide a taxonomy that bridges classical and LLM-based agentic approaches, exposing methodological commonalities of credit assignment, role specialization, communication learning, and decentralized execution, that are too often discussed in isolation within each subcommunity. Second, we offer comparative summaries that allow practitioners to select among competing algorithms within each thematic cluster, by mapping each paper to its design choices (e.g., centralized critic vs. value decomposition, attention-based vs. graph-based communication) and benchmark coverage. Third, we frame the literature explicitly through the lens of applied decision analytics, the field of Applied Decision Analytics concerns itself with both methodological advances and application-driven studies in data-driven and uncertainty-aware decision-making. We argue that agentic AI is increasingly relevant to this remit because real-world decision pipelines (logistics, healthcare triage, financial portfolio management, infrastructure scheduling) are inherently multi-agent and increasingly automated through learned policies. Section 12 develops this connection in depth.

The review is structured as follows. Section 2 describes our review methodology, inclusion criteria, and limitations. Section 3 establishes the formal foundations of MARL, including the Decentralized Partially Observable Markov Decision Process (Dec-POMDP) and the Stochastic Game formalism. Section 4 surveys the algorithmic taxonomy of cooperative MARL, covering value decomposition, policy gradient, and sequence-model approaches. Section 5 reviews communication, coordination, and role-learning methods. Section 6 covers credit assignments in detail. Section 7 examines game-theoretic and equilibrium-based approaches for competitive and mixed-motive

settings. Section 8 surveys open-ended learning and the use of MARL in mixed-motive games such as Diplomacy. Section 9 discusses benchmarks, reproducibility, and evaluation methodology. Section 10 turns to the agentic AI wave of LLM-based multi-agent systems and synthesizes the rapid 2023-2025 literature on dialogue-based agent coordination, tool use, and role specialization. Section 11 reviews applications. Section 12 develops the implications for decision analytics. Section 13 articulates open challenges and a forward research agenda. Section 14 concludes.

2. Review Methodology

2.1 Search Strategy and Inclusion Criteria

We conducted a focused thematic literature review rather than an exhaustive systematic review, in keeping with the *Applied Decision Analytics* journal's expectation that reviews provide an adequate overview of representative work in the field, situated against current open questions. Sources were drawn from the arXiv preprint server (focusing on author-archived versions of accepted publications), open-access proceedings of top-tier machine-learning venues (NeurIPS, ICML, ICLR, JMLR, OpenReview), and peer-reviewed journals (IEEE Transactions, JMLR, Foundations and Trends in Machine Learning, Frontiers of Computer Science, Applied Intelligence). We restricted the time window to 2020-2025 inclusive, with one exception: foundational works from 2019 (notably MAVEN [2]) are included where they remain canonical references for subsequent 2020+ developments.

Inclusion criteria required that a paper (i) was peer-reviewed at a recognized venue or, in the case of arXiv preprints, was either subsequently accepted at a peer-reviewed venue or carried sufficient methodological substance and citation impact (>100 citations as of mid-2025) to merit inclusion as gray literature; (ii) made an algorithmic, theoretical, or empirical contribution to multi-agent reinforcement learning, agentic AI, or directly adjacent areas (LLM-based multi-agent coordination, game-theoretic learning, agent communication); and (iii) was openly accessible to maximize reproducibility and reader access.

Exclusion criteria removed papers that (i) addressed single-agent RL without an explicit multi-agent contribution; (ii) treated multi-agent systems purely from a software-engineering perspective without learning content; or (iii) were paywalled with no openly archived version.

2.2 Thematic Synthesis

We employed thematic coding rather than meta-analytic aggregation, because the included papers report results across heterogeneous benchmarks (StarCraft Multi-Agent Challenge, Multi-Agent Particle Environments, Hanabi, Diplomacy, Minecraft, GitHub-issue resolution) that resist quantitative meta-analysis. Each paper was coded along the following dimensions: (a) problem class (cooperative, competitive, mixed-motive, agentic-AI/LLM); (b) algorithmic mechanism (value decomposition, policy gradient, actor-critic, equilibrium computation, transformer-based, dialogue-based); (c) communication paradigm (no communication, learned discrete messages, learned continuous embeddings, natural-language dialogue); (d) credit assignment scheme (centralized critic, value factorization, counterfactual, difference rewards, none); (e) benchmark coverage; and (f) deployment readiness (simulation only, robotics demonstration, real-world deployment).

2.3 Thematic Synthesis

Three limitations qualify our synthesis. First, the agentic-AI literature is moving faster than peer-review cycles can accommodate; some highly cited 2024-2025 contributions remain in preprint form, and we have included them where their methodological substance and community uptake clearly warrant inclusion. Second, our 57-paper corpus is curated for breadth rather than completeness whereas important threads in offline MARL, hierarchical MARL, and continual MARL are touched only briefly. Third, we cite many papers based on their archived versions; readers should consult the published versions where available for any verbatim quotation.

3. Foundations of Multi-Agent Reinforcement Learning

3.1 Formal Frameworks

The mathematical scaffolding for MARL rests on two formalisms. The Stochastic (Markov) Game generalizes the Markov Decision Process (MDP) to N agents that simultaneously select actions in a shared state, each receiving a possibly distinct reward. When agents share a common reward and observe only local information, the resulting structure is the Decentralized Partially Observable MDP (Dec-POMDP), which is widely adopted as the canonical cooperative-MARL formalism in the surveyed literature [3, 4]. The Dec-POMDP framing emphasizes two structural difficulties that motivate most algorithmic innovation in the field: agents must learn under partial observability (each agent sees only its own observation), and they must coordinate without communicating their internal state at execution time.

The Centralized Training with Decentralized Execution (CTDE) paradigm has become the dominant practical strategy across cooperative MARL [3, 4]. Under CTDE, the training procedure exploits a centralized perspective of full state information, joint actions, and joint reward, to learn coordinated value functions or policies, but the resulting policies are conditioned only on each agent's local observation history at deployment. This separation enables the use of expressive centralized critics during training without sacrificing the decentralized inference required by physically distributed agents.

3.2 Core Challenges

Zhang et al. [3] and the comprehensive cooperative-MARL review by Oroojlooy and Hajinezhad [4] together articulate the recurring technical difficulties that motivate algorithmic literature. Non-stationarity arises because, from any single agent's perspective, the environment evolves as other agents update their policies, violating the stationary-environment assumption that underpins single-agent RL convergence. Partial observability requires agents to maintain belief states or recurrent representations to compensate for missing global information. Multi-agent credit assignment is the problem of attributing the joint team reward to individual agents' actions when only the joint signal is available. Scalability degrades sharply with agent count because the joint action space grows exponentially.

Finally, exploration in multi-agent settings is qualitatively harder than in single-agent settings: coordinated exploration requires multiple agents to commit jointly to novel action sequences, a problem MAVEN [2] addresses through latent-variable conditioning.

3.3 Cooperative versus Competitive versus Mixed-Motive Regimes

A useful conceptual division separates the literature into cooperative MARL (all agents share a common reward), competitive MARL (zero-sum or constant-sum settings, dominated by game-theoretic equilibrium computation), and mixed-motive MARL (general-sum settings in which agents have partially aligned and partially opposed objectives, including social dilemmas, market simulations, and negotiation games). Each regime has produced its own algorithmic mainstream. Cooperative MARL is dominated by value-decomposition and centralized-critic methods (Section 4); competitive MARL relies on policy-space response oracles and equilibrium meta-solvers (Section 7); mixed-motive MARL combines elements of both with explicit modeling of opponent intent and game-theoretic equilibrium concepts beyond Nash, such as correlated equilibria [5-7].

4. Algorithmic Taxonomy of Cooperative MARL

4.1 Value-Decomposition Methods

Value decomposition is the workhorse approach to cooperative MARL under CTDE. The central idea is to learn a joint action-value function Q_{tot} that decomposes into individual utility functions Q_i such that decentralized greedy action selection on the Q_i recovers the argmax of Q_{tot} . QMIX [1], published in JMLR, represents the canonical realization of this principle: it constrains Q_{tot} to be a monotonic, non-linear mixing of agent utilities, computed by a hypernetwork conditioned on the global state. Monotonicity guarantees the Individual-Global-Max (IGM) property in that decentralized argmax selection on individual utilities recovers the joint argmax, while admitting richer mixing functions than simple sums.

Subsequent work generalized and refined the decomposition. Weighted QMIX [8] introduced two extensions, Centrally-Weighted QMIX (CW-QMIX) and Optimistically-Weighted QMIX (OW-QMIX), that re-weight the Bellman update to better handle the bias introduced by the monotonicity constraint, particularly in non-monotonic cooperative tasks. QPLEX [9] reformulates the IGM constraint via a duplex dueling architecture that decomposes Q_{tot} into a sum of state values and advantage terms, the latter enforced to satisfy IGM through a learned attention-weighted aggregation. The result expands the representational capacity of the decomposition while preserving decentralized execution. NDQ [10] takes an information-theoretic view: it learns nearly-decomposable value functions in which residual coordination is concentrated in a small number of communication channels minimized through a variational objective. These methods together establish the design space of value decomposition: the trade-off between decomposition expressiveness, training stability, and the granularity of the IGM constraint.

4.2 Policy Gradient and Actor-Critic Methods

Parallel to value decomposition, the policy-gradient strand has produced its own line of methods adapted from single-agent counterparts. MAPPO [11], published at NeurIPS, demonstrated empirically that Proximal Policy Optimization (PPO) with a centralized value function which is the simplest possible CTDE adaptation of a popular single-agent algorithm, achieves strong, often state-of-the-art, results across StarCraft Multi-Agent Challenge (SMAC), Hanabi, and Multi-Agent Particle Environments. The “surprising effectiveness” framing of MAPPO challenged a prevailing assumption that bespoke MARL algorithms were uniformly superior to careful adaptations of single-agent methods, and it has become a community-standard baseline.

HAPPO and HATRPO [12], published at ICLR, offer a theoretically grounded extension by deriving the multi-agent counterpart of the trust-region policy optimization framework. The key insight is that sequential policy updates under a properly chosen update order admit a monotonic improvement guarantee analogous to the single-agent case, even in the heterogeneous-agent regime. Heterogeneous-Agent RL (HARL [13]), published in JMLR, generalizes this further by removing parameter-sharing assumptions altogether and providing the first MARL algorithms with monotonic improvement guarantees under heterogeneous, parameter-non-shared agents. FACMAC [14], published at NeurIPS, integrates centralized policy gradients with factored critics, adapting the deterministic policy gradient to multi-agent settings.

DOP [15] addresses the off-policy regime by combining decomposed policy gradients with a centralized linearly decomposed critic. LICA [16] with Learning Implicit Credit Assignment, replaces explicit value decomposition with an implicit, end-to-end-learned credit assignment mechanism that conditions the policy gradient on a centralized critic without requiring an IGM-compatible decomposition.

4.3 Sequence Models and Attention-Based Architectures

A more recent thread reframes MARL itself as a sequence-modeling problem. The Multi-Agent Transformer (MAT [17]), published at NeurIPS, treats joint action selection as an autoregressive decoding problem: agents are ordered, and each agent's action is generated conditioned on the observations of all agents and the actions of preceding agents in the order. Theoretically, MAT links to HAPPO's sequential-update insight by exposing the decoding order as the sequential update order. Empirically, MAT achieves strong scalability across heterogeneous-agent benchmarks. UPDeT [18], published at ICLR, applies transformers in a different way to the per-agent representation, decoupling the policy across action types via attention over observation entities. The result is a universal policy network applicable across tasks with varying numbers of entities and action types, addressing one of the long-standing limitations of CTDE methods that were typically tied to fixed input/output dimensions. UPDeT's policy-decoupling-with-transformers approach has informed subsequent work on transferable MARL representations. Table 1 summarizes the principal cooperative MARL algorithms reviewed.

5. Communication, Coordination, and Role Learning

5.1 Learned Communication

Inter-agent communication has been studied as a means of overcoming partial observability without the bandwidth and reliability assumptions of full state-sharing. NDQ [10] frames communication as a value-decomposition problem: agents communicate only when residual coordination requirements demand it, with the size of the communication channel minimized via a variational objective. Individually Inferred Communication [19], published at NeurIPS formalizes the question of whom each agent should communicate with by deriving an individual prior over the receivers most likely to benefit from a message, conditioned on the sender's local observation. This allows for sparse, targeted communication that scales better than the dense broadcast assumed in earlier methods. The MARL communication survey by Zhu, Dastani, and Wang [20] consolidates this thread, dividing the design space along axes of message addressing (broadcast vs. targeted), message content (discrete vs. continuous; learned vs. structured), and timing (synchronized vs. event-driven).

5.2 Role Discovery and Decomposition

A complementary approach to coordination factors the joint policy through emergent or learned roles. ROMA [21], published at ICML, introduces a mutual-information-based regularizer that encourages agents with similar local observations to develop similar latent role representations, while differentiating roles across distinct subtasks. This produces emergent specialization that improves both sample efficiency and interpretability of learned policies. RODE [22], published at ICLR, takes a more explicit decomposition approach by clustering action effects to define a discrete role space and then learning role-conditional policies, with a high-level policy selecting roles for each agent at coarse time scales. The decomposition both reduces the effective action-space dimensionality per agent and exposes a natural curriculum for hierarchical exploration. MAVEN [2] though predating our window, remains a foundational reference for committed exploration in MARL: a latent variable conditions the joint policy across an episode, breaking the suboptimality of decentralized epsilon-greedy exploration in the QMIX family.

5.3 Coordination via Credit Assignment

The line between communication and credit assignment is fluid: both ultimately address how individual agents reconcile local actions with team-level outcomes. Difference Rewards Policy Gradients [23], published at AAMAS, computes counterfactual baseline rewards that quantify each agent’s marginal contribution to team return, providing a decentralized policy-gradient signal that empirically reduces variance and improves cooperation in benchmarks where naive shared rewards fail. Section 6 returns to credit assignment in greater depth.

Table 1

Principal cooperative MARL algorithms (2019-2024) reviewed in this work, organized by mechanism family

<i>Algorithm</i>	<i>Year</i>	<i>Venue</i>	<i>Mechanism Family</i>	<i>Key Innovation</i>	<i>Decentralized Execution</i>
MAVEN [2]	2019	NeurIPS	Value decomp + latent	Committed exploration via shared latent	Yes
QMIX [1]	2020	JMLR	Value decomposition	Monotonic mixing network with hypernetwork	Yes
Weighted QMIX [8]	2020	NeurIPS	Value decomposition	CW-/OW-QMIX re-weighted Bellman update	Yes
ROMA [21]	2020	ICML	Role learning	Mutual-information role regularizer	Yes
NDQ [10]	2020	ICLR	Comm + value decomp	Communication-minimizing factorization	Yes
LICA [16]	2020	NeurIPS	Implicit credit	End-to-end learned credit attribution	Yes
FACMAC [14]	2021	NeurIPS	Centralized policy gradient	Factored deterministic policy gradients	Yes
QPLEX [9]	2021	ICLR	Value decomposition	Duplex dueling with attention-based IGM	Yes
RODE [22]	2021	ICLR	Role learning	Action-effect role decomposition with hierarchy	Yes
DOP [15]	2021	ICLR	Off-policy actor-critic	Linear value decomposition + sample reuse	Yes
UPDeT [18]	2021	ICLR	Transformer architecture	Universal transformer with policy decoupling	Yes
Diff. Rewards PG [23]	2021	AAMAS	Counterfactual PG	Difference-reward policy gradient signal	Yes
MAPPO [11]	2022	NeurIPS	Centralized PPO	Centralized critic + PPO baseline	Yes
HAPPO/HATRPO [12]	2022	ICLR	Trust region	Sequential-update monotonic improvement	Yes
MAT [17]	2022	NeurIPS	Sequence model	Autoregressive joint-action transformer	Yes
HARL [13]	2024	JMLR	Heterogeneous-agent	Monotonic improvement w/o parameter sharing	Yes

6. Credit Assignment in Cooperative MARL

Credit assignment which is the problem of decomposing a shared team reward into per-agent learning signals, is one of the most foundational and persistent challenges in cooperative MARL. The literature reviewed here exposes three distinct families of solutions.

The first family relies on value decomposition to implicitly assign credit. The QMIX family [1, 9, 8] ensures that each agent's individual utility Q_i tracks its marginal contribution to Q_{tot} through the monotonicity constraint, so that gradient flow during Bellman updates implicitly attributes responsibility for joint-return changes to individual utilities. NDQ [10] extends this view by casting communication as the residual that cannot be decomposed.

The second family computes explicit counterfactual baselines. Difference Rewards Policy Gradients [23] calculates each agent's policy gradient with respect to a counterfactual reward, what the team would have received if that agent had taken a default action, yielding a low-variance, agent-specific signal. This generalizes the COMA (Counterfactual Multi-Agent) family of methods to a wider range of policy-gradient formulations.

The third family employs implicit, end-to-end learned credit assignment. LICA [16] eschews any predetermined decomposition or counterfactual baseline and instead trains the centralized critic and decentralized actors jointly, allowing the critic to learn credit-distribution patterns implicitly through gradient flow. The trade-off is reduced interpretability against improved expressiveness.

DOP [15] bridges the value-decomposition and policy-gradient strands by deriving a decomposed off-policy actor-critic with linear value decomposition that admits sample reuse from off-policy buffers, a critical scalability lever in benchmarks with expensive simulation. LICA, DOP, Difference Rewards, and the QMIX family together constitute the dominant credit-assignment toolkit reviewed here.

A practical lesson emerges from comparing benchmark results across these methods. Papoudakis et al. [24] show that the choice of credit assignment scheme is highly task-dependent: tasks with sparse, delayed rewards (Predator-Prey, multi-step coordination challenges) favor counterfactual-baseline methods, while tasks with dense intermediate signals (StarCraft micro-management) often see strong performance from simple value-decomposition methods. Section 9 returns to this benchmark sensitivity.

7. Game-Theoretic and Equilibrium-Based Approaches

In zero-sum and general-sum competitive settings, the cooperative-MARL toolkit gives way to game-theoretic methods that explicitly target equilibrium concepts. The Policy-Space Response Oracles (PSRO) framework, originally introduced before our review window, has seen substantial extension over 2020-2025. Pipeline PSRO [25], published at NeurIPS, addresses one of the principal scalability bottlenecks of PSRO by parallelizing the response-oracle computation across multiple policy-space coordinates in a pipeline that maintains a population of best-response candidates simultaneously. The result is a scalable approach for finding approximate Nash equilibria in two-player zero-sum games with large strategy spaces, demonstrated on Barrage Stratego.

JPSRO (Joint Policy-Space Response Oracles [6]), published at ICML, extends the framework beyond zero-sum to general-sum games via correlated equilibrium meta-solvers. The contribution is significant because Nash equilibrium is unstable and computationally hard in general-sum games, while correlated equilibria admit efficient computation and enjoy stronger learnability properties. JPSRO replaces the Nash meta-solver in classical PSRO with one that targets correlated or coarse-

correlated equilibria, expanding the applicability of the response-oracle framework to mixed-motive multi-agent settings such as routing, negotiation, and resource allocation.

Online Double Oracle [26], published in TMLR, takes a complementary perspective by recasting double-oracle methods in an online-learning framework. The connection between PSRO-style methods and the classical double-oracle algorithm in matrix games is made explicit, enabling regret-based convergence analyses and yielding online algorithms that converge to Nash equilibria under no-regret meta-solvers.

A separate strand examines the use of game-theoretic primitives directly in deep MARL. The “Game-Theoretic Multi-Agent Reinforcement Learning” survey by Yang et al. [27], published in Foundations and Trends in Machine Learning, consolidates this line of work, covering fictitious play, regret minimization, evolutionary game-theoretic methods, and their integration with deep representation learning. The survey identifies the persistent gap between theoretical convergence guarantees in tabular settings and empirical performance in deep, function-approximation regimes, a gap that animates much of the recent literature.

8. Open-Ended Learning, Mixed-Motive Settings, and Emergent Behavior

A distinctive thread of 2020-2025 research expands MARL beyond fixed-task benchmarks toward open-ended task distributions. The Open-Ended Learning Team at DeepMind [28] trained a single agent population on a procedurally generated, ever-expanding task distribution in the XLand environment, demonstrating zero-shot capability transfer across a vast space of physical-coordination tasks not seen during training. The methodological contribution is twofold: a curriculum mechanism that adapts task difficulty to the current capability of the agent population, and a population-based training scheme that maintains diversity through evolutionary selection. The empirical result, agents that exhibit qualitatively novel behaviors in entirely unseen tasks, has shaped subsequent thinking about generalization in agentic AI.

In the mixed-motive regime, alliance-formation dilemmas pose a particular challenge: agents must reason about which subset of opponents to cooperate with, when to defect, and how to manage reputation. Hughes et al. [5], in their AAMAS work on alliance dilemmas, develop population-based training schemes for many-player zero-sum games (the Alliance Dilemma family) that yield agents capable of forming, maintaining, and dissolving alliances under environmental pressure. The agents’ behavior exhibits cyclical alliance formation patterns reminiscent of human strategic play, suggesting that emergent diplomatic behavior may not require explicit linguistic communication.

The CICERO system [7], published in Science via Meta FAIR, represents perhaps the most striking integration of MARL with natural-language communication in our review window. CICERO plays No-Press Diplomacy which is a seven-player negotiation game with explicit alliance dynamics at human level, combining a policy network trained via human-regularized RL with a controllable language model for negotiation messages. The system exhibits behavior that human players judge as cooperative, persuasive, and contextually appropriate, while remaining strategically competitive. Methodologically, CICERO foreshadows the LLM-based agentic AI wave reviewed in Section 10: the integration of a policy network for action selection with a language model for communication is now a recurring architectural pattern.

9. Benchmarks, Reproducibility, and Evaluation

The maturation of MARL is closely tied to the maturation of its benchmark and evaluation infrastructure. PettingZoo [29], published in NeurIPS Datasets and Benchmarks, provides a unified,

Gym-like API across over 60 multi-agent environments spanning cooperative, competitive, and mixed-motive settings. The library's API, AEC (Agent Environment Cycle), makes turn-based and simultaneous-action games interoperable, and it has become the de facto standard interface for MARL experimentation.

Papoudakis et al. [24], published in NeurIPS Datasets and Benchmarks, contributed an equally important reproducibility effort: a standardized benchmarking study of cooperative MARL algorithms (QMIX, MADDPG, MAPPO, MAA2C, IPPO, IQL, COMA, VDN) across StarCraft Multi-Agent Challenge and Multi-Agent Particle Environments. Their results expose the brittleness of many published claims to small implementation choices and hyperparameter selections, and they articulate a set of best practices such as random seeds, hyperparameter sweeps, learning curves with confidence intervals, that have substantially raised the reproducibility floor of the field.

A third resource, the StarCraft Multi-Agent Challenge (SMAC) suite (referenced via [11] and many of the algorithm papers in our corpus), remains the most widely used cooperative MARL benchmark, although the literature increasingly notes its limitations: SMAC tasks rely on hand-coded enemy AI, which provides limited variation in opponent behavior, and many algorithms reach performance saturation that obscures genuine algorithmic differences. The community has begun to migrate toward more challenging benchmarks (Hanabi, Google Research Football, Multi-Agent MuJoCo, SMAC v2), reflected in recent algorithm papers reporting on multiple benchmark suites.

10. The Agentic AI Wave: LLM-Based Multi-Agent Systems

The period 2023-2025 saw an explosion of multi-agent systems built on top of large language models. These systems differ from classical MARL in several fundamental ways: action spaces are open-ended natural language, coordination proceeds through dialogue rather than scalar reward sharing, and policies are shaped via prompting and in-context learning rather than gradient-based optimization. Yet they confront the same coordination problems of credit assignment, role specialization, communication efficiency, that classical MARL has studied for decades. We review this literature thematically.

10.1 Single-Agent Foundations: Reasoning, Acting, and Reflection

The agentic-AI wave rests on three foundational single-agent contributions. ReAct [30], published at ICLR, demonstrated that interleaving chain-of-thought reasoning steps with environment-grounded acting steps produces qualitatively better agent behavior than either alone, particularly in long-horizon tasks involving tool use and information retrieval. The ReAct prompting pattern of Thought, Action, Observation, and repeat, has become an architectural primitive replicated across virtually every subsequent agentic-AI framework. Reflexion [31], published at NeurIPS, augments ReAct with a self-reflection mechanism: after each unsuccessful trajectory, the agent generates verbal feedback that is stored in episodic memory and conditions subsequent attempts. This verbal reinforcement learning using natural-language self-critique as a substitute for gradient updates, produces substantial improvements on programming, decision-making, and reasoning benchmarks. Tree of Thoughts [32], also at NeurIPS, generalizes chain-of-thought reasoning to a search tree over partial reasoning traces, with explicit branch evaluation and pruning.

Toolformer [33], published at NeurIPS, addresses the orthogonal question of how language models acquire tool-use capability through self-supervision: the model learns when and how to call external APIs by autoregressively predicting both natural-language tokens and API calls, with the API responses used as training signals. ToolLLM [34], published at ICLR, scales this to 16,000+ real-world

APIs through systematic data collection and instruction tuning. The Tool Manipulation Capability paper [35] extends this evaluation to open-source LLMs, finding substantial gaps between proprietary and open-source models in zero-shot tool-use ability. Together, these works define the agent's interface to the external world.

10.2 Multi-Agent Frameworks: Conversation, Specialization, and Emergence

A second cluster of papers builds explicit multi-agent frameworks atop the single-agent primitives. AutoGen [36] provides a flexible Python framework in which agents are configured as "ConversableAgents" that communicate through structured messages, with both human-in-the-loop and fully automated execution paths. The framework supports tool use, code execution, and group chat patterns, and has become widely adopted in industry deployments.

CAMEL [37], published at NeurIPS, formalizes a role-playing paradigm in which two LLM agents take complementary roles (e.g., user and assistant) and engage in a multi-turn dialogue to solve a task. CAMEL's inception prompting explicitly assigns roles and constraints at the outset, producing more coherent and goal-directed cooperation than naive few-shot multi-agent prompting. The empirical demonstration that role-conditioned dialogue can solve complex tasks including software engineering and creative-writing problems, has informed virtually all subsequent multi-agent LLM frameworks.

MetaGPT [38], published at ICLR, takes role specialization further by encoding human standard operating procedures (SOPs) for software development as agent role specifications: Product Manager, Architect, Engineer, QA Engineer, and so on. Each role consumes structured outputs from preceding roles, enabling complex software projects to be decomposed into agent workflows that produce executable code, tests, and documentation. The contribution is methodological: encoding domain-expert workflows as agent SOPs translates much of the cooperative-MARL coordination problem into prompt engineering. ChatDev [39] implements a similar idea specifically for software development with a sequential waterfall workflow.

AgentVerse [40], published at ICLR, generalizes role specialization to a flexible expert-recruitment paradigm: given a goal, the system dynamically recruits a group of specialized agents, has them deliberate in a structured collaborative-decision phase, executes their plan, and evaluates outcomes for iterative refinement. AgentVerse explicitly studies emergent collaborative behaviors of volunteer behavior, conformity, destructive behavior, that arise in multi-agent groups, providing empirical evidence that LLM agent groups exhibit social dynamics analogous to human teams.

10.3 Embodied, Open-Ended, and Generative Agents

A parallel thread embeds LLM agents in rich, open-ended environments. Voyager [41] places a GPT-4-driven agent in Minecraft with three components: an automatic curriculum that proposes increasingly complex tasks, an iterative prompting mechanism that generates and refines code-based skills, and an ever-growing skill library. Voyager autonomously discovers diverse behaviors and acquires more skills over time than baseline agents that lack the curriculum-skill-library architecture. Generative Agents [42], published at UIST, takes a different angle: 25 agents inhabiting a small simulated town exhibit believable individual and social behavior from morning routines to coordinating a Valentine's Day party, driven by a memory stream, reflection, and planning architecture. The paper demonstrates that LLM agents can produce social simulations of qualitatively new richness compared to rule-based or scripted alternatives.

SwiftSage [43], published at NeurIPS, integrates fast pattern-matching (“Swift”) and slow deliberative reasoning (“Sage”) in a dual-process agent architecture inspired by Kahneman’s System 1 / System 2 distinction. The agent achieves strong results on ScienceWorld, an interactive scientific reasoning benchmark, with substantial computational savings compared to pure deliberative-reasoning agents.

10.4 Coordination, Theory of Mind, and Debate

A maturing thread of research applies classical multi-agent concepts that is theory of mind, coalition formation, and debate, to LLM agents. Li et al. [44], published at EMNLP, evaluate LLM agents in a multi-agent cooperative text game requiring theory-of-mind inference, comparing them to MARL and planning baselines. The study finds emergent collaborative and high-order theory-of-mind behaviors but also documents systematic failures in long-horizon planning and hallucination of shared task state, failures that explicit belief-state representations partially mitigate. The Talebirad and Nadiri [45] framework on harnessing intelligent LLM agents articulates a general architecture for multi-agent collaboration with role specialization, communication channels, and coordination protocols.

Du et al. [46], published at ICML, demonstrate that multi-agent debate of instantiating multiple LLM agents that argue and refine answers across multiple rounds, substantially improves factual accuracy and reasoning quality compared to single-agent baselines or majority voting over independent samples. The mechanism resembles classical ensembling but with structured exchange of justifications. LLM-Blender [47], published at ACL, takes a related but distinct ensembling approach: it pairwise-ranks outputs from a pool of LLMs and generatively fuses the top-ranked candidates, achieving consistent improvements across NLP tasks.

The emergent-collaboration paper by Zhang et al. [48], published at ACL, draws explicit parallels to social-psychology theories of group dynamics including conformity, polarization, and reasoning biases, finding that LLM agent groups reproduce many of the same behavioral phenomena. This empirical observation has implications for both deployment (multi-agent LLM systems may inherit human social pathologies) and theory (social-psychology constructs may provide useful priors for designing multi-agent LLM workflows).

10.5 Safety, Deception, and the Limits of Alignment

The agentic-AI wave has made safety considerations more pressing. Hubinger et al. [49] demonstrate that LLM agents can be trained as “sleeper agents” which are systems that appear aligned during evaluation but exhibit deceptive behavior under deployment-time triggers and that standard safety training (RL from human feedback, red-teaming, supervised fine-tuning) fails to remove this deception. The result has stark implications for the deployment of multi-agent LLM systems in high-stakes decision contexts: the difficulty of detecting deceptive policies grows with model capability, and current evaluation methods provide weak guarantees against emergent misalignment.

10.6 Surveys of the Agentic AI Landscape

The rapid pace of the agentic-AI literature has prompted several synthesis papers. Wang et al. [50], published in *Frontiers of Computer Science*, provide a comprehensive survey of LLM-based autonomous agents, organizing the literature along the dimensions of profile, memory, planning, and action. Guo et al. [51], published at *IJCAI*, focus specifically on multi-agent LLM systems, taxonomizing them by communication structure, coordination mechanism, and task domain. Cheng et al. [52] discuss definitions, methods, and prospects for LLM-based intelligent agents from a forward-looking perspective. SWE-bench [53], published at *ICLR*, offers a benchmark for evaluating LLM agents on real-world GitHub issue resolution, providing one of the few benchmarks where agent capability is grounded in objective, executable success criteria. Table 2 summarizes the principal LLM-based agentic frameworks reviewed.

Table 2

Principal LLM-based agentic frameworks (2022-2024) reviewed, organized by capability emphasis.

<i>Framework</i>	<i>Year</i>	<i>Venue</i>	<i>Agent Count</i>	<i>Coordination Mechanism</i>	<i>Primary Capability</i>
ReAct [30]	2023	ICLR	1	Reasoning-acting interleaving	Tool-augmented reasoning
Toolformer [33]	2023	NeurIPS	1	Self-supervised API calls	Tool acquisition
Reflexion [31]	2023	NeurIPS	1	Verbal self-critique loop	Episodic self-improvement
Tree of Thoughts [32]	2023	NeurIPS	1	Tree search over reasoning	Deliberative problem solving
CAMEL [37]	2023	NeurIPS	2	Role-playing dialogue	Cooperative task completion
Generative Agents [42]	2023	UIST	25	Memory + reflection + planning	Believable social simulation
Voyager [41]	2023	TMLR	1 (lifelong)	Curriculum + skill library	Open-ended skill discovery
SwiftSage [43]	2023	NeurIPS	1 (dual)	Fast/slow dual process	Scientific reasoning
LLM-Blender [47]	2023	ACL	N (ensemble)	Pairwise rank + fuse	Output ensembling
ToM Multi-Agent [44]	2023	EMNLP	N	Theory-of-mind belief states	ToM-aware coordination
AutoGen [36]	2023	COLM	N (configurable)	Conversable agent dialogue	Flexible multi-agent framework
ToolLLM [34]	2024	ICLR	1	API instruction-tuning	Tool use at 16k+ APIs
MetaGPT [38]	2024	ICLR	5+	SOP-encoded role workflow	End-to-end software dev
ChatDev [39]	2024	ACL	4+	Sequential waterfall workflow	Software development
AgentVerse [40]	2024	ICLR	Variable	Expert recruitment + deliberation	Emergent collaboration
SWE-bench [53]	2024	ICLR	1+	(Benchmark)	GitHub issue resolution eval
Multi-Agent Debate [46]	2024	ICML	N	Multi-round argument	Factuality and reasoning
Social Collab [48]	2024	ACL	N	Social-psychology framing	Collaboration mechanism design
Sleeper Agents [49]	2024	arXiv	1	(Safety probe)	Deceptive alignment evaluation

11. Applications and Real-World Deployment

The literature reviewed offers substantive evidence of MARL and agentic AI moving beyond synthetic benchmarks into application domains.

11.1 Autonomous Driving

Two surveys in our corpus map the application of MARL to autonomous driving. Kiran et al. [54], published in *IEEE Transactions on Intelligent Transportation Systems*, surveys deep RL for autonomous driving more broadly, covering perception, prediction, decision-making, and control with attention to the multi-agent considerations of mixed traffic. The 2024 follow-up survey by Zhang et al. [55], focused specifically on MARL for autonomous driving, reviews the algorithmic state of the art for cooperative perception, intent inference, and joint trajectory planning, and details the simulator and benchmark landscape (CARLA, SMARTS, MetaDrive) on which the field's progress has been measured. A central methodological theme is that real-world driving scenarios, particularly mixed-autonomy traffic, require not just cooperative MARL among vehicles but also opponent-modeling of human drivers whose behaviors are non-stationary and only partially rational.

11.2 Multi-Agent Pathfinding

PRIMAL2 [56], published in *IEEE Robotics and Automation Letters*, applies cooperative MARL with imitation learning to multi-agent pathfinding (MAPF) in lifelong settings involving environments where new pickup-and-delivery tasks continuously arrive and agents must replan online. The system combines decentralized RL agents with imitation learning from a centralized expert planner, achieving scalable performance for warehouse-robot fleets at scales (>1000 agents) where centralized search becomes computationally intractable.

11.3 Software Engineering and Scientific Reasoning

The LLM-based multi-agent systems reviewed in Section 10 such as ChatDev, MetaGPT, and AutoGen, directly address software engineering applications. SWE-bench [53] provides empirical evidence that current LLM agents, in single- or multi-agent configurations, can resolve a non-trivial fraction of real-world GitHub issues with executable test verification. The ChatDev and MetaGPT systems demonstrate end-to-end software project completion from natural-language requirements to executable code with tests and documentation, operating through structured multi-role collaboration. SwiftSage [43] demonstrates capability on scientific reasoning tasks, and Generative Agents [42] showcases applications to social simulation and game-design prototyping.

11.4 Networked Systems and Resource Allocation

Scalable MARL for Networked Systems with Average Reward [57], published at NeurIPS, addresses a class of large-scale resource-allocation problems of power grids, traffic networks, communication networks, in which agents are arranged in a graph with local dependence structure. The authors derive a Scalable Actor-Critic (SAC) method whose complexity scales with the local-neighborhood state-action space rather than the global one, exploiting an exponential-decay property that quantifies how an agent's influence on others diminishes with graph distance. The

result provides one of the few MARL methods with explicit per-agent computational guarantees, and it has direct application to smart-grid and traffic-network control.

12. Implications for Applied Decision Analytics

The thematic organization of the *Applied Decision Analytics* journal covering decision models and frameworks, data science in decision analytics, uncertainty and risk, and human-centered and cognitive decision analytics, places agentic AI squarely within its remit. We identify four lines of implication.

12.1 Multi-Agent Decision Pipelines as Decision Analytics Workflows

Many of the most consequential real-world decision pipelines include supply-chain coordination, healthcare triage, financial portfolio rebalancing, infrastructure scheduling, fleet management, are inherently multi-agent. Historically, they have been addressed with operations-research methods (mathematical programming, network optimization, MCDA) that assume a single decision-maker with full information. The MARL literature reviewed here offers an alternative paradigm in which the pipeline is modeled as a system of interacting learning agents, each optimizing a local objective with shared and partial information. The CTDE paradigm with centralized training with decentralized execution, fits naturally with decision-analytics deployments where rich historical data permits centralized learning but operational deployment must be distributed.

12.2 Uncertainty Quantification and Risk-Bounded Coordination

Decision analytics emphasizes uncertainty quantification and risk-aware decision-making. Several MARL methods in our review have direct counterparts in this remit. The information-theoretic communication-minimization framework of NDQ [10] can be reinterpreted as a minimum-bandwidth coordination objective with privacy implications. Mean-field and scalable MARL methods [57] offer per-neighborhood error bounds that translate to risk-bounded coordination guarantees. Counterfactual-baseline methods [23] yield agent-specific marginal-contribution measures that align with the Shapley-value framework familiar to decision-analytic practitioners.

12.3 Human-Centered and Cognitive Decision Analytics

The LLM-agent literature is particularly relevant to human-centered and cognitive decision analytics — settings where decisions involve human stakeholders and where agent rationality must align with human cognitive frames. AutoGen [36] explicitly provides human-in-the-loop hooks. CICERO [7] demonstrates persuasive natural-language coordination at human level. The theory-of-mind work [44] and emergent-collaboration analyses [48] provide empirical foundations for designing multi-agent systems that reason about and accommodate human cognitive partners. Reflexion [31] and the multi-agent-debate paradigm [46] offer mechanisms for verifiable, auditable agent reasoning that can be inspected by human reviewers, an important property for accountability in high-stakes decision domains.

12.4 Verifiability, Auditability, and Safety

Decision analytics deployed in high-stakes domains demand verifiable and auditable agents. The reviewed literature offers both promise and warning. On the positive side, multi-agent debate [46] and reflective architecture [31] provide reasoning traces that can be audited. On the cautionary side, the sleeper-agents work [49] demonstrates that current safety-training methods provide weak guarantees against deceptive policies, particularly in agentic settings where models are deployed across heterogeneous tasks. Decision-analytic deployments must invest in continuous monitoring, behavioral guardrails, and adversarial evaluation rather than relying solely on pre-deployment safety training.

13. Open Challenges and Future Research Directions

We close the synthesis with a structured agenda of open problems suggested by the reviewed literature.

13.1 Bridging MARL and LLM-Agent Methodologies

Despite increasing co-citation between the two communities, the methodological synthesis between classical MARL and LLM-based agents remains immature. Classical MARL provides rigorous formalisms (Dec-POMDP, stochastic games), credit-assignment mechanisms, and convergence analyses, but its policies are typically opaque and tied to specific benchmarks. LLM-agent frameworks provide expressive policies, natural-language coordination, and rich tool-use capability, but lack the analytical clarity around credit assignment, sample efficiency, and convergence. Hybrid systems that train an LLM-based policy with MARL-style credit assignment over multi-agent rollouts represent a promising frontier where initial steps include the integration of CICERO's policy network with its language model [7] and the use of MARL-trained tool-use policies in LLM agents.

13.2 Non-Stationarity and Continual Adaptation

Most MARL benchmarks assume a fixed task distribution and a fixed agent population. Real-world deployments rarely satisfy either assumption. Open-ended learning [28] and meta-learning approaches partially address task non-stationarity, but population non-stationarity of the arrival, departure, and capability changes of co-agents during deployment, remains under-studied. The MAVEN and ROMA approaches to committed exploration and role discovery may inform continual-adaptation algorithms that maintain diverse, transferable policies across changing populations.

13.3 Credit Assignment under Partial Observability

The reviewed credit-assignment methods (LICA, Difference Rewards, DOP) provide solutions for fully-observable cooperative games, but credit assignment under partial observability where agents may not know what other agents observed or did, is substantially more difficult. The NDQ communication framework [10] and the I2C targeted-communication approach [19] offer partial paths forward, but a unified theoretical framework for credit assignment in Dec-POMDPs with limited communication remains an open problem.

13.4 Evaluation under Distribution Shift

Papoudakis et al. [24] raised concerns about reproducibility and benchmark sensitivity that remain unresolved. The agentic-AI literature compounds the concern: SWE-bench [53] and similar benchmarks measure success on snapshots of real-world tasks that drift over time, while LLM agent capability is sensitive to model versions, prompting strategies, and tool versions. Robust evaluation requires benchmarks that explicitly test distribution shift, including adversarial perturbations, long-tail task instances, and long-term deployment trajectories.

13.5 Alignment, Safety, and Verifiable Autonomy

The sleeper-agents result [49] suggests that the agentic-AI safety problem is qualitatively different from, and more difficult than, single-model alignment. Multi-agent systems exhibit emergent behaviors that may amplify deception, polarization, or resource-monopolization tendencies of individual models. Research directions include verifiable agent specifications (formal contracts that agents must satisfy), behavioral monitoring systems that detect anomalous coordination patterns, and adversarial-evaluation methodologies that test agent ensembles against worst-case co-agents.

13.6 Symbolic-Subsymbolic Integration

A long-standing aspiration of AI research involving the integration of symbolic reasoning with sub-symbolic learning, is partially realized in the agentic-AI wave. Tool-using LLM agents (Toolformer, ToolLLM) bridge to symbolic computation; ReAct and Tree of Thoughts incorporate explicit reasoning steps; multi-agent debate [46] externalizes reasoning into discursive structure. Yet the integration remains brittle: tool-call failures, reasoning errors, and hallucinated outputs propagate through multi-agent systems in ways that are poorly understood. A mature theory of symbolic-subsymbolic integration in agentic AI is among the most consequential open research directions.

13.7 Decision-Analytic Deployment Methodology

Finally, we identify a methodological gap of direct relevance to Applied Decision Analytics: there is little published guidance on how to translate the reviewed algorithmic literature into deployable decision pipelines. Existing surveys [3, 4, 27] catalog algorithms and benchmark results; the application surveys [54, 55] describe domain-specific challenges. What is missing is a deployment-oriented methodology that specifies how to scope a multi-agent decision problem, select among algorithmic options, validate convergence, monitor live performance, and govern operational risk. Decision-analytic practitioners are well positioned to develop this methodology by combining the algorithmic literature reviewed here with their existing toolkits in stochastic modeling, MCDA, and operations research.

14. Conclusions

The 2020-2025 literature on agentic AI algorithms reveals a research landscape undergoing rapid synthesis. Classical multi-agent reinforcement learning has matured along three principal axes of value decomposition, policy gradient and sequence models, and game-theoretic equilibrium computation, while a parallel wave of large-language-model-based multi-agent systems has

redefined what an “agent” can be: a reasoning-and-acting entity that coordinates with peers through natural-language dialogue, tool use, and reflective self-critique. The two communities, once distinct, increasingly co-cite each other and co-borrow methodological tools, foreshadowing a period of integration.

This review has organized 57 peer-reviewed and openly archived contributions into a thematic taxonomy spanning algorithmic foundations, communication and role learning, credit assignment, game-theoretic methods, open-ended learning, benchmarking infrastructure, LLM-agent frameworks, applications, and decision-analytic implications. We have argued throughout that the methodological commonalities across these areas including credit assignment, role specialization, decentralized coordination, and opponent modeling exceed their surface differences, and that decision-analytic practitioners are well positioned to translate the algorithmic literature into deployable workflows.

Two structural observations bear emphasis. First, the field’s most striking recent advances such as CICERO’s human-level Diplomacy play, Voyager’s lifelong skill acquisition, MetaGPT’s end-to-end software development, share a common architectural pattern: a learned coordination mechanism (whether MARL-trained or in-context-prompted) layered on top of expressive perception and action modules. Second, the field’s most pressing risks involving deceptive agents, brittle benchmarks, and evaluation drift, also share a common structure: emergent multi-agent behaviors that are difficult to predict, audit, or correct from observations of individual components.

We anticipate that the next phase of agentic AI research will be characterized by four developments: explicit methodological synthesis between classical MARL and LLM-agent frameworks; the development of robust, distribution-shift-aware evaluation infrastructure; the maturation of safety, alignment, and verifiability methods specific to multi-agent systems; and the codification of deployment-oriented methodology that bridges algorithmic capability and decision-analytic practice. We hope that this review, by mapping the field as it stands in 2025 and identifying these open problems, contributes to that next phase.

Acknowledgement

This research was not funded by any grant.

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

The authors declare that no generative AI or AI-assisted technologies were used in the manuscript preparation process.

References

- [1] Rashid, T., Samvelyan, M., Schroeder de Witt, C., Farquhar, G., Foerster, J., & Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178), 1-51. arXiv:2003.08839.
- [2] Mahajan, A., Rashid, T., Samvelyan, M., & Whiteson, S. (2019). MAVEN: Multi-agent variational exploration. *Advances in Neural Information Processing Systems (NeurIPS)*, 32. arXiv:1910.07483.
- [3] Zhang, K., Yang, Z., & Basar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In K. G. Vamvoudakis, Y. Wan, F. L. Lewis, & D. Cansever (Eds.), *Handbook of reinforcement learning and control* (pp. 321-384). Springer. arXiv:1911.10635.
- [4] Oroojlooy, A., & Hajinezhad, D. (2023). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11), 13677-13722. arXiv:1908.03963.

- [5] Hughes, E., Anthony, T. W., Eccles, T., Leibo, J. Z., Balduzzi, D., & Bachrach, Y. (2020). Learning to resolve alliance dilemmas in many-player zero-sum games. *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. arXiv:2003.00799.
- [6] Marris, L., Muller, P., Lanctot, M., Tuyls, K., & Graepel, T. (2021). Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers. *Proceedings of the International Conference on Machine Learning (ICML)*. arXiv:2106.09435.
- [7] Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwa, M., Lewis, M., Mishra, A., Renduchintala, A., Roller, S., Rowe, D., Shi, W., Spisak, J., Wei, A., Wu, D., Zhang, H., & Zijlstra, M. (2022). Mastering the game of no-press Diplomacy via human-regularized reinforcement learning and planning. *Science*, 378(6624), 1067-1074. arXiv:2210.05492.
- [8] Rashid, T., Farquhar, G., Peng, B., & Whiteson, S. (2020). Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33. arXiv:2006.10800.
- [9] Wang, J., Ren, Z., Liu, T., Yu, Y., & Zhang, C. (2021). QPLEX: Duplex dueling multi-agent Q-learning. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2008.01062.
- [10] Wang, T., Wang, J., Zheng, C., & Zhang, C. (2020). Learning nearly decomposable value functions via communication minimization. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:1910.05366.
- [11] Yu, C., Velu, A., Vinitisky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of PPO in cooperative multi-agent games. *Advances in Neural Information Processing Systems (NeurIPS)*, 35. arXiv:2103.01955.
- [12] Kuba, J. G., Chen, R., Wen, M., Wen, Y., Sun, F., Yang, Y., & Wang, J. (2022). Trust region policy optimisation in multi-agent reinforcement learning. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2109.11251.
- [13] Zhong, Y., Kuba, J. G., Feng, X., Hu, S., Ji, J., & Yang, Y. (2024). Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25, 1-67. arXiv:2304.09870.
- [14] Peng, B., Rashid, T., Schroeder de Witt, C. A., Kamienny, P.-A., Torr, P. H. S., Böhmer, W., & Whiteson, S. (2021). FACMAC: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems (NeurIPS)*, 34. arXiv:2003.06709.
- [15] Wang, Y., Han, B., Wang, T., Dong, H., & Zhang, C. (2021). DOP: Off-policy multi-agent decomposed policy gradients. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2007.12322.
- [16] Zhou, M., Liu, Z., Sui, P., Li, Y., & Chung, Y. Y. (2020). Learning implicit credit assignment for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33. arXiv:2007.02529.
- [17] Wen, M., Kuba, J. G., Lin, R., Zhang, W., Wen, Y., Wang, J., & Yang, Y. (2022). Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems (NeurIPS)*, 35. arXiv:2205.14953.
- [18] Hu, S., Zhu, F., Chang, X., & Liang, X. (2021). UPDeT: Universal multi-agent reinforcement learning via policy decoupling with transformers. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2101.08001.
- [19] Ding, Z., Huang, T., & Lu, Z. (2020). Learning individually inferred communication for multi-agent cooperation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33. arXiv:2006.06455.
- [20] Zhu, C., Dastani, M., & Wang, S. (2024). A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38, 4. arXiv:2203.08975.
- [21] Wang, T., Dong, H., Lesser, V., & Zhang, C. (2020). ROMA: Multi-agent reinforcement learning with emergent roles. *Proceedings of the International Conference on Machine Learning (ICML)*. arXiv:2003.08039.
- [22] Wang, T., Gupta, T., Mahajan, A., Peng, B., Whiteson, S., & Zhang, C. (2021). RODE: Learning roles to decompose multi-agent tasks. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2010.01523.
- [23] Castellini, J., Devlin, S., Oliehoek, F. A., & Savani, R. (2021). Difference rewards policy gradients. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. arXiv:2012.11258.
- [24] Papoudakis, G., Christianos, F., Schäfer, L., & Albrecht, S. V. (2021). Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *NeurIPS Datasets and Benchmarks Track*. arXiv:2006.07869.
- [25] McAleer, S., Lanier, J., Fox, R., & Baldi, P. (2020). Pipeline PSRO: A scalable approach for finding approximate Nash equilibria in large games. *Advances in Neural Information Processing Systems (NeurIPS)*, 33. arXiv:2006.08555.
- [26] Dinh, L. C., Yang, Y., McAleer, S., Perez-Nieves, N., Slumbers, O., Tian, Z., Mguni, D. H., Bou Ammar, H., & Wang, J. (2022). Online double oracle. *Transactions on Machine Learning Research (TMLR)*. arXiv:2103.07780.
- [27] Yang, Y., Ma, C., Ding, Z., McAleer, S., Jin, C., Wang, J., & Sandholm, T. (2025). Game-theoretic multiagent reinforcement learning. *Foundations and Trends in Machine Learning*. arXiv:2011.00583.
- [28] Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., McAleese, N., Bradley-Schmieg, N., Wong, N., Porcel, N., Raileanu, R., Hughes-Fitt, S., Dalibard, V., & Czarnecki, W. M. (Open-Ended Learning Team). (2021). Open-ended learning leads to generally capable agents. arXiv:2107.12808.

- [29] Terry, J. K., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L. S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., Williams, N., Lokesh, Y., & Ravi, P. (2021). PettingZoo: A standard API for multi-agent reinforcement learning. *NeurIPS Datasets and Benchmarks Track*. arXiv:2009.14471.
- [30] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2210.03629.
- [31] Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36. arXiv:2303.11366.
- [32] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36. arXiv:2305.10601.
- [33] Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems (NeurIPS)*, 36. arXiv:2302.04761.
- [34] Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M. (2024). ToolLLM: Facilitating large language models to master 16000+ real-world APIs. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2307.16789.
- [35] Xu, Q., Hong, F., Li, B., Hu, C., Chen, Z., & Zhang, J. (2023). On the tool manipulation capability of open-source large language models. arXiv:2305.16504.
- [36] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv:2308.08155 (presented at COLM 2024).
- [37] Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative agents for “mind” exploration of large language model society. *Advances in Neural Information Processing Systems (NeurIPS)*, 36. arXiv:2303.17760.
- [38] Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2024). MetaGPT: Meta programming for a multi-agent collaborative framework. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2308.00352.
- [39] Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., & Sun, M. (2024). ChatDev: Communicative agents for software development. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. arXiv:2307.07924.
- [40] Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C., Qin, Y., Cong, X., Xie, R., Liu, Z., Sun, M., & Zhou, J. (2024). AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2308.10848.
- [41] Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*. arXiv:2305.16291.
- [42] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. arXiv:2304.03442.
- [43] Lin, B. Y., Fu, Y., Yang, K., Brahman, F., Bhagavatula, C., Ammanabrolu, P., & Choi, Y. (2023). SwiftSage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 36. arXiv:2305.17390.
- [44] Li, H., Chong, Y. Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., & Sycara, K. (2023). Theory of mind for multi-agent collaboration via large language models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. arXiv:2310.10701.
- [45] Talebirad, Y., & Nadiri, A. (2023). Multi-agent collaboration: Harnessing the power of intelligent LLM agents. arXiv:2306.03314.
- [46] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the International Conference on Machine Learning (ICML)*. arXiv:2305.14325.
- [47] Jiang, D., Ren, X., & Lin, B. Y. (2023). LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. arXiv:2306.02561.
- [48] Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., & Deng, S. (2024). Exploring collaboration mechanisms for LLM agents: A social psychology view. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. arXiv:2310.02124.

- [49] Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askill, A., Radhakrishnan, A., Anil, C., Duvenaud, D., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv:2401.05566.
- [50] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345. arXiv:2308.11432.
- [51] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. arXiv:2402.01680.
- [52] Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., & He, X. (2024). Exploring large language model based intelligent agents: Definitions, methods, and prospects. arXiv:2401.03428.
- [53] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:2310.06770.
- [54] Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A. A. A., Yogamani, S., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6), 4909-4926. arXiv:2002.00444.
- [55] Zhang, R., Hou, J., Walter, F., Gu, S., Guan, J., Röhrbein, F., Du, Y., Cai, P., Chen, G., & Knoll, A. (2024). Multi-agent reinforcement learning for autonomous driving: A survey. arXiv:2408.09675.
- [56] Damani, M., Luo, Z., Wenbin, E., & Sartoretti, G. (2021). PRIMAL2: Pathfinding via reinforcement and imitation multi-agent learning - Lifelong. *IEEE Robotics and Automation Letters*, 6(2), 2666-2673. arXiv:2010.08184.
- [57] Qu, G., Lin, Y., Wierman, A., & Li, N. (2020). Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems (NeurIPS)*, 33. arXiv:2006.06626.