Autonomous AI systems make numerous daily decisions impacting human welfare. While this scalability increases capabilities, it introduces systemic hazards if AI behaviours fail to align with human safety and societal values. As AI agents and multi-agent systems gain autonomy, new risks emerge from complex interactions among agents, human operators, and dynamic environments [1, 2]. Rapid evolution of multimodal foundation models, such as Vision-Language Models (VLMs), Large Action Models (LAMs), has driven research toward expanding capabilities, enforcing safety alignment, and reducing computational resource needs. A critical frontier is, therefore, finding a sweet spot that balances algorithmic improvement with resource efficiency. Addressing these challenges requires rigorous evaluation of the agent capability lifecycle, spanning task-agnostic perception, autonomous policy optimisation, and multi-agent coordination. This research program categorises these capabilities and challenges across three core topics:
(1) **Task-Agnostic Spatial Reasoning:** Investigating how autonomous agents construct, maintain, and update visual-spatial representations and memory without explicit goal specifications.
(2) **Autonomous Self-Improvement & Red-Teaming:** Analysing iterative self-optimisation mechanisms, such as self-play, and characterising their theoretical and computational failure modes.
(3) **Multi-Agent Coordination & Safety Alignment:** Evaluating multi-agent coordination dynamics and decision convergence to distinguish genuine cooperative strategy from shared architectural bias.

Topic 1: Task-Agnostic Spatial Reasoning and Visual Memory

In novel environments, autonomous agents must efficiently gather spatial information to build reliable internal models. Robotics research shows that letting agents learn from an environment before task assignment improves robustness and reliability [3]. To evaluate intrinsic spatial understanding, exploration must be decoupled from extrinsic rewards. With goal-directed optimisation, agents use narrow visual policies and often omit topological features irrelevant to immediate goals. In task-agnostic exploration, the challenge shifts to intrinsic value assignment and selective memory encoding. Without explicit objectives, the agent must rely on intrinsic prioritisation logic to determine which environmental visual features warrant long-term memory retention versus background suppression.

Vision-Language Models (VLMs) are promising for task-agnostic representation learning thanks to their zero-shot semantic reasoning. However, VLMs are typically benchmarked in goal-oriented settings, not unguided spatial exploration. The *Theory of Space* framework [4] evaluates how VLMs construct spatial representations and maintain memory during task-agnostic traversal. However, key questions remain about how VLMs allocate attention and manage memory constraints.

This research focuses on three primary failure modes in VLM spatial representation:
(I) **Priority Inversion:** Lacking intrinsic hierarchical filtration, VLMs overlook or forget important features, while unimportant data is retained, leading to degraded spatial understanding.
(II) **Contextual Hallucination (Semantic Priors):** VLMs are biased by training data. In ambiguous situations, they often hallucinate expected objects or layouts, favouring prior expectations over direct sensory input.
(III) **Topological Instability (Spatial Amnesia):** VLMs may fail to maintain spatial consistency or recognise previously mapped locations from new viewpoints, indicating topological instability.

Future directions. Building robust spatial representations requires filtering irrelevant features, grounding models in sensory data, and preserving topological continuity. To address Priority Inversion, I will evaluate how stateless VLMs structure visual memory within limited context windows. I will assess whether VLMs can build memory hierarchies that prioritise navigationally relevant features over background noise. Second, to address Contextual Hallucination, I will explore methods to force VLMs to anchor spatial representations in sensory data, overriding conflicting priors. Third, to eliminate Topological Instability, I will model spatial feature invariance across viewpoints to ensure consistent topological mapping. Strongly inspired by the *Theory of Space* exploratory approach, my work [5] already deepens some of the topics above in relation to safety-critical scenarios.

Topic 2: Autonomous Policy Optimisation via Self-Play After building an initial environmental representation, an agent can optimise its policy through self-improvement before external deployment. Self-play is a proven framework for policy optimisation in reinforcement learning [6, 7]. Integrating Large Language Models (LLMs) expands self-play: a single foundation model can assume multiple roles to generate dynamic, adversarial training without human annotation.

Self-play is useful across reasoning domains such as mathematics [8], program synthesis [9], and safety alignment [10]. A typical setup uses one base model in three roles: Proposer (task generation), Solver (solutions), and Verifier (evaluation and rewards). This closed-loop creates an automated data pipeline for continuous policy optimisation.

However, self-play frameworks suffer from game-theoretic and computational limitations that degrade policy optimisation. This research isolates three critical vulnerabilities:
(I) **Performance Optimisation Plateaus:** Autonomous self-improvement has capacity limits. Progress depends on generating novel training instances, and the reasons for performance plateaus remain unclear.
(II) **Parameter-Sharing Collusion:** When roles like attacker and defender share model parameters, self-consistency bias arises: adversarial agents avoid generating exploits unsolvable by the defender, causing undetected policy collapse.
(III) **Computational Bottlenecks:** Iterative self-play is computationally intensive, requiring multi-role inference and ongoing fine-tuning without causing instability.

Future directions. Effective self-play requires adversarial pressure, avoiding parameter-sharing collusion, and maximising efficiency. My recent work [11] introduced Anchored Bipolicy Self-Play (ABS) to address these challenges. ABS freezes the base model and trains role-specific LoRA adapters for self-play roles. By decoupling role gradients, ABS eliminates parameter-sharing collusion, restores adversarial pressure, and reduces overhead compared to full fine-tuning. My future research extends this framework in the following directions. First, building on Yue et al. [12], I will investigate causes of performance plateaus in self-play, specifically isolating the effects of: (i) base model capacity for generating out-of-distribution data; (ii) reward formulation; and (iii) RL algorithm dynamics (PPO, GRPO, Reinforce++) under data bias. Second, I will expand diagnostics for parameter-sharing collusion. Using separate models for each role ensures independence but greatly increases memory use. I aim to develop lightweight methods to advance self-play to evolving populations of agents maintaining balanced competition between them.

Topic 3: Multi-Agent Alignment and High-Stakes Coordination

When deployed with other agents or humans, agents must coordinate under shared resource and ethical constraints. Most alignment research focuses on single-agent tuning via RLHF, but multi-agent systems require coordinated constraint adherence under uncertainty [13]. A key goal is building multi-agent coordination mechanisms that meet human safety and fairness standards. Constrained multi-agent reinforcement learning (MARL) algorithms (e.g., MACPO [14], SOTO [15], and Fair-PPO [16]) offer mathematical tools for safe optimisation, but need integration with foundation model architectures.

Incorporating multi-agent game theory into foundation model training embeds social structures into agent protocols [17, 18]. This research targets two unresolved structural challenges:
(I) **Bias vs. Genuine Coordination:** When agents converge on the same strategy, we must distinguish real coordination from shared pre-training data bias. Shared training data may cause apparent alignment, creating a cognitive monoculture.
(II) **Safety and Fairness Under Constraints:** Heterogeneous agents with different capabilities make enforcing baseline safety and fairness challenging. Understanding how agents balance local reward with systemic safety in competitive or scarce-resource settings is key to preventing hazardous behaviours.

Future directions. Reliable multi-agent deployment requires shared situational awareness and verified safety boundary compliance. Coordination strategies must be stress-tested in dynamic multi-agent environments before real-world use. I will evaluate multi-agent coordination in high-stakes scenarios, such as emergency building evacuations with autonomous agents and humans. To distinguish true coordination from pre-training bias, I will test agents with out-of-distribution disruptions (e.g., unexpected barriers on evacuation routes). By analysing whether agents create novel cooperative routes or default to shared failures, I can measure dynamic coordination versus pre-training artefacts. To address multi-agent safety and fairness, I will measure behavioural alignment between agent optimisation and safety protocols under resource competition. I aim to develop evaluation metrics to verify agents enforce safety and fairness constraints, not unconstrained self-optimising behaviours, in high-stress situations.

Conclusion: My research program investigates how autonomous AI agents construct visual world models, execute self-directed policy optimisation, and coordinate within complex multi-agent environments. By explicitly evaluating internal state representations, parameter collusion failure modes, and multi-agent decision dynamics, I aim to map the structural limitations of autonomous models. Leveraging my background in multi-agent reinforcement learning, algorithmic safety and fairness, self-play red-teaming, and visual-spatial reasoning, I aim to advance the computational foundations of verifiable and safe AI agent deployment.

References

[1] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023.

[2] **Gabriele La Malfa**, Lakmal Meegahapola, Edyta Bogucka, Jie M. Zhang, Michael Luck, Elizabeth Black, and Daniele Quercia. Unaccountable delegation, fading skills: Mapping the risks of workplace ai agents, 2026.

[3] Hao Zhu, Raghav Kapoor, So Yeon Min, Winson Han, Jiatai Li, Kaiwen Geng, Graham Neubig, Yonatan Bisk, Aniruddha Kembhavi, and Luca Weihs. Excalibur: Encouraging and evaluating embodied exploration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14931–14942, 2023.

[4] Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *International Conference on Learning Representations (ICLR)*, 2026.

[5] **Gabriele La Malfa**, Nitay Alon, Emanuele La Malfa, Reuth Mirsky, and Stefan Sarkadi. Explore, map, remember, decide: Are embodied vlms ready for safety-critical scenarios?, 2026.

[6] David Silver, Aja Huang, Christopher J. Maddison, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.

[7] OpenAI, Christopher Berner, Greg Brockman, et al. Dota 2 with large scale deep reinforcement learning, 2019.

[8] Thomas Hubert, Rishi Mehta, Laurent Sartran, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651:607–613, 2025.

[9] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[10] Mickel Liu, Liwei Jiang, Yancheng Liang, Simon Shaolei Du, Yejin Choi, Tim Althoff, and Natasha Jaques. Chasing moving targets with online self-play reinforcement learning for safer language models, 2025.

[11] **Gabriele La Malfa**, Emanuele La Malfa, Saar Cohen, Jie M. Zhang, Michael Luck, Michael Wooldridge, and Elizabeth Black. The attacker in the mirror: Breaking self-consistency in safety via anchored bipolicy self-play, 2026.

[12] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[13] Lewis Hammond, Alan Chan, Jesse Clifton, et al. Multi-agent risks from advanced ai, 2025.

[14] Shangding Gu, Jakub Grudzien Kuba, Munning Wen, et al. Multi-agent constrained policy optimisation, 2022.

[15] Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2021.

[16] **Gabriele La Malfa**, Jie M. Zhang, Michael Luck, and Elizabeth Black. Fairness aware reinforcement learning via proximal policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 22725–22733, 2026.

[17] **Gabriele La Malfa**, Emanuele La Malfa, Samuele Marro, Jie M. Zhang, Elizabeth Black, Michael Luck, Philip Torr, and Michael J. Wooldridge. Large language models miss the multi-agent mark. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025.

[18] Lynnette Hui Xian Ng, Iain J. Cruickshank, Adrian Xuan Wei Lim, and Kathleen M. Carley. Social theory should be a structural prior for agentic ai: A formal framework for multi-agent social systems, 2026.