AI already makes a vast number of decisions daily, affecting numerous aspects of human life. This offers great opportunities, but also poses risks if AI does not align with human needs and values. As AI agents and multi-agent systems become increasingly autonomous, we face a new frontier of risk born from how these systems interact with one another, with humans and the surrounding environment [1, 2]. The advent and evolution of Large Language Models (LLMs), including Vision-Language Models (VLMs), Large Action Models (LAMs), and Large Audio Language Models (LALMs), has driven a new wave of research dedicated to both expanding their capabilities, limiting their potential threats, and controlling their high computational resource requirements. The critical frontier is, therefore, finding a sweet spot that balances algorithmic improvement with resource efficiency. To navigate this perimeter safely and efficiently, we need a fundamental understanding of the AI's cognitive lifecycle, from isolated perception to collective action. Using the metaphor of the Expedition, known in psychology as the difficult but purposeful steps into the unknown, I map this developmental journey of autonomous agents across three critical phases.

(1) **The Cartographer (Unguided Memory):** The investigation of how agents construct, retain and update their cognitive map and spatial information without predefined goals and for pure environmental understanding.

(2) **The Grandmaster (Self-Optimisation):** The understanding of agents' autonomous improvement, through mechanisms such as self-play, and its limitations.

(3) **The Journey Together (High-Stakes Coordination):** The exploration of agents' coordination logics and the convergence to common decisions, revealing whether they are the product of isolated shared bias or real coordination.

The Cartographer: How Agents Build their Cognitive Map by Exploring

Analogous to a cartographer entering an uncharted city, the initial action required of an agent deployed in an unfamiliar environment is to gather information. This process enables the agent to perform rational actions and construct a robust internal cognitive map. Foundational robotics research, including the explore-before-execute paradigms [3] enforced by the reinforcement learning community, has demonstrated that permitting an agent to thoroughly map a space prior to task assignment enhances reliability and success. To assess an agent's spatial understanding, it is necessary to fully separate exploration from specific, extrinsic tasks. When guided by an objective, an agent interprets its environment through a narrow perspective, often disregarding or omitting topological details irrelevant to its immediate mission. In contrast, when the directive is an open-ended mandate to explore, the primary challenge transitions from task execution to autonomous value assignment. In the absence of a predefined target, the agent must employ intrinsic logic to determine which elements warrant attention, which should be dismissed as background, and which should be permanently encoded within its constrained memory.

Currently, Vision-Language Models (VLMs) represent a promising paradigm shift for this meta-goal because of their advanced zero-shot semantic reasoning capabilities, which enable intrinsic understanding of visual input. Historically, VLMs have been primarily deployed and tested in goal-driven environments rather than evaluated on their capacity to purely explore. The recent *Theory of Space* framework [**?**] addresses this gap and establishes a methodology to evaluate how VLMs build cognitive maps and maintain baseline memory stability during undirected exploration. Yet, a vast frontier remains in understanding how these models allocate semantic attention and manage memory boundaries.

I want to focus on three aspects/vulnerabilities of VLMs' spatial understanding:
(I) **Priority Inversion:** Without an intrinsic hierarchical filter, some VLMs suffer from catastrophic priority inversion, retaining meaningless visual noise while discarding important navigational anchors.
(II) **The Mirage:** VLMs are burdened by the semantic priors from their training data [4]. When triggered by certain environmental contexts, they can suffer from cognitive mirages, i.e., a bias where they "see" what they are conditioned to expect, hallucinating objects or layouts while ignoring physical evidence from the actual environment.
(III) **Spatial Amnesia:** VLMs exhibit severe topological fragility, suffering from spatial amnesia and losing track of their surroundings the moment they re-explore a previously visited space.

Future directions. A cartographer must prioritise critical structural landmarks over irrelevant background clutter, mapping the landscape as it truly exists rather than how they expect or hope to find it, and maintain a record of where they have already travelled. My future research will target three specific frontiers in VLM spatial reasoning. First, to investigate Priority Inversion, I will analyse how these models structure their memory. Although VLMs are fundamentally stateless, their bounded context window functions as a dynamic memory workspace. Beyond tracking whether objects are captured, I aim to evaluate whether VLMs possess the capacity to construct a memory hierarchy. This involves aligning different levels of environmental objects to immediate sub-goals, final goals, or exploration trajectories to ensure models prioritise structural landmarks over background clutter. Second, to evaluate the Mirage problem, I will explore the tension between objective physical evidence and a VLM's ingrained training bias. I plan to evaluate mechanisms that force models to map the landscape as it truly exists, rather than hallucinating expected layouts triggered by semantic priors. Third, to address Spatial Amnesia, I will examine how VLMs relate objects to their surrounding space and across different visual orientations. This step is necessary to ensure agents maintain topological consistency upon re-exploring a previously visited area.

I will benchmark these three vulnerabilities by analysing how VLM spatial memory behaves when transitioning between task-agnostic exploration and high-stakes execution. While open-ended exploration requires models to structure memory under uniform attention, goal-oriented constraints introduce cognitive overhead that alters an agent's internal value assignment. I will investigate whether these targeted objectives mitigate or accelerate context-window decay. Furthermore, I will conduct these evaluations across both proprietary and open-source models, as differing architectural scales exhibit distinct failure modes. Large proprietary models achieve baseline map stability but lack selective filtering, whereas open-source architectures rely on resource-constrained memory drops, leaving them vulnerable to amnesia if their context decays rapidly. 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.

The Grandmaster: How Agents can Self-Improve Their Skills

Once the cartographer pinpoints the environmental characteristics, it focuses on self-improving before going back to the world outside, becoming the grandmaster. Self-play has long served as a foundational mechanism for autonomous improvement in classical reinforcement learning [6, 7]. The emergence of Large Language Models (LLMs) has transformed this approach. Since LLMs operate through language, they can assume distinct, asym-

metric roles, enabling autonomous generation of adversarial training curricula. As model size increases, self-play provides a way to save the slow and costly process of human data annotation.

Self-play promisingly applies to different domains, like mathematics [8], coding [**?**] or safety [9]. It is often deployed through a single shared model partitioned into three personas: a Proposer that synthesises novel tasks or adversarial prompts, a Solver that addresses them, and a Verifier that evaluates the outputs to generate internal reward signals. This continuous interaction creates a self-synthetic data pipeline for autonomous policy refinement and logical expansion without human intervention.

Self-play presents several areas for further improvement. These areas include both conceptual and game-theoretical challenges and computational limitations. Such challenges threaten the integrity of the autonomous learning loop. I want to focus on three of them:
(I) **The Limits of Self-Improvement:** Autonomous self-improvement does not scale indefinitely. A critical challenge involves identifying the upper limits of self-play capabilities. An agent's improvement is related to its ability to generate novel, high-quality adversarial data. Understanding the reasons of performance plateau is not trivial, and can potentially enable new form of creative interaction.
(II) **The Persona Collusion:** When a single shared model assumes opposing personas, such as an attacker and a defender, it becomes highly susceptible to the risk of false self-consistency. Since the personas share identical underlying weights, they may know each other's latent representations. This telepathic phenomenon can result in persona collusion, which poses significant risks and is challenging to detect, in particular in a naturally human-unsupervised setting.
(III) **Computational Requirements:** Fine-tuning through self-play remains a significant computational challenge, as it requires training large-scale foundation models, synthesising continuous inference across the three personas simultaneously that may grow over time. Alternative paths to optimise computational requirements exist, however, these should not compromise the learning stability.

Future directions. The grandmaster must pursue continuous learning, avoiding mental loopholes that limit creativity, or retreating into established beliefs, pacing themselves to avoid consuming all the resources at their disposal before time. In my recent work, *The Attacker in the Mirror: Breaking Self-Consistency in Safety via Anchored Bipolicy Self-Play* [10], I have already tackled some of these problems. I introduced Anchored Bipolicy Self-Play (ABS), an architecture that, by freezing the base model and training role-specific Low-Rank Adaptation (LoRA) adapters for the self-play personas, explicitly separates gradient updates. This breaks the self-consistency bias inherent in shared-parameter models, restoring genuine adversarial tension while offering higher parameter efficiency than full fine-tuning.

My future research will push these boundaries by addressing three specific frontiers in self-play. First, building upon the foundational work of Yue et al. [**?**], I plan to investigate the root causes behind self-play limits to better understand the generation of novel, high-quality adversarial data. In particular, I will isolate whether these limits are driven by: (i) model capacity, determining if larger architectures produce more out-of-distribution data; (ii) game-theoretic formulations, evaluating reward specification and exploring new theoretical frameworks; or (iii) reinforcement learning and biasing constraints, examining the comparative performance of algorithms like PPO, GRPO, and Reinforce++, alongside how prompt or data bias compromises generative quality. Second, I intend to investigate persona

collusion within single-model self-play. While using distinct models for different personas ensures structural separation and guarantees self-consistency, it defeats the core computational efficiency of self-play. Furthermore, maintaining a strength equilibrium between opposing personas is critical; if one persona significantly outpaces the other, the training loop collapses. To address this, my research will focus on establishing an analytical framework specifically tailored to detect and mitigate collusion in self-play environments.

The Journey Together: Exploring The Foundation of Agents Coordination

Out in the world again, the agent encounters other agents, artificial or human, with different experiences but interdependent goals. Successful coordination in this phase relies on peaceful coexistence and the alignment of core ethical values. Recent work in AI alignment has predominantly focused on single-agent scenarios, relying on techniques like Reinforcement Learning from Human Feedback (RLHF) to ensure individual models adhere to human instructions and safety guidelines. However, as we deploy multiple autonomous agents into shared environments, the paradigm must shift from isolated compliance to high-stakes coordination [11]. The new frontier is exploring agents' coordination logics and how they converge on common decisions respecting value alignment. Various examples of safe and fair multi-agent RL coordination, such as CPO [**?**], MACPO [12], FEN [13], SOTO [**?**], and Fair-PPO [14], can be used for foundation models but require adequate structural integration.

By leveraging insights from multi-agent systems literature, we can embed social logics at both the surface level of agent interaction and the foundational level of model training [**?**, 15]. In this context, my research focuses on two critical challenges:
(I) **The Shared Bias Illusion:** When multiple agents converge on a common decision, we must reveal whether this is the product of real, dynamic coordination or merely an isolated shared bias. Because foundation models are often trained on the same massive internet corpora, their "agreement" may simply be a reflection of a cognitive monoculture rather than genuine negotiation or rational compromise.
(II) **High-Stakes Coordination, Safety and Fairness:** When heterogeneous agents, with different architectures, capabilities, and human-aligned preferences, interact in the environment, establishing a baseline of safety and fairness becomes challenging. We must understand how these agents negotiate conflicting ethical boundaries and whether they can uphold strict safety constraints when their local sub-goals conflict.

Future directions. A safe collective journey requires humans and AI agents to share both situational awareness and underlying values. While foundation models can be fine-tuned for alignment, we must rigorously evaluate their behavioural baselines and limitations before deploying them in multi-agent spatial scenarios. In line with my earlier focus on spatial reasoning, I plan to test high-stakes coordination within safety-critical simulated environments. Specifically, I will use high-stress spatial scenarios, such as a coordinated building evacuation involving humans and embodied AI agents, to investigate both of the aforementioned challenges.

First, to reveal the Shared Bias Illusion, I will assess how multiple agents respond when standard spatial expectations are stressed, such as during an emergency, when a primary evacuation route is blocked. By analysing whether agents develop a novel, coordinated strategy or consistently default to identical failure modes, I can assess whether their consensus indicates genuine collaboration or merely reflects shared training data.

Second, to address High-Stakes Coordination, Safety, and Fairness, I will measure the

alignment between AI decision-making and human-established protocols under pressure. Misalignment between an agent's internal optimisation logic and human safety rules can present significant physical risks. I intend to develop dynamic behavioural metrics to quantify whether agents adhere to strict safety constraints, rather than opting for hazardous, self-optimising behaviours during human-AI interactions.

Conclusion: I am interested in understanding how AI agents explore, evolve, and coordinate in complex real-world systems. I am particularly curious to investigate the internal representation of the external world that AI agents build when learning independently or adapting to the values and behaviours of other agents and humans, and I aim to analyse their failures to map their limitations. I am confident that my experience with multi-agent reinforcement learning, safety and fairness alignment, self-play red teaming, and VLM exploration will allow me to address crucial alignment questions and help to better define the boundaries of AI agents' 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] Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, et al. Mirage: The illusion of visual understanding, 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] 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.

[10] **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.

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

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

[13] Jiechuan Jiang and Zongqing Lu. *Learning fairness in multi-agent systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[14] **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.

[15] 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.