# A Conceptual Framework for Shared Autonomy

Extended Abstract

Shashank Shekhar
Université de Caen Basse Normandie
Caen, France
shashank.shekhar@unicaen.fr

Laurent Jeanpierre
Université de Caen Basse Normandie
Caen, France
laurent.jeanpierre@unicaen.fr

Abdel-Illah Mouaddib
Université de Caen Basse Normandie
Caen, France
abdel-illah.mouaddib@unicaen.fr

## ABSTRACT

Real human-AI systems do not operate at a single level of autonomy. Their effectiveness depends critically on how, when, and under what engagement conditions the human operator intervenes. Inspired by the Pilot Authorization and Control of Tasks (PACT)-style paradigms, we present a decision-theoretic framework for mixed-initiative interaction under shared autonomy. The key idea is to treat the *interaction context* as an object relevant for planning, rather than a collection of ad-hoc autonomy modes and hard-coded constraints. At a high level, our framework enables reasoning about rich interaction protocols within a Markovian planning model, without enumerating autonomy modes or relying on recursive models of human reasoning. We prove that the resulting formulation reduces to a standard Markov decision process (MDP), while remaining compact and scalable, and admits solution via classical MDP solvers such as value iteration. Empirically, this structure enables qualitatively richer policies that adapt not only to the task state but also to evolving interaction context, which yields behaviors beyond the scope of the models available today. All formal proofs and experimental details are deferred to a longer version.

## KEYWORDS

Human-AI teaming; Mixed-initiative planning; Joint actions; MDPs

**ACM Reference Format:**
Shashank Shekhar, Laurent Jeanpierre, and Abdel-Illah Mouaddib. 2026. A Conceptual Framework for Shared Autonomy : Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/QRXH6744

## 1 INTRODUCTION

In this paper, we take the perspective of decision-theoretic planning for mixed-initiative interactions under shared autonomy. Here, we focus on the fundamental case of a *single* human coordinating with a *single* AI agent (e.g., a robot), while the proposed framework is conceptually adaptable to larger multi-agent settings.

Despite significant advances in planning and autonomy, real-world human-AI systems, especially in safety-critical, high-stakes, and nonroutine settings, continue to rely on human involvement [15], yet, existing decision-theoretic models underrepresent when such involvement is solicited and how its effectiveness depends on engagement conditions [17].

Although prior models, such as by Vanhée et al. [18], incorporate constructs like probabilistic support availability and autonomy constraints, they do not treat human engagement and responsiveness factors as an endogenous, planning-relevant variable. They also do not provide an explicit machinery for handling mixed-initiative interactions. We make these capabilities explicit in our framework in a principled manner.

We address this gap by elevating interaction structure to the level of planning model itself, allowing mixed-initiative protocols to be reasoned about, rather than engineered externally.

### Overview of Our Approach

We introduce a decision-theoretic framework that operationalizes PACT-style paradigms (e.g., see [2]) as a structured MDP, where interaction context evolves alongside task state. Importantly, the framework specifies when interaction semantics matter, which is without prescribing how actual task dynamics or preferences are computed. This separation allows interaction-aware planning while remaining agnostic to whether environment dynamics are learned, specified, or treated as black-box.

## 2 THE DECISION-THEORETIC FRAMEWORK FOR SHARED AUTONOMY (CONCEPTUAL ONLY)

The interaction context abstracts who holds initiative, how the counterpart may respond given responsiveness constraints, and whether execution proceeds nominally or transitions into recovery. Precisely, this context is relevant for decision and forward looking: policies reason not only about immediate coordination, but also about how present interaction choices shape future engagement and authority.

### 2.1 Why This Matters

Suppose we are in the context of human-robot interaction, such as a bridge crossing scenario, in which different agents have varying capability dynamics. (We assume that the core primitive actions are implemented on the robot, and it is only the agent's decision that distinguishes.) In a state, it may be the case that the deferred payoff of a particular task done by one agent is better than that of the other agent. For example, suppose there is snow on the bridge, so the human must teleoperate to cross it, even if this choice in this state does not fully align with the executor's internal understanding of efficiency. Our framework effectively captures such nuanced interactions, where a decision to act or delegate must conciliate contextual awareness and differing evaluations of efficiency. Thus,



*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.* © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/QRXH6744

let us say, under declining human engagement, a policy may prefer to delay action, seek confirmation, or proceed autonomously. And, most importantly, these choices are evaluated prospectively and in a principled way: *policies reason about how present interaction decisions affect future authority, engagement, and execution quality.* In conclusion, this enables the policy to make a unified decision-theoretic trade-off or balance between maintaining attentiveness & engagement and ensuring high-quality outcomes.

### 2.2 Positioning

The interaction semantics in our framework is grounded in cognitive science, bounded rationality [14], dual-process theory [10], and heuristic reasoning [4]. Rather than relying on recursive Theory-of-Mind (ToM) models [11], ours achieves context-dependent mixed-initiative teaming via structured interaction protocols. This design aligns with decision-support and semantic-grounding principles [1, 8, 13], where a key factor is attention, which acts as a proxy for cognitive engagement and accountability, both critical in safety-sensitive domains.

While earlier multi-agent MDPs [3, 7, 16] capture cooperation via joint policies or local dependencies, they typically assume symmetric agents with comparable goals and authority. Our framework departs from this assumption by proposing model-asymmetric and semantic structure for interaction, where initiative and responsiveness are jointly determined by interaction contexts. This provides a principled, scalable foundation for mixed-initiative teaming under shared autonomy, where initiative and accountability shift dynamically, beyond what symmetric multi-agent formulations can succinctly represent.

### 2.3 Our Contributions

One main contribution of our work is the reframing of mixed-initiative interaction under shared autonomy as a planning problem over evolving interaction context, rather than as a collection of externally enforced protocols as attempted by prior work. By encoding agent initiative and human responsiveness directly into the decision-theoretic model, our framework enables policies to reason explicitly about when to involve a human, how that shapes outcomes, and whether engagement should be preserved, increased, or discontinued over time.

This reframing expands the scope of planning objectives. Policies are no longer limited to optimizing task performance under fixed authority and-or autonomy assumptions, but they can instead balance execution quality, decision quality, and future coordination capacity. As a result, interaction choices such as seeking confirmation, deferring execution, or acting autonomously emerge as part of optimal planning behavior rather than ad-hoc control logic.

Note that we achieve this expressiveness without the need to introduce recursive reasoning about human beliefs or intentions. Instead, interaction semantics are managed by interaction contexts that encode initiative and engagement constraints directly. This design choice preserves computational tractability and interpretability, while it avoids the complexities associated with ToM-based approaches.

From a modeling perspective, our framework occupies a middle ground between fully symmetric multi-agent formulations [3] and purely teleoperated or autonomous control models [5, 6]. It supports asymmetric authority and responsibility, thus it reflects realistic human-AI teaming assumptions, and it remains compatible with standard MDP algorithms. The framework thus provides a principled foundation for studying mixed-initiative behavior across a range of domains, including safety-critical, high-stakes, and decision-support settings.

In the longer version of this work, we formalize this new framework, analyze its theoretical properties, and instantiate it in a concrete interaction setting to show how richer policies emerge naturally from the proposed structure. These results illustrate that interaction-aware planning need not sacrifice rigor or scalability, and that meaningful human-AI coordination can be captured within a unified decision-theoretic model.

## 3 DESIGN INSIGHTS AND IMPLICATIONS

In this section, our goal is to speak with a modeler, and not just with a theorist. A key implication of our framework is a shift in how one can design a mixed-initiative interactive system. Rather than bringing together distinct elements such as autonomy modes, external control logic, and escalation rules, these elements become an inherent part of the planning problem.

This perspective enables domain modelers to specify interaction principles declaratively [19], while leaving the system to reason about when and how they should be applied, as we show in Section 2.3.

Another key aspect we note is that this improves transparency. As interaction semantics are explicitly represented, we can explain the system behavior in terms of interaction context rather than opaque policy choices [9, 12]. This is particularly relevant for safety-critical domains.

## 4 SCOPE, LIMITATIONS, AND OUTLOOK

We intentionally focus on the fundamental case of a single human coordinating with a single AI agent. We capture the core challenges of mixed-initiative interaction while allowing the interaction structure to be studied in isolation from multi-agent coordination effects.

Our intention is not to model belief states or recursive reasoning [11]. Instead, we adopt a principled abstraction for initiative and engagement as an interaction context, and that is sufficient for planning while remaining tractable and interpretable. Although not required for the intended planning domains, one could always layer richer epistemic or cognitive models on top of this.

Broadly speaking, the proposed framework in this paper should be seen as a modeling foundation and not a fixed instantiation. Different domains may instantiate interaction semantics or dynamics differently, while retaining the same decision-theoretic backbone. One such interaction instantiation appears in the longer version.

One of the promising directions is to extend this model to real multi-agent teams with several humans and AI agents [5].

## ACKNOWLEDGMENTS

This work was supported by the AMI CMA France 2030 NORMANTHIIA project.

## REFERENCES


[1] Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, Edward C Williams, Mina Rhee, Lawson LS Wong, and Stefanie Tellex. 2019. Grounding natural language instructions to semantic goal representations for abstraction and generalization. *Autonomous Robots* 43 (2019), 449–468.

[2] Michael Bonner, Robert Taylor, Keith Fletcher, and Christopher Miller. 2000. Adaptive automation and decision aiding in the military fast jet domain. *Proceedings of Human Performance, Situation Awareness, and Automation* (2000), 154–159.

[3] Craig Boutilier. 1996. Planning, learning and coordination in multiagent decision processes. In *TARK*, Vol. 96. Citeseer, 195–210.

[4] Shelly Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology* 39, 5 (1980), 752.

[5] Clarissa Costen, Anna Gautier, Nick Hawes, and Bruno Lacerda. 2024. Multi-Robot Allocation of Assistance from a Shared Uncertain Operator. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 400–408.

[6] Clarissa Costen, Marc Rigter, Bruno Lacerda, and Nick Hawes. 2022. Shared Autonomy Systems with Stochastic Operator Models. In *Proceedings of IJCAI*. 4614–4620.

[7] Dmitri Dolgov and Edmund Durfee. 2004. Graphical models in local, asymmetric multi-agent Markov decision processes. In *International Conference on Autonomous Agents: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-*, Vol. 2. 956–963.

[8] Stefan Fenz. 2020. Supporting complex decision making by semantic technologies. In *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17*. Springer, 632–647.

[9] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable planning. In *Proceedings of IJCAI Workshop on Explainable AI*. https://arxiv.org/pdf/1709.10256

[10] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[11] Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in artificial intelligence* 5 (2022), 778852.

[12] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[13] Julia V Rogushina and Anatoly Y Gladun. 2022. Semantic Approach to Decision Making in Comparison of Complex Objects. In *Information Technologies and Security (ITS)*. 102–114.

[14] Herbert Simon. 1957. A behavioral model of rational choice. *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting* 6, 1 (1957), 241–260.

[15] Leo van Breda and Timothy Barry. 2020. *Human-Autonomy Teaming: Supporting Dynamically Adjustable Collaboration*. Technical Report. NATO and Science and Technology Organization, TR-HFM-247. https://apps.dtic.mil/sti/citations/AD1183655

[16] Elise van der Pol, Herke van Hoof, Frans A. Oliehoek, and Max Welling. 2022. Multi-Agent MDP Homomorphic Networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

[17] Loïs Vanhée, Laurent Jeanpierre, and Abdel-Illah Mouaddib. 2019. Augmenting Markov Decision Processes with Advising. In *AAAI*. 2531–2538.

[18] Loïs Vanhée, Laurent Jeanpierre, and Abdel-Illah Mouaddib. 2021. Optimizing Requests for Support in Context-Restricted Autonomy. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021*. IEEE, 6434–6440.

[19] Kaiyu Zheng and Stefanie Tellex. 2020. pomdp_py: A framework to build and solve POMDP problems. In *ICAPS 2020 Workshop on Planning and Robotics (PlanRob)*. https://icaps20subpages.icaps-conference.org/wp-content/uploads/2020/10/14-PlanRob_2020_paper_3.pdf