Towards AI-Sympathy Using Agents

Blue Sky Ideas Track

Sebastian Rodriguez
RMIT University
Melbourne, Australia
sebastian.rodriguez@rmit.edu.au

Brian Logan
University of Aberdeen
Aberdeen, UK
Utrecht University
Utrecht, The Netherlands
brian.logan@abdn.ac.uk

John Thangarajah
RMIT University
Melbourne, Australia
john.thangarajah@rmit.edu.au

ABSTRACT

As AI systems become collaborative partners rather than passive tools effective, human–AI teams require reciprocal understanding between systems and their users. Across established team-efficiency models, four foundations consistently emerge: shared understanding, decision-making models, effective communication, and trust. However, these foundations are difficult to achieve as humans often lack accurate mental models of the AI and AI systems have limited representations of human cognitive structures and their own capabilities. As with any new tool, humans must be trained to use AI effectively. Yet AI is unique in its ability to interact and potentially train humans to use itself. In this paper, we introduce three new concepts as essential for genuine human–AI teaming: AI-sympathy (humans understanding AI capabilities and limitations), Human-sympathy (AI systems understanding human context and constraints) and Self-sympathy (the AI system's understanding of its own limitations and capabilities), and present a roadmap for the development of effective human-AI teamwork based on these concepts which identifies four main stages in this evolution.

KEYWORDS

AI-Sympathy, Human-Sympathy, Human-Agent Teams

ACM Reference Format:
Sebastian Rodriguez, Brian Logan, and John Thangarajah. 2026. Towards AI-Sympathy Using Agents: Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 5 pages. https://doi.org/10.65109/TJGL1277

1 INTRODUCTION

Many tools require training to use safely, responsibly, and effectively. We teach children how to handle knives because proper instruction prevents harm and empowers them to use the tool with confidence. As tools become more powerful, we increase the level of training and oversight. For example, driving a car not only requires a formal license but also an understanding of how to coordinate with others on the road, follow shared rules, and behave predictably in a complex social environment.

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/TJGL1277

Artificial Intelligence is no different. As a transformative tool that shapes decisions and interactions, AI demands its own form of literacy. Users must learn not only how to use AI systems, but how to work with them — understanding their strengths, limitations, and the collective behaviours that emerge when humans and AI collaborate. Just as society created driving schools, safety standards, and road rules for cars, we now need structured guidance, training, and norms for AI-enabled work. Well-trained users become empowered users, capable of harnessing the full potential of these technologies while minimizing risks to themselves and others.

The term *mechanical sympathy* was originally coined by Sir Jackie Stewart, the three-time Formula One World Champion, to describe a driving philosophy in which the driver develops a deep understanding of how the car behaves and deliberately drives in ways that minimise mechanical stress while maximising performance [36]. The concept was later adopted and popularised in software engineering by Martin Thompson, who used it to characterise high-performance software design in which programmers understand and respect the underlying hardware architecture (memory hierarchies, CPU pipelines, and data locality), arguing that software systems perform best when engineers respect the hardware's internal dynamics rather than abstracting them away.

In this paper, we define a *cognitive sympathy* framework based on three core ideas: AI-sympathy, human-sympathy and self-sympathy. *AI-sympathy* is a technically grounded, intuition-driven sensitivity to how an AI system processes information, learns, reasons, and fails — allowing humans to interact with, guide, and integrate the AI in ways that maximise performance, stability, transparency, and safety. Developing AI-sympathy is essential because effective, safe, and reliable engagement with artificial intelligence requires a nuanced understanding of how these systems actually process information, generalise, and fail. However, AI-sympathy must be learnt. As most skills this will require a combination of formal training (e.g., core AI paradigm understanding) and practice (i.e., real world experience).

Traditional technologies require external instruction, e.g., manuals, courses, certifications, etc. What makes AI fundamentally different from every tool that came before it is its capacity to teach its human partners how to use it, by providing adaptive, real-time guidance tailored to each user's skills, goals, and context. However, for this potential to be realised, AI must also adapt to its human partners — recognising their capabilities, preferences, limitations, and levels of expertise. For an AI system to be able to engender AI-sympathy in its human users in turn requires that the system has both human-sympathy and self-sympathy. *Human-sympathy* is the capacity of an AI system to model, adapt to, and respect the

cognitive, emotional, contextual, and operational constraints of human users, so that the AI interacts with humans in ways that minimise cognitive friction, support human intentions, and align with human limitations, values, and expectations. Finally, in order for the system to instruct the human in its effective use requires that it has an awareness of its own limitations and capabilities, which we term *self-sympathy*.

In this paper, we define these concepts, review progress toward AI-, human-, and self-sympathy, and propose a research agenda which the AAMAS community is uniquely equipped to address.

2 PRINCIPLES OF COGNITIVE SYMPATHY

Several decades of research has investigated the foundational factors which underpin effective human teamwork. Across well-established frameworks, four characteristics consistently emerge as essential to high-performing teams [4, 16, 21]:

Shared Understanding refers to the team's collective clarity about its mission, goals, priorities, and working norms.

Decision-Making Model defines how choices are made, who holds responsibility for specific outcomes, and the distribution of roles and authority within the team.

Communication & Collaboration encompass the methods and protocols used to share information and work together.

Trust is the confidence team members have in one anothers' integrity, competence, and reliability.

Human–agent teams must embody these characteristics if they are to operate successfully in complex, real-world environments and we argue that this requires fostering *cognitive sympathy*.[1] In the remainder of this section, we present the core principles that shape each type of sympathy.

2.1 AI-Sympathy

AI-sympathy refers to the user's sensitivity to how an AI system processes information, learns, reasons, and fails. Inspired by mechanical sympathy, the core principles are:

(1) *Working with the model, not against it.* Understanding how the AI represents knowledge, handles context, and reacts to inputs, data structures, and system design.
(2) *Aligning behaviours with internal constraints.* Just as mechanical sympathy respects torque curves and heat dissipation, AI-sympathy respects each AI paradigm constraints, such as token limits for Large Language Models (LLM), the specification of rewards in Reinforcement Learning (RL), etc.
(3) *Minimising "friction" in human-AI interaction.* Designing workflows and interfaces that provide the right type of input in the right format at the right time.
(4) *Anticipating emergent failure modes.* Understanding where models fail, e.g,. overfit, or misgeneralise, and adapting engineering and communication practices accordingly.
(5) *Respecting the AI's "computational ergonomics".* Choosing representations, constraints, and instructions that make tasks cognitively "easy" for the underlying AI paradigm.

[1] In the most basic scenario, a team is composed of a single human and a single agent. However, in general, teams may involve multiple humans and agents grouped in multiple (cooperating) teams.

2.2 Human-Sympathy

As AI-sympathy asks humans to understand the nature of AI systems, Human-sympathy asks AI systems to understand the nature of humans. The core principles of Human-sympathy are:

(1) *Working with the user, not against them.* AI must adapt its behaviour, interface, language, and explanations to human cognitive strengths and limitations. This includes clarity, reduced ambiguity, personalised assistance, and supporting incremental understanding.
(2) *Respecting human cognitive constraints and minimising friction.* AI systems should account for limits in memory, attention, workload, and time pressure by adapting outputs to reduce cognitive load—summarising when appropriate, maintaining context, presenting information at the right level of abstraction, providing clear error messages, and supporting natural, low-effort interaction aligned with human workflows.
(3) *Anticipating human reasoning patterns and failure modes.* AI must recognise when humans are likely to misunderstand, misinterpret, overtrust, or misuse the system. This includes modelling: human misunderstanding of AI output, misunderstanding of probabilities or uncertainty, tendencies toward automation bias.
(4) *Respecting human values, emotions, and ethical boundaries.* Human-sympathy requires AI to adjust behaviour to align with human social norms, moral expectations, and emotional context. This includes: safe and respectful communication, transparency about limitations, adapting tone to the user's state, supporting human goals without manipulation.
(5) *Supporting human agency and decision-making.* Rather than replacing human judgement, AI systems must enhance it by offering explanations, alternatives, and uncertainty indicators. Human-sympathy ensures that AI does not dominate or obscure decision processes.

2.3 Self-Sympathy

Self-sympathy is the capacity of an AI system to understand its own internal states, capabilities, and limitations. The core principles of self-sympathy are:

(1) *Self-Representation.* AI maintains an interpretable model of its own reasoning processes, available knowledge, uncertainty, operational boundaries, and performance limitations. This includes recognising when it lacks data, expertise, or the capacity to perform reliably.
(2) *Self-Assessment.* AI continuously evaluates the adequacy of its own outputs relative to task demands, human expectations, and ethical constraints. Critically, it understands when its limitations affect task performance — for example, when uncertainties raise beyond (team-)acceptable limits.
(3) *Self-Adjustment.* AI adapts its behaviour in response to recognised limitations. This includes deferring decisions to humans, asking for clarification, increasing transparency, switching strategies, or reducing autonomy in conditions where it cannot operate safely or effectively.

A key difference between *Self-sympathy* and guardrails specified by a human engineer who anticipates potential problems or limitations of the system and introduces rules to constrain the AI, is that

the agent's own models allows it to detect its own limitations in a particular context.

3 STATE OF THE ART

The agent research community has produced a substantial and diverse literature relevant to the concepts of sympathy we propose. Because this body of work is too extensive to survey comprehensively, we focus here on illustrative examples that capture the community's core contributions.

AI-Sympathy. Ontologies and other knowledge representation approaches contribute to shared understanding by grounding agent knowledge in formal, human-interpretable semantic structures, ensuring that concepts used by agents align with human conceptual frameworks [15, 20]. In addition, there has been considerable work in Agent-Oriented Software Engineering (AOSE) on modelling techniques which make agent architectures, behaviours, and interaction patterns explicitly understandable to human designers and users [6, 10, 23, 31, 35, 37]. Similarly, research on norms [5, 12, 22, 39] provides explicit representations of the expectations, obligations, and constraints guiding agent behaviour, enabling humans to anticipate how agents will respond under different conditions. There is also work on decision models which aim to make agent decision-making both principled and interpretable (e.g. Belief-Desire-Intention (BDI) [30] and goal-oriented architectures). Because these models align closely with human concepts, they naturally support AI-sympathy by exposing the underlying logic of agent choices. Research on agent communication provides additional mechanisms for helping humans understand AI behaviour. Commitment-based interaction models offer structured representations of what an agent has promised to do, under what conditions, and with what consequences—concepts that map naturally to human expectations about cooperation [40, 42]. Other key contributions to building trust in AI systems are formal methods and verification for agents [8]. A final area contributing to AI-sympathy is the extensive work on explainability [2, 17] and responsible AI [9]. Explainable agent reasoning techniques aim to expose the rationale behind actions, plan selections, or belief updates in forms appropriate for human interpretation [32, 33, 41] making them a central piece for AI-sympathy. Clearly, LLM-based agents (aka Agentic AI) will play a major role in enabling AI-sympathy. However, given the many cases of "anti-sympathy" of LLMs, integrating multiple AI techniques is necessary to enable AI systems to develop human- and self-sympathy and ultimately to support the emergence of AI-sympathy in human users.

Human-Sympathy. Ontologies and norms have been used not only to structure agent knowledge but also to represent human concepts, expectations, and conventions, enabling agents to align their interpretations with human reasoning patterns. Complementing this, Theory of Mind-based approaches provide mechanisms for modelling human beliefs, desires, and intentions, allowing agents to form predictive representations of human mental states and adapt their behaviours accordingly [13, 26]. Human-aware practical reasoning builds on BDI-style frameworks to incorporate human conceptualization of goals, values, and situational constraints into the agent's deliberation process, including preference and value modelling [27], human behaviour models that capture likely patterns of action [14], and goal recognition techniques that infer human objectives from observations [1, 24, 25, 29]. Work on assessing cognitive load gives agents the ability to adjust decision-making based on human capacity, attention, and performance, supporting more adaptive and context-sensitive teaming [38]. While there has been considerable work on helping humans trust AI systems, there has been comparatively little research on how agents calibrate their trust in humans. Early directions suggest that combining cognitive load estimation, goal recognition, and behavioural monitoring could allow agents to evaluate human reliability, predict mistakes, and adjust collaboration strategies. This remains a largely open challenge and a key gap for achieving full Human-sympathy.

Self-Sympathy. The concept of introspection in agents has a long history [3, 19]. Work on meta-cognition and self-monitoring equips agents with mechanisms to assess their knowledge gaps, detect reasoning failures, and evaluate plan viability [7, 28]. Research on bounded rationality and resource-aware reasoning contributes techniques for agents to recognise computational or perceptual limits and adjust strategies accordingly [18]. In addition, verification, runtime monitoring, and autonomic computing offer tools for agents to diagnose inconsistencies, detect unexpected behaviours, and adapt plans or goals to maintain stability [8, 11, 34]. Collectively, this body of work lays important groundwork for Self-sympathy, but it remains fragmented and typically oriented toward system robustness rather than deeper self-understanding or the ability to communicate limitations to human or AI teammates.

4 A ROADMAP TO ACHIEVING COGNITIVE SYMPATHY

In this section we present a roadmap for research to achieve cognitive sympathy.

4.1 Stage 1 — Foundations

The first stage aims to create the representational and behavioural bedrock for sympathy-enabled teaming. We imagine AI systems that see human intent, cognitive load, uncertainty, and behavioural patterns — not as heuristics but as first-class modelled entities. Likewise, humans gain access to transparent, legible signals describing the AI's internal states and limitations. We introduce Self-sympathy as a crucial element: AI capable of articulating its own limitations, blind spots, degradations, and failure modes. At this stage, most components are static or defined at design time.

Imagine a human–agent disaster-response team, where first-responders (humans and robots) must attend to a collapsing building. In this stage, the agent-drone can recognise that dust levels exceed its sensor reliability and alerts the human team that its object-detection accuracy is degraded, while humans share their mission priorities (e.g., "locate survivors first, structural mapping second"). Both sides understand basic goals and limitations, but coordination remains simple and mostly manual.

This stage requires: (1) Shared Understanding: Representations of human mental models, AI competence limits, and shared world models. (2) Decision-Making: Early joint decision primitives, uncertainty-aware self-assessment, and human cognitive models. (3) Communication: Bidirectional, interpretable channels to communicate goals, limitations, and expectations. (4) Trust: Foundational trust metrics and safe delegation rules.

4.2 Stage 2 — Integration

The second stage envisions continuous co-adaptation, where humans and AI develop mutual sympathy that evolves during collaboration. AI models incorporate not just human behaviour but human variability, shifting expertise, and moment-to-moment cognitive conditions. Humans, in turn, are trained by the AI to better understand its reasoning, constraints, and strategies — leveraging the unique property that AI is the first tool in history capable of teaching humans how to use the tool itself.

In our example, the drone now adapts its search patterns based on the human team's stress levels, changing commands, and shifting tactical priorities, while also teaching humans how to interpret its uncertainty signals. Humans adjust their behaviour based on real-time AI feedback (e.g., "I need you to move 10 meters north so I can recalibrate my signal"). The team co-adapts during the mission.

This stage requires: (1) Shared Understanding: Adaptive modelling that synchronises human goals, AI goals, and team-level representations. (2) Decision-Making: Mixed-initiative role negotiation driven by real-time human and AI states. (3) Communication: Multimodal interactions channels with adaptive explanations, uncertainty signalling, and self-aware communication. (4) Trust: Dynamic trust calibration to avoid over-reliance and under-reliance, integrating self-sympathy disclosures.

4.3 Stage 3 — Co-Agency

This stage imagines teams where humans and AI operate as true co-agents, each capable of understanding the other and themselves with nuance and self-awareness. AI agents leverage full-spectrum sympathy to become reliable teammates — predictable but adaptive, transparent but efficient, self-aware but collaborative.

In our running example, the AI autonomously identifies structural risks, proposes an evacuation strategy, then defers to human judgement when ethical trade-offs arise (e.g., choosing between two risky routes). Humans and AI fluidly decide who leads which part of the mission, and the AI transparently explains its reasoning and limits. The team operates as coordinated co-agents with shared autonomy — but within known strategies.

This stage requires: (1) Shared Understanding: Continuously evolving joint mental models capturing risk, strategy, context, and human state. (2) Decision-Making: Fluid shared autonomy, where agents knows when to lead, follow, and teach. (3) Communication: Collaborative communication—negotiation of roles, strategies, intent, and uncertainty. (4) Trust: Frameworks integrating behavioural history, team norms, and self-sympathy as source of guardrails.

4.4 Stage 4 — Co-Evolution

In the final, aspirational stage, humans and AI agents transcend coordination, sympathy, and co-agency to enter a regime of co-evolution. In this stage, human–AI teams do not merely collaborate —they mutually shape each other's abilities, developing new forms of teamwork, cognition, and strategy that neither could achieve alone. The team becomes a living, adaptive ecosystem where capabilities continuously emerge and refine across missions, domains, and versions of systems.

At this stage, during repeated missions, the human–agent team discovers and progressively refines a new hybrid rescue tactic, e.g. a coordinated drone–human "triage funnel" method where drones guide humans movement in real time. This method was not designed, programmed, or imagined in advance; it emerges through joint experimentation, learning, and adaptation across deployments. Over time, the new tactic becomes a novel capability of the team itself, not of humans or drones individually.

(1) Shared Understanding: Human and AI partners jointly construct evolving joint mental models that accumulate experience, strategies, and insights. (2) Decision-Making Models: Decision–making becomes a reciprocal learning process in which humans and agents refine each other's heuristics, challenge assumptions, and synthesize new decision policies. The team gradually develops new, emergent decision-making paradigms irreducible to either species alone. (3) Communication & Collaboration: Through sustained co-learning, interaction channels evolve from explanation and negotiation into collaborative ideation, where humans and agents co-create new communication protocols, shorthand languages, and mixed-initiative workflows. (4) Trust Trust becomes a dynamic, bi-directional construct that adapts not only to performance but to growth trajectories, shared experiences, and emerging capabilities.

5 CONCLUSION

This paper has proposed a new framework to conceptualise human-AI teaming through the complementary notions of Human-sympathy, AI-sympathy, and Self-sympathy. Building on the four pillars of high-performing human teams—shared understanding, decision-making, communication, and trust—we argue that sympathy in these three forms is essential for enabling the next generation of adaptive, resilient, and safe human-agent partnerships.

By reframing intelligent agents as teammates that must understand humans, be understood by humans, and understand themselves, we introduced a four-stage research agenda that charts a path from foundational representational capabilities, through dynamic co-adaptation, toward mature co-agency, and ultimately co-evolution. The progression illustrates how human-agent teams will evolve from systems that merely exchange information to teams capable of fluid joint autonomy and the development of entirely new capabilities. This trajectory highlights how agent can actively teach, shape, and learn from their human partners.

In this work, we have highlighted major stepping stones of this evolution. However, the community will have build the scaffolding to support this progress. In this context, some important research direction include: (1) Self-sympathy as an AI-reflective model to enable systems to reason about themselves; (2) definition of metrics and benchmarks to evaluate performance of cognitive sympathy (e.g. efficiency of communication, human mental models); (3) standards and reference architectures for human-agent teams (e.g. goal representation, protocol models); (4) shared interpretable decision–making models (e.g. human interpretable models, adaptive social and cooperation norms) (5) security and safety in cognitive sympathy (e.g. how to prevent deception or unethical use of AI-sympathy) (6) resilience and fault tolerance of models.

Our hope is that this Blue Sky agenda catalyses a new research direction for the AAMAS community–one where agents are designed with built-in sympathy mechanisms to train and cooperate with humans to its true potential.

REFERENCES

[1] Leonardo Amado, Sveta Paster Shainkopf, Ramon Fraga Pereira, Reuth Mirsky, and Felipe Meneguzzi. 2024. A Survey on Model-Free Goal Recognition. In *Thirty-Third International Joint Conference on Artificial Intelligence*, Vol. 9. 7923–7931. https://doi.org/10.24963/ijcai.2024/877

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[3] John Batali. 1983. Computational Introspection. (Feb. 1983).

[4] Richard Beckhard. 1972. Optimizing Team-Building Efforts. *Journal of contemporary business* 1, 3 (1972), 23–32.

[5] Cristiano Castelfranchi. 2003. Formalising the Informal?: Dynamic Social Order, Bottom-up Social Control, and Spontaneous Normative Relations. *Journal of Applied Logic* 1, 1 (Feb. 2003), 47–92. https://doi.org/10.1016/S1570-8683(03)00004-1

[6] Massimo Cossentino, Vincent Hilaire, Ambra Molesini, and Valeria Seidita (Eds.). 2014. *Handbook on Agent-Oriented Design Processes*. Springer-Verlag, Berlin Heidelberg.

[7] Michael T. Cox. 2005. Metacognition in Computation: A Selected Research Review. *Artificial Intelligence* 169, 2 (Dec. 2005), 104–141. https://doi.org/10.1016/j.artint.2005.10.009

[8] Louise A. Dennis and Michael Fisher. 2023. *Verifiable Autonomous Systems: Using Rational Agents to Provide Assurance about Decisions Made by Machines*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY, USA.

[9] Virginia Dignum. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*.

[10] Jürgen Dix, Koen V. Hindriks, Brian Logan, and Wayne Wobcke. 2012. Engineering Multi-Agent Systems (Dagstuhl Seminar 12342). *Dagstuhl Reports* 2, 8 (2012), 74–98. https://doi.org/10.4230/DagRep.2.8.74

[11] Debora C. Engelmann, Angelo Ferrando, Alison R. Panisson, Davide Ancona, Rafael H. Bordini, and Viviana Mascardi. 2022. RV4JaCa – Runtime Verification for Multi-Agent Systems. *Electronic Proceedings in Theoretical Computer Science* 362 (July 2022), 23–36. https://doi.org/10.4204/EPTCS.362.5 arXiv:2207.09708 [cs]

[12] Marc Esteva, Julian Padget, and Carles Sierra. 2002. Formalizing a Language for Institutions and Norms. In *Intelligent Agents VIII*, G. Goos, J. Hartmanis, J. Van Leeuwen, John-Jules Ch. Meyer, and Milind Tambe (Eds.). Vol. 2333. Springer Berlin Heidelberg, Berlin, Heidelberg, 348–366. https://doi.org/10.1007/3-540-45448-9_26

[13] Chris Frith and Uta Frith. 2005. Theory of Mind. *Current Biology* 15, 17 (Sept. 2005), R644–R645. https://doi.org/10.1016/j.cub.2005.08.041

[14] Andrew Fuchs, Andrea Passarella, and Marco Conti. 2023. Modeling, Replicating, and Predicting Human Behavior: A Survey. *ACM Transactions on Autonomous and Adaptive Systems* 18, 2 (May 2023), 4:1–4:47. https://doi.org/10.1145/3580492

[15] F. Gandon. 2002. *Ontology Engineering a Survey and a Return of Experience*. Technical Report RR-4396. INRIA, France.

[16] Google. 2025. Google Re:Work - Guides: Understand Team Effectiveness. https://rework.withgoogle.com/intl/en/guides/understanding-team-effectiveness.

[17] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable Artificial Intelligence. *Science Robotics* 4, 37 (Dec. 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[18] Eric J. Horvitz. 1987. Reasoning about Beliefs and Actions under Computational Resource Constraints. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence (UAI'87)*. AUAI Press, Arlington, Virginia, USA, 429–447.

[19] Kurt Konolige. 1985. A Computational Theory of Belief Introspection. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1 (IJCAI'85)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 502–508.

[20] H J Levesque. 1986. Knowledge Representation and Reasoning. *Annual Review of Computer Science* 1, 1 (June 1986), 255–287. https://doi.org/10.1146/annurev.cs.01.060186.001351

[21] Pat MacMillan. 2001. *The Performance Factor: Unlocking the Secrets of Teamwork*. Broadman & Holman Publishers, Nashville, Tenn.

[22] Moamin A. Mahmoud, Mohd Sharifuddin Ahmad, Mohd Zaliman Mohd Yusoff, and Aida Mustapha. 2014. A Review of Norms and Normative Multiagent Systems. *The Scientific World Journal* 2014, 1 (2014), 684587. https://doi.org/10.1155/2014/684587

[23] Viviana Mascardi, Danny Weyns, Alessandro Ricci, Clara Benac Earle, Arthur Casals, Moharram Challenger, Amit Chopra, Andrei Ciortea, Louise A. Dennis, Álvaro Fernández Díaz, Amal El Fallah-Seghrouchni, Angelo Ferrando, Lars-Åke Fredlund, Eleonora Giunchiglia, Zahia Guessoum, Akin Günay, Koen Hindriks, Carlos A. Iglesias, Brian Logan, Timotheus Kampik, Geylani Kardas, Vincent J. Koeman, John Bruntse Larsen, Simon Mayer, Tasio Méndez, Juan Carlos Nieves, Valeria Seidita, Baris Tekin Teze, László Z. Varga, and Michael Winikoff. 2020. Engineering Multi-Agent Systems: State of Affairs and the Road Ahead. *SIGSOFT Softw. Eng. Notes* 44, 1 (Oct. 2020), 18–28. https://doi.org/10.1145/3310013.3322175

[24] Peta Masters and Sebastian Sardina. 2021. Expecting the Unexpected: Goal Recognition for Rational and Irrational Agents. *Artificial Intelligence* 297 (Aug. 2021), 103490. https://doi.org/10.1016/j.artint.2021.103490

[25] Felipe Meneguzzi and Ramon Fraga Pereira. 2021. A Survey on Goal Recognition as Planning. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, Vol. 5. 4524–4532. https://doi.org/10.24963/ijcai.2021/616

[26] Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, and Carles Sierra. 2023. Combining Theory of Mind and Abductive Reasoning in Agent-Oriented Programming. *Auton Agent Multi-Agent Syst* 37, 36 (2023). https://doi.org/10.1007/s10458-023-09613-w

[27] Nieves Montes, Nardine Osman, Carles Sierra, and Marija Slavkovik. 2023. Value Engineering for Autonomous Agents. https://doi.org/10.48550/arXiv.2302.08759 arXiv:2302.08759

[28] Jorg P. Muller. 1997. *The Design of Intelligent Agents: A Layered Approach* (1st ed.). Springer-Verlag, Berlin, Heidelberg.

[29] Artem Polyvyanyy, Zihang Su, Nir Lipovetzky, and Sebastian Sardina. 2020. Goal Recognition Using Off-The-Shelf Process Mining Techniques. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1072–1080.

[30] Anand S. Rao and Michael P. Georgeff. 1995. BDI Agents : From Theory to Practice. In *Proceedings of the First International Conference on Multi-agent Systems (Icmas-95)*. 312–319.

[31] Sebastian Rodriguez, Akhila Bairy, Matteo Baldoni, Patrick Benjamin, Constantin Blessing, Nicolas Brandstetter, Amit K. Chopra, Thomas Clemen, Louise A. Dennis, Ahmad Esmaeili, Lu Feng, Angelo Ferrando, Zahra Ghorrati, Victor Guillet, Önder Gürcan, Soham Hans, James Herber, Viviana Mascardi, Marcel Mauri, Jörg P. Müller, John Thangarajah, Rafał Tyl, and Yi Yang. 2026. Engineering the Next Generation of Multi-agent Systems: A Community Roadmap from EMAS 2025. In *Engineering Multi-Agent Systems*, Sebastian Rodriguez, Lu Feng, and Jörg P. Müller (Eds.). Springer Nature Switzerland, Cham, 238–258. https://doi.org/10.1007/978-3-032-18011-7_14

[32] Sebastian Rodriguez and John Thangarajah. 2024. Explainable Agents (XAg) by Design. In *Proceedings of the 2024 International Conference on Autonomous Agents and Multiagent Systems (Blue Sky) (AAMAS '24)*. Auckland, New Zeland, 2712–2716.

[33] Sebastian Rodriguez, John Thangarajah, and Andrew Davey. 2024. Design Patterns for Explainable Agents (XAg). In *Proceedings of the 2024 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24)*. Auckland, New Zeland, 1621–1629.

[34] Sebastian Rodriguez, John Thangarajah, and Michael Winikoff. 2026. Requirements-Based Explainability for Multi-agent Systems. In *AI 2025: Advances in Artificial Intelligence*, Miaomiao Liu, Xin Yu, Chang Xu, and Yiliao Song (Eds.). Springer Nature, Singapore, 246–259. https://doi.org/10.1007/978-981-95-4969-6_19

[35] Onn Shehory and Arnon Sturm (Eds.). 2014. *Agent-Oriented Software Engineering*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-54432-3

[36] Sir Jackie Stewart and Peter Manso. 1972. *Faster!* HarperCollins Distribution Services, London.

[37] Arnon Sturm and Onn Shehory. 2014. Agent-Oriented Software Engineering: Revisiting the State of the Art. In *Agent-Oriented Software Engineering*. Springer, Berlin, Heidelberg, 13–26. https://doi.org/10.1007/978-3-642-54432-3_2

[38] Maria Vukovic. 2022. *Using Speech Interfaces to Support Human Task Performance in Complex Training Environments*. Ph.D. Dissertation. RMIT University, Melbourne, Australia.

[39] Michael Winikoff, Frank Dignum, Sebastian Rodriguez, and John Thangarajah. 2026. Engineering Norm-Aware BDI Agents (Extended Abstract). In *25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*. Paphos, Cyprus.

[40] Michael Winikoff, Wei Liu, and James Harland. 2005. Enhancing Commitment Machines. In *Declarative Agent Languages and Technologies II*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, João Leite, Andrea Omicini, Paolo Torroni, and pInar Yolum (Eds.). Vol. 3476. Springer Berlin Heidelberg, Berlin, Heidelberg, 198–220. https://doi.org/10.1007/11493402_12

[41] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. 2021. Why Bad Coffee? Explaining BDI Agent Behaviour with Valuings. *Artificial Intelligence* 300 (Nov. 2021), 103554. https://doi.org/10.1016/j.artint.2021.103554

[42] Pınar Yolum and Munindar P. Singh. 2002. Commitment Machines. In *Intelligent Agents VIII*, John-Jules Ch. Meyer and Milind Tambe (Eds.). Springer, Berlin, Heidelberg, 235–247. https://doi.org/10.1007/3-540-45448-9_17