Theoretical and Empirical Evidence for Trust in Embodied Agents: Integrating Conscious Agent Theory and Computational Trust Models (Short Paper)

Kevin Babashov[1] and Maria Gini[1][0000-0001-8841-1055]

University of Minnesota, Minneapolis, MN 55455, USA
{babas007,gini}@umn.edu

Abstract. This paper connects abstract theoretical ideas about trust, grounded in Conscious Agent Theory (CAT), with probabilistic computational trust models such as TRAVOS. Sufficient conditions for the stability of trust dynamics in embodied multi-agent systems are derived, where embodiment modulates both perception and cooperation. In particular, a convergence condition for trust updates is presented, incorporating an embodiment similarity function $\phi(i, j)$. The theory is complemented by simulations using learning agents that estimate trust across varying levels of embodiment similarity. The results indicate that high morphological and sensorimotor similarity leads to faster trust convergence and higher asymptotic trust values, whereas low similarity produces lower plateaus. These trends align with cognitive bias analogs: agents with high similarity quickly establish mutual trust (akin to in-group bias), while agents with low similarity remain more skeptical (analogous to out-group bias), absent significant evidence. Overall, the findings motivate integrating embodiment factors into trust models and provide theoretical and empirical support for their impact on trust dynamics.

Keywords: Computational trust · Embodied agents · Multi-agent systems · TRAVOS · Conscious Agent Theory · Trust dynamics · Open systems governance

1 Introduction

Trust is a foundational component of decision-making in autonomous multi-agent systems, especially in contexts involving cooperation under uncertainty [15,5,16]. Traditional computational trust models, including Beta-distribution reputation systems and specific frameworks such as TRAVOS [25,24], estimate trustworthiness from interaction outcomes [14,11,2]. In many real deployments, agents are embodied [18,19], in large numbers [13,10], and can be heterogeneous: they differ in morphology, sensors, and actuation. These differences influence what agents can observe, how they interpret outcomes, and which interactions

occur. When embodiment is ignored, trust formation and convergence can be mischaracterized, particularly in heterogeneous systems [7,8].

In human societies and embodied agent teams, perceived similarity can bias trust. An agent that looks or senses similarly to another may be treated as more trustworthy even before sufficient interaction evidence accumulates. We model this effect by introducing an embodiment-modulated trust update. The core idea is to weight evidence by an embodiment-similarity function $\phi(i, j)$ that quantifies the compatibility between the embodiments of a trustor i and a trustee j. We analyze how this weighting constrains equilibrium trust values and affects convergence behavior, and illustrate the resulting dynamics through simulations in which learning agents form trust estimates over time.

This perspective is relevant to human-robot collaboration, decentralized robotic teams, and cyber-physical systems in which agents differ not only in performance but also in form and perception [22,9]. Accounting for embodiment in trust updates supports more robust coordination and yields trust dynamics that are easier to interpret in heterogeneous open systems.

2 Related Work

Computational trust in multi-agent systems has a long history spanning formal models of trust and delegation, Bayesian reputation systems, and probabilistic trust estimation [15,5,14]. Within this line, TRAVOS-style approaches [25,24,23] treat trust as an estimate of a trustee's future reliability based on evidence, typically aggregated via Bayesian updates. These models are effective when interaction outcomes are observable and comparable across agents, and when observational processes are aligned across the population.

A separate body of work in robotics and Human Robot Interaction emphasizes that embodiment changes the interaction dynamics and perception. Differences in sensors, morphology, and action capabilities influence which events are observed, how actions are interpreted, and what counts as success or failure [8,9,22]. In heterogeneous teams, these differences can create systematic asymmetries: two agents can experience the same world state but extract different observations, which changes both decision-making and trust evidence.

Work on cooperation, repeated interaction, and game-theoretic learning clarifies how trust-like variables shape equilibrium selection and the emergence of cooperation in repeated social dilemmas [3,4,20,17]. In these settings, small biases in early interactions can lead to persistent path dependence, clustering, and polarization. We treat embodiment similarity as a structured bias acting on evidence integration rather than directly on partner selection.

Conscious Agent Theory (CAT) frames perception and action as fundamental primitives and emphasizes that an agent's experienced reality depends on perceptual mapping and internal dynamics [12]. We use CAT as a conceptual motivation for making embodiment explicit in trust updates: if perception is embodiment-dependent, then evidence incorporation should be embodiment-dependent as well. We do not require full CAT commitments for the mechanics of the update

rule; rather, CAT motivates a principled path for modeling perception-dependent trust evidence in embodied open systems.

COINE emphasizes coordination, organizations, institutions, norms, and ethics as governance elements for open multi-agent systems. Trust affects delegation, compliance with norms, information sharing, and coordination under uncertainty. In open systems where agents differ in form and capability, trust mechanisms that ignore embodiment can unintentionally produce systematic disadvantages for certain groups of agents. We return to this governance perspective in Section 6.

3 Methodology and Approach

3.1 Mathematical Framework for Embodied Trust

We draw on CAT [12] to conceptually ground the modeling choice that embodiment should shape the integration of trust evidence. In an embodied multi-agent context, we represent each agent a_i by

$$C_i = (X_i, A_i, W, M_i, \pi_i, T_i, E_i),$$

where X_i is the experience space, A_i the action space, W the world state space, M_i and T_i are perceptual and transition kernels, π_i is the decision policy, and E_i is an embodiment descriptor capturing physical and sensorimotor properties (morphology, sensor modalities, autonomy level). This framing motivates explicit embodiment dependence in trust dynamics.

We define directed trust $T_{ij}(t)$ as the trust that agent i places in agent j at time t. A standard Bayesian approach aggregates evidence with a Beta distribution: after α_{ij} successes and β_{ij} failures observed by i about j, trust can be represented by $\text{Beta}(\alpha_{ij}+1, \beta_{ij}+1)$ [14]. The expected value, $\frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}}$, is often used as a scalar trust estimate and underlies TRAVOS-style approaches [25,24].

We introduce an embodiment similarity factor $\phi(i,j) = \phi(E_i, E_j) \in [0,1]$ as a compatibility weight on evidence: when embodiments are similar, evidence impacts trust more strongly; when embodiments differ, evidence is discounted. Trust is updated as:

$$T_{ij}(t+1) = (1-\eta)\, T_{ij}(t) + \eta\, \phi(i,j)\, \frac{\alpha_{ij}(t)}{\alpha_{ij}(t) + \beta_{ij}(t)}, \tag{1}$$

where $\eta \in (0,1]$ is the step size. When $\phi(i,j) = 1$, (1) reduces to the classic form. When $\phi(i,j) = 0$, evidence is ignored, and trust remains at its initial value under this update.

3.2 Embodiment Similarity Functions

In the simplest scalar case, we use:

$$\phi(i,j) = \max\{0,\ 1 - |E_i - E_j|\}. \tag{2}$$

For vector descriptors $E_i \in \mathbb{R}^d$, natural choices include a Radial Basis Function (RBF) kernel or cosine similarity mapped to $[0, 1]$:

$$\phi_{\mathrm{rbf}}(i,j) = \exp\left(-\frac{\|E_i - E_j\|_2^2}{2\sigma^2}\right), \quad \phi_{\cos}(i,j) = \frac{1 + \cos(E_i, E_j)}{2}. \tag{3}$$

ϕ is treated as a modeling choice that reflects the extent to which embodiment alignment affects evidence integration.

3.3 Phenomenological Perspective

$\phi(i,j)$ is interpreted as a formal encoding of perceived similarity or identification. High ϕ corresponds to settings in which the trustor treats the trustee as "like itself" in relevant ways, mirroring similarity-driven trust effects in human contexts. Low ϕ corresponds to otherness or misalignment, in which repeated positive outcomes may be discounted and trust may plateau far below objective reliability. This view suggests that improving inter-embodiment interpretability can effectively increase ϕ over time and mitigate persistent mistrust [6].

3.4 Closed-Form Trajectory Under Stationary Evidence

Assume that, after sufficient interaction, the evidence estimate becomes effectively stationary in expectation:

$$\frac{\alpha_{ij}(t)}{\alpha_{ij}(t) + \beta_{ij}(t)} \approx p_{ij},$$

where $p_{ij} \in [0, 1]$ is the long-run probability of success as perceived by i. Then (1) becomes

$$T_{ij}(t+1) = (1-\eta)T_{ij}(t) + \eta\,\phi(i,j)p_{ij},$$

with a closed-form solution

$$T_{ij}(t) = (1-\eta)^t T_{ij}(0) + \left(1 - (1-\eta)^t\right)\phi(i,j)p_{ij}. \tag{4}$$

Equation (4) implies an embodiment-limited plateau at $\phi(i,j)p_{ij}$ and geometric convergence governed by η.

Theorem 1 (Embodiment-Limited Equilibrium). *Under stationary evidence with success probability p_{ij}, the trust sequence defined by* (1) *converges to*

$$T_{ij}^* = \phi(i,j)p_{ij}$$

for all $i \neq j$.

Proof. Taking the limit of (4) as $t \to \infty$ yields $(1-\eta)^t \to 0$, which implies $T_{ij}(t) \to \phi(i,j)p_{ij}$.

3.5 Stochastic Evidence and Consistency

In finite simulations, $\frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}}$ is stochastic because it is estimated from Bernoulli outcomes. Under standard conditions, this estimator converges in probability to p_{ij} as interaction counts grow. The trust update (1) can be interpreted as a stochastic approximation tracking the stationary recursion, so trust concentrates near $\phi(i,j)p_{ij}$ as evidence accumulates.

4 Data and Simulation Setup

To validate the theoretical behavior, we simulate a population of interacting agents and track how embodiment similarity affects trust trajectories and final trust values.

We simulate $N = 20$ agents. Each agent has a scalar embodiment feature $E_i \in [0,1]$ sampled uniformly at random. Embodiment similarity is defined as:

$$\phi(i,j) = \max\{0,\ 1 - |E_i - E_j|\}.$$

Each agent j is assigned an inherent reliability $r_j \in [0,1]$, sampled uniformly from $[0.5, 0.9]$.

Agents interact over discrete time steps. At each step, we sample a random ordered pair (i,j) with $i \neq j$. Agent i decides whether to trust j using an ϵ-greedy Q-learning rule [21,1]. If i trusts j, the interaction succeeds with probability r_j (reward $+1$) and fails with probability $1 - r_j$ (reward -1). If i does not trust j, the reward is 0.

When a trust attempt occurs, we update α_{ij} or β_{ij} and then update trust using (1) with $\eta = 0.1$. We initializa $T_{ij}(0) = 0.5$ for $i \neq j$ and set $T_{ii} = 0$. We run the simulation for 25,000 steps. To illustrate the effect of embodiment, representative high-similarity and low-similarity directed pairs are logged, along with the final trust matrix.

5 Results and Analysis

The simulations support the theoretical predictions. For high-similarity pairs (Fig. 1), trust typically rises rapidly from 0.5 to a high plateau and stabilizes after relatively few interactions. This aligns with Theorem 1: when $\phi(i,j) \approx 1$, equilibrium trust approaches p_{ij}, the trustee's effective reliability.

For low-similarity pairs (Fig. 2), trust grows slowly or decreases initially and plateaus well below what would be expected without embodiment modulation. This matches the equilibrium $T^*_{ij} = \phi(i,j)p_{ij}$: low ϕ caps achievable trust. Finite sampling and ϵ-greedy exploration produce small deviations, but the plateau effect persists.

The final trust matrix (Fig. 3) shows consistent global patterns: directed asymmetries due to stochastic interaction histories, high-trust regions when reliability and similarity are both high, and cluster structure in which agents with

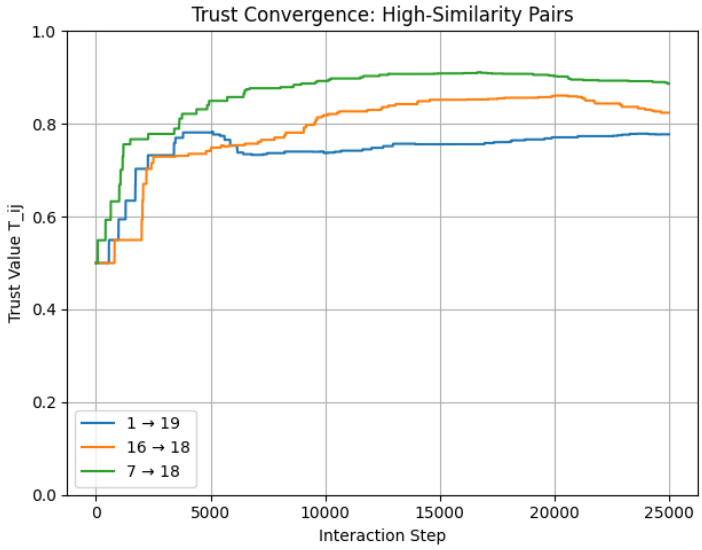

Fig. 1. Trust convergence for high-similarity pairs. Each curve shows T_{ij} over time for a selected directed pair $(i \to j)$ with $\phi(i,j) \approx 1.0$. Trust rises quickly and stabilizes near a high plateau.

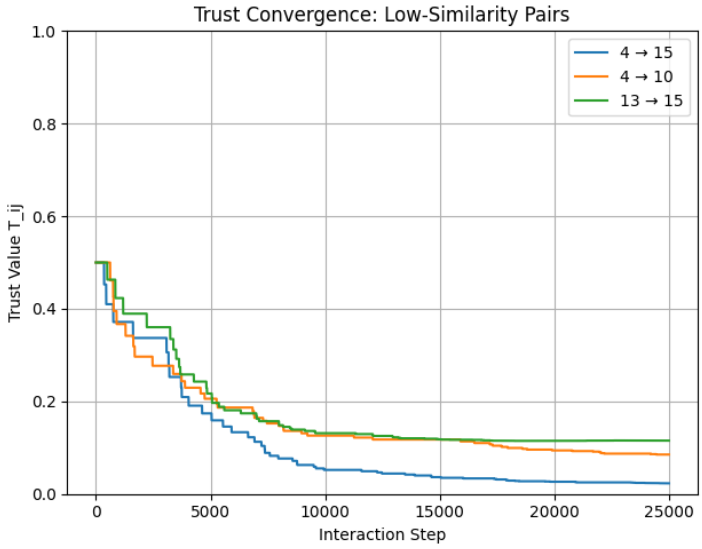

Fig. 2. Trust convergence for low-similarity pairs. For pairs with $\phi(i,j)$ near 0, trust grows slowly and plateaus at lower values. Early negative outcomes can cause persistent decreases because low ϕ limits later positive recovery.

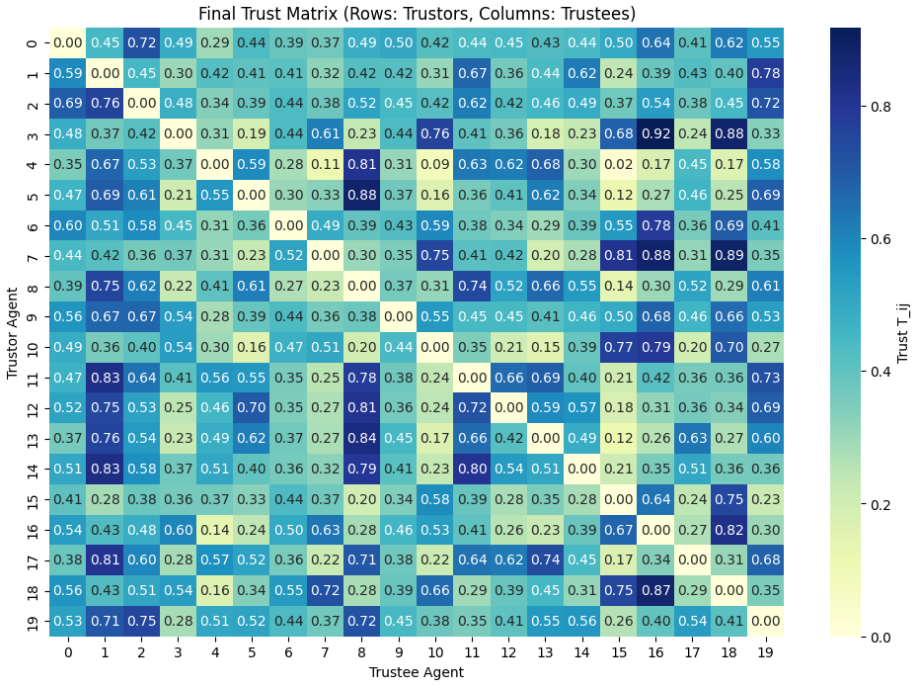

Fig. 3. Final trust matrix heatmap. Cell (i, j) shows trustor i's final trust in trustee j after 25,000 steps. The structure reflects both trustee reliability and embodiment similarity, including clustering among agents with similar E.

similar embodiments develop mutual trust. This suggests a system-level consequence: embodiment differences can induce persistent segmentation in trust relationships even when objective reliability is moderate. In open systems, this kind of segmentation can reduce coordination efficiency and lead to underutilization of capable agents who happen to be dissimilar.

5.1 Sensitivity Analyses and Extensions

We evaluate several variations to clarify which components drive the observed behavior.

Step size. Varying η changes convergence speed and smoothness. Larger η yields faster adaptation but higher variance under stochastic evidence; smaller η yields smoother curves with slower convergence. The plateau structure remains governed by $\phi(i, j)$ because the limiting value depends on $\phi(i, j)p_{ij}$.

Population heterogeneity. Increasing the spread of embodiments increases the frequency of low-similarity pairs and strengthens trust clustering. Decreasing the spread increases average similarity and weakens segmentation. This connects directly to open-systems governance, since heterogeneity can induce fragmentation even in the absence of explicit group labels.

Non-stationary trustees. When a subset of agents changes reliability mid-simulation, embodiment-modulated updates can be advantageous or risky depending on evidence availability. High similarity can accelerate adaptation because evidence is weighted strongly, but sparse interactions can still yield transient over-trust. This motivates the addition of governance constraints, such as minimum-evidence requirements for high-stakes delegation.

6 Governance Implications for Open MAS

COINE emphasizes governance elements – coordination, organizations, institutions, norms, and ethics – in open multi-agent systems. Embodiment-aware trust interacts with each element and can either improve diagnosability or introduce bias amplification if left unconstrained.

6.1 Coordination and Delegation

In many coordination protocols, trust determines delegation: agents allocate tasks, share resources, and accept commitments based on perceived reliability. In heterogeneous embodied teams, observational misalignment can make certain agents systematically less trusted, even when they are reliable. The similarity term $\phi(i, j)$ makes this effect explicit and measurable. When coordination fails due to systematic distrust across embodiment gaps, the failure can be traced to issues of compatibility and evidence integration rather than to malicious behavior.

6.2 Organizations and Institutions

Organizations and institutions shape interaction patterns through roles, authority, and structured communication. Embodiment can be broadly interpreted to include interfaces and sensing modalities. Two agents can become institutionally compatible (e.g., through shared protocols, audit trails, and standardized reporting) even if they are physically separated. In such settings, E_i can include institutional interface features, and $\phi(i, j)$ becomes a measure of institutional interoperability. Improving interoperability increases effective similarity and raises the trust ceiling for cross-embodiment interactions.

6.3 Norms, Compliance, and Trust Calibration

Norm compliance depends on credible monitoring and enforcement. If monitoring capabilities differ across embodiments, compliance evidence can be systematically biased. The formulation highlights that trust depends on both the quality of evidence and behavior. A governance mechanism can treat $\phi(i, j)$ as a proxy for evidence comparability and can require additional verification when similarity is low.

6.4 Ethics: Bias Amplification and Mitigation

Similarity-weighted trust can encode and amplify bias: dissimilar agents can remain permanently under-trusted even when reliable, leading to fragmentation and reduced collective performance. This motivates mitigation strategies: constrain how strongly similarity can suppress trust, increase interoperability through institutional interfaces, impose minimum-evidence requirements before extreme trust differentials are allowed, and explicitly monitor for segmentation that is not explained by objective performance differences.

7 Limitations

The work presented has several limitations. First, we use a simplified embodiment descriptor and similarity function; realistic deployments require multi-dimensional descriptors and careful measurement of interoperability and perceptual alignment. Second, we treat similarity as exogenous and mostly static. In real systems, communication and interaction protocols can adapt and effectively increase compatibility over time. Third, the simulation uses a simple binary outcome and one-step rewards; richer tasks involve partial observability, delayed credit assignment, and richer notions of success and failure. Finally, trust is only one governance signal; effective open MAS design also requires explicit norms, enforcement, accountability, and transparency mechanisms that can override purely experience-based trust.

8 Conclusions and Future Work

We introduced an embodiment-aware trust model that integrates a CAT-motivated embodiment perspective with TRAVOS-style Bayesian trust estimation. By incorporating an embodiment similarity factor $\phi(i, j)$ into the trust update, we obtain an embodiment-limited trust plateau and an interpretable characterization of the equilibrium. The simulations demonstrate that greater similarity leads to faster trust formation and higher asymptotic trust, whereas lower similarity leads to slower growth and lower plateaus.

We plan on extending the model in four directions: (1) learn $\phi(i, j)$ from interaction traces and cross-modal alignment signals rather than specifying it by hand, (2) incorporate structured institutions that modify evidence quality and auditing, (3) explore strategic trustees that manipulate trust under embodiment constraints, and (4) evaluate the approach in richer embodied simulations where sensing and actuation differences change what outcomes are observed and how they are interpreted.

Code Availability

The code is available at the following link: GitHub Link

References

1. Alpaydin, E.: Introduction to Machine Learning. MIT Press, Cambridge, MA (2020)
2. Awan, K.A., Din, I.U., Almogren, A., Han, Z., Guizani, M.: TrustAware-GNN: Graph Neural Network-based trust management for IoT anomaly detection. arXiv:2203.56789 (2022), https://arxiv.org/abs/2203.56789
3. Axelrod, R.: Effective choice in the prisoner's dilemma. Journal of Conflict Resolution **24**(1), 3–25 (1980)
4. Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984)
5. Castelfranchi, C., Falcone, R.: Trust and control: A dialectic link. Applied Artificial Intelligence **14**(8), 799–823 (2000)
6. Chen, Z., et al.: Deep learning to interpret autism spectrum disorder behind the camera. arXiv:2105.12345 (2021), https://arxiv.org/abs/2105.12345
7. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T.: Cooperative AI: Machines must learn to find common ground. Nature **593**, 33–36 (2021). https://doi.org/10.1038/s41586-021-03595-2
8. Eriksen, K.T., Bodenhagen, L.: Understanding human-robot teamwork in the wild: The difference between success and failure for mobile robots in hospitals. In: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2023). pp. 277–284. IEEE (2023). https://doi.org/10.1109/RO-MAN57019.2023.10309638
9. Esterbauer, R., Kubicek, B.: Relational embodiment and intentionality in trust dynamics. Humanities and Social Sciences Communications **11**(1420), 1–8 (2024)
10. Ferrante, E., Turgut, A.E., Duéñez-Guzmán, E., Dorigo, M., Wenseleers, T.: Evolution of self-organized task specialization in robot swarms. PLoS Computational Biology **11**(8), e1004273 (2015). https://doi.org/10.1371/journal.pcbi.1004273
11. Hernandez, E., Wunsch, D.: Graphical trust models for agent-based systems. IEEE Potentials **37**(5), 25–33 (2018). https://doi.org/10.1109/MPOT.2018.2823860
12. Hoffman, D.D., Prakash, C.: Objects of consciousness. Frontiers in Psychology **5**, 82279 (2014)
13. Hou, C., Pinter-Wollman, N., et al.: Costs of task allocation with local feedback: Effects of colony size and extra workers in social insects and multi-agent systems. PLoS Computational Biology **14**(9), e1006275 (2018). https://doi.org/10.1371/journal.pcbi.1006275
14. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision support systems **43**(2), 618–644 (2007)
15. Marsh, S.P.: Formalising trust as a computational concept. University of Stirling, Stirling, Scotland (1994)
16. McKnight, D.H., Carter, M., Thatcher, J.B., Clay, P.F.: Trust in a specific technology: An investigation of its components and measures. ACM Transactions on Management Information Systems **2**(2), 12:1–12:25 (2011). https://doi.org/10.1145/1985347.1985353
17. Nowak, M.A.: Five rules for the evolution of cooperation. Science **314**, 1560–1563 (2006). https://doi.org/10.1126/science.1133755
18. Pippin, B., Christensen, H.: Perfpatrol: Performance monitoring of mobile robot teams. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 4459–4466. IEEE (2014). https://doi.org/10.1109/ICRA.2014.6907645
19. Pippin, C., Christensen, H.: Trust modeling in multi-robot patrolling. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 59–66. IEEE (2014). https://doi.org/10.1109/ICRA.2014.6907506

20. Press, W.H., Dyson, F.J.: Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. Proceedings of the National Academy of Sciences **109**(26), 10409–10413 (2012). https://doi.org/10.1073/pnas.1206569109
21. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson, Upper Saddle River, NJ (2016)
22. Schäfer, A., Esterbauer, R., Kubicek, B.: Trusting robots: a relational trust definition based on human intentionality. Humanities and Social Sciences Communications **11**(1412), 1–10 (2024)
23. Sylvester, A., Gini, M.: Detecting adversarial interference in cooperative tasks through multi-dimensional Bayesian-informed trust metric. In: Intelligent Autonomous Systems 19 (IAS-19). Springer Nature, The Campus, 4 Crinan Street, London, N1 9XW (2025)
24. Teacy, W.T.L., Luck, M., Rogers, A., Jennings, N.R.: An efficient and versatile approach to trust and reputation using hierarchical Bayesian modelling. Artificial Intelligence **193**, 149–185 (2012)
25. Teacy, W.T.L., Patel, J., Jennings, N.R., Luck, M.: Travos: Trust and reputation in the context of inaccurate information sources. Autonomous Agents and Multi-Agent Systems **12**, 183–198 (2006). https://doi.org/10.1007/s10458-006-5952-x