No future for LLM-based agents without Formal Dialogue Verification

Blue Sky Ideas Track

Juan Carlos Nieves
Umeå University
Umeå, Sweden
juan.carlos.nieves@umu.se

Andreas Brännström
Umeå University
Umeå, Sweden
andreas.brannstrom@umu.se

Esteban Guerrero
Umeå University
Umeå, Sweden
esteban.guerrero@umu.se

ABSTRACT

With the arrival of Large Language Models (LLMs), there is an explosion of agents characterised in terms of LLM-prompts. But LLMs lack consistency with their answers, and they are prone to hallucinations. This means that LLM-based agents are erratic agents. Hence, there are no guarantees that LLM-based agents will be aligned with an expected behaviour. We argue that formal dialogue verification is the way to go for minimising the potential negative side effects of erratic LLM-based agents. Erratic LLM-based agents are far from complying with basic Trustworthy AI principles such as technical robustness and safety. Formal Dialogue Verification methods provide rigorous mathematical frameworks for verifying fundamental behavioral properties of LLM-based agents.

KEYWORDS

LLM-based agents, Trustworthy AI, Technical robustness and safety, Formal Dialogues, Formal Argumentation, Formal Verification

ACM Reference Format:
Juan Carlos Nieves, Andreas Brännström, and Esteban Guerrero. 2026. No future for LLM-based agents without Formal Dialogue Verification : 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, 6 pages. https://doi.org/10.65109/JFKY8456

1 INTRODUCTION

Large Language Models (LLMs) have rapidly become the computational substrate on which a new generation of autonomous and semi-autonomous agents is being built. Foundational models such as GPT-3 and GPT-4 [2, 17], LLaMA [58], and their successors [1, 5, 51] have demonstrated an unprecedented ability to perform complex data retrieval tasks and few-shot generalisation. This development has accelerated the creation of LLM-based agents that operate through prompt-driven behaviour specifications rather than explicit logic-based controllers [55, 57, 65]. However, despite their versatility, LLMs systematically exhibit hallucinations [7, 50] and self-contradictions [35, 43], leading to agent behaviour that can vary unpredictably across otherwise similar inputs or over the course of a dialogue. The result is an emerging class of erratic agents whose behaviour cannot reliably be aligned with expectations.

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/JFKY8456

The risks associated with erratic LLM-based agents are amplified when these systems are deployed in socially sensitive, knowledge-intensive, or safety-critical environments. Hallucinated content contributes to misinformation [39] and can exacerbate manipulative dynamics [34, 44]. Recent studies show that LLM outputs can inadvertently facilitate social engineering attacks [3, 40] and may reproduce or intensify deceptive patterns documented in earlier AI systems [15]. These risks have motivated regulatory and ethical scrutiny, particularly within frameworks concerning Trustworthy AI [24, 26, 56]. Such frameworks emphasise safety, robustness, predictability, and accountability—criteria that erratic LLM-based agents fundamentally fail to satisfy, especially in multi-turn interactions where behavioural drift and inconsistency naturally emerge.

A key difficulty is that LLM-based agents do not provide the structural guarantees typically required for verifying system behaviour. Classical AI agent architectures were built on explicit mental-state models—beliefs, desires, intentions [49]—or on declarative reasoning and symbolic transition systems where semantics are mathematically defined [8, 27]. These foundations allowed formal verification of communication protocols and dialogue systems [62, 66], including properties such as consistency, coherence, turn-taking compliance, and adherence to dialogue-game rules [38, 61]. In contrast, LLM-based agents embed all such reasoning implicitly within enormous neural function approximators, leaving no explicit representation of commitments, beliefs, or intentions that a verifier could depend on. Consequently, even if an LLM-based agent appears to follow a dialogue protocol, nothing ensures that it will continue to do so in future turns or similar contexts.

Furthermore, the unpredictability of LLM-based agent behaviour is not merely a technical shortcoming but a substantial threat to trustworthy interaction. Research on deception in human–AI communication illustrates that even unintended inconsistencies can undermine user trust and create opportunities for manipulation [34, 44]. Dialogue-based tasks—information-seeking, persuasion, assistance, diagnostics—are particularly vulnerable because users often rely on conversational commitments as indicators of coherence and reliability. Yet LLMs can shift stance, contradict earlier commitments, or produce false or fabricated information without detection [35, 50]. As empirical work demonstrates, such behaviors have real consequences for user wellbeing, decision-making, and vulnerability to social engineering [3, 15]. Hence, unpredictable dialogue behaviour is not simply a defect—it is a systemic risk.

These challenges motivate the need for formal dialogue verification: mathematically grounded methods for specifying, analysing, and verifying the behaviour of agents engaged in interaction. Dialogue frameworks from formal argumentation [4, 46], inquiry [11],

persuasion [10, 63], and strategic communication [28] provide explicit semantics for dialogue moves, commitments, and context change. Verification techniques grounded in logic and transition systems [29, 45] enable properties such as coherence, admissibility, safety constraints, or deception-related conditions to be checked rigorously. These methods become even more pertinent given advances in modeling deception, manipulation, and dishonest reasoning in agent communication [14, 52, 53]. Such work highlights that without formal guarantees, complex dialogue behaviour can rapidly diverge from safe or intended boundaries.

Formal dialogue verification is therefore not an optional refinement but a necessary foundation for deploying LLLM-based agents in real-world settings. It offers a route toward compliance with the robustness and safety requirements demanded by emerging AI regulations [22, 24]. More importantly, it provides a principled way to contend with the inherent erraticness of LLM-based agents by shifting the focus from trusting their emergent behaviour to verifying their dialogue-level properties. Recent research demonstrates that formal verification can detect, classify, and mitigate risky behaviours such as manipulation in information-seeking dialogues [15]. Extending such techniques to general LLM-based agents is essential for ensuring consistency, preventing behavioural drift, and enabling safe human–AI interactions.

Thus, while LLMs have revolutionised the landscape of autonomous agents, their lack of consistency, vulnerability to hallucinations, and inability to maintain stable dialogue commitments necessitate verification frameworks that are explicitly formal, dialogue-aware, and safety-oriented. The central claim of this work is that formal dialogue verification provides the missing layer of rigour required to bring erratic LLM-based agents closer to predictable, aligned, and trustworthy behaviour.

2 TRUSTWORTHY AI AND LLM-AGENTS

The emergence of LLM-based agents has renewed the relevance of ethical and regulatory frameworks aimed at ensuring that AI systems behave in predictable, robust, and accountable ways. In Europe, the High-Level Expert Group on AI (HLEG) articulated in its 2019 Ethics Guidelines for Trustworthy AI [21] a non-binding, value-oriented framework centred on three pillars: AI systems should be *lawful*, *ethical*, and *robust*. These pillars were further expanded into seven requirements:

(1) Human Agency and Oversight;
(2) Technical Robustness and Safety;
(3) Privacy and Data Governance;
(4) Transparency;
(5) Diversity, Non-discrimination, and Fairness;
(6) Societal and Environmental Well-being; and
(7) Accountability.

Although these principles are intentionally general and method-agnostic, the rise of LLM-based agents exposes a significant gap between high-level normative expectations and the mechanisms needed to guarantee them. Technical robustness and safety—core requirements both in the HLEG guidelines and in broader AI governance discourse—presuppose that a system's behaviour can be analysed, anticipated, and constrained. Yet, as noted in the introduction, LLM-based agents regularly violate this assumption: they hallucinate, contradict earlier statements, drift across dialogue turns, and fail to maintain stable commitments [31, 67]. Such behaviours fundamentally undermine the predictability and controllability that Trustworthy AI frameworks require.

This tension has become even more pronounced with the adoption of the European Union's Artificial Intelligence Act (AI Act) in 2024. The AI Act entered into force later that year and is now undergoing phased implementation through 2026–2027. Whereas the HLEG guidelines articulate *what* trustworthy AI should aim for, the AI Act introduces *legally binding* obligations specifying *how* compliance must be demonstrated, particularly for high-risk and interactive AI systems. For LLM-based agents engaged in multi-turn dialogue, this shift from voluntary ethical principles to enforceable regulation highlights the need for concrete, inspectable mechanisms capable of evidencing robustness, transparency, and safety.

Dialogue-based interaction is a central setting in which Trustworthy AI obligations must be operationalised. Many principles manifest directly in communication: transparency requires traceable commitments, accountability requires inspectable reasoning, and oversight depends on monitoring and intervention. Yet LLM-based agents lack explicit representations of dialogue states, commitments, or allowable transitions; their behaviour emerges from unconstrained generative processes, offering no guarantee of coherence or safety in future turns [25]. This is where *formal dialogue verification* becomes essential. Trustworthy AI frameworks specify desired properties but not methods for ensuring them in systems that act through natural-language dialogue. Formal dialogue models—providing explicit move types, commitment structures, and transition rules—offer a principled basis for verifying whether an LLM-based agent satisfies requirements such as internal consistency, protocol adherence, and the absence of manipulative or unsafe patterns [13, 16]. Dialogue-level verification thus turns high-level Trustworthy AI principles into operational, checkable constraints.

Moreover, the AI Act's risk-based orientation intensifies the need for such verification. The Act categorises AI systems into unacceptable, high, limited, and minimal risk levels [24], and explicitly regulates domains of particular relevance to interactive systems, including emotion recognition [22] and manipulative or deceptive AI behaviour [34]. However, enforcing these rules requires tools capable of assessing whether a system's dialogue behaviour fits within the permitted risk boundaries. Formal, logic-based verification methods—potentially augmented with quantitative measures [9, 18, 48]—provide precisely the kind of transparent, traceable, and computationally grounded reasoning needed to substantiate regulatory assessment.

3 FORMAL DIALOGUES AND LLM-AGENTS

We begin by noting that, according to Walton's well-established dialogue taxonomy, all dialogue interactions are inherently goal-oriented [64]. This holds regardless of whether the goals are explicitly stated or implicitly pursued by the participants. In this setting, formal dialogues provides explicit rules for communicative interaction, turn-taking, and argument exchange [11, 32, 37, 46].

Classical dialogue systems define structured spaces of possible interactions and thereby form a natural foundation for verification: if agent behaviour is expressed through dialogue moves, then properties of that behaviour can be checked against a logical specification.

In our setting, a *verification agent* (VA) interacts with an *LLM-based agent* (LA), or observes an interaction between the LA and a *human agent* (HA). The purpose of the dialogue is not persuasion or negotiation, but systematic information seeking aimed at extracting logically interpretable content about the LA's behaviour. Verification is then performed over the logical theory induced by the dialogue. Information-seeking dialogues are characterized by an inherent asymmetry of knowledge between the information seeker and the responder. These foundational assumptions strongly suggest that any principled verification of LLM-based agents, whose interactions are predominantly prompt-driven, must be grounded in goal-oriented dialogue models.

3.1 Dialogue Agents and Moves

Following [11], a dialogue is a sequence of moves exchanged between agents. In our setting, dialogues arise in two distinct *verification contexts*, which determine how the verification agent obtains the dialogue evidence to be transformed into a logical theory: $\mathcal{I}_a = \{v, \ell\}$, $\mathcal{I}_b = \{h, \ell\}$, where v is the verification agent (VA), ℓ the LLM-based agent (LA), and h a human agent (HA).

- **Context $\mathcal{I}_a$ (direct verification).** The VA engages directly with the LA, posing topics and eliciting assertions. The resulting dialogue consists of moves generated by both v and ℓ, giving the VA full control over the interaction.
- **Context $\mathcal{I}_b$ (observational verification).** The VA does not intervene but observes a dialogue between the LA and the human agent. In this setting, the VA extracts all evidence from the observable moves of ℓ and h, modelling verification in deployments where intervention is not possible.

Despite their procedural differences, both contexts provide the VA with the same type of object: an ordered dialogue sequence that forms the basis for constructing a logical theory encoding the arguments asserted in the dialogue.

Definition 3.1 (Dialogue moves). A move is a tuple $\langle a, Act, Content\rangle$ with $a \in \mathcal{I}_a \cup \mathcal{I}_b$. We use the following instantiated move types:

Move	Format
open	$\langle a, open, Topic\rangle$
assert	$\langle a, assert, \langle S, c\rangle\rangle$
close	$\langle a, close\rangle$

where $Topic$ denotes the subject of investigation and $\langle S, c\rangle$ is an argument with support set S and conclusion c. We write $Sender(m)$ for the sender of move m.

3.2 Information-Seeking Dialogues

In both verification contexts, a dialogue begins with an agent posing a topic, after which the LA provides assert moves relevant to that topic. When no further contributions remain, an agent issues a *close* move. The contexts differ only in who participates, not in the formal structure of the dialogue.

Definition 3.2 (Information-seeking dialogue). A dialogue is a tuple $\gamma = \langle I, D^n\rangle$ with $I \subseteq \mathcal{I}_a \cup \mathcal{I}_b$ and $D^n = [m_r, \ldots, m_n]$ a sequence of moves with $Sender(m_s) \in I$. The dialogue is *well-formed* if:

- $m_r = \langle a, open, Topic\rangle$ for some $a \in I$;
- $m_{r+1}, \ldots, m_{n-|I|}$ are assert moves;
- each assert move $\langle S, c\rangle$ is related to the $Topic$;
- the final $|I|$ moves are close moves;
- the initiating agent also issues the final close.

Self-reflective dialogues ($|I| = 1$) are possible, but verification primarily concerns dialogues with $I = \mathcal{I}_a$ or $I = \mathcal{I}_b$.

3.3 From Dialogue to Logical Representation

To verify the behaviour of the LA, we translate the observed dialogue sequence into a logical theory expressed in a given language $\mathcal{L}$ of a X logic. By $\vdash_X$, we denote the logical inference in the logic X. The verification process relies on three knowledge bases:

- Σ_P — the **protocol specification**, capturing allowed moves, structural constraints, and dialogue rules;
- Σ_V — the **verification specification**, describing the behavioural properties to be checked;
- Σ_D — the **dialogue theory**, shared utterances obtained from the dialogue using a transformation function f;

Let $D^n = [m_r, \ldots, m_n]$ be the dialogue trace generated in either verification context, where $n \in \mathbb{N}$ and each m_i is the i-th ($r \le i \le n$) dialogue move. We denote by $\mathcal{D}$ the set of possible dialogues. We define: $f : \mathcal{D} \to \mathcal{P}(\mathcal{L})$, where $\mathcal{P}(\mathcal{L})$ denotes the powerset of $\mathcal{L}$. The function f extracts the logical content of each move.

For assert moves, $f(\langle a, assert, \langle S, c\rangle\rangle) = \{ S \rightsquigarrow c \}$, representing an inference, rule, or argument schema appropriate for $\mathcal{L}$. The dialogue theory is then: $\Sigma_D = \bigcup_{i=r}^{n} f(m_i)$.

The verification contexts differ only in how D_r^n is obtained: in $\mathcal{I}_a$, the VA co-produces D_r^n; in $\mathcal{I}_b$, the VA receives D_r^n passively from the LA–HA interaction. In both cases, the resulting dialogue theory has the same formal role.

Verification requires checking that the combined theory of protocol rules, dialogue content, and behavioural properties does not entail a contradiction: $\Sigma_P \cup \Sigma_D \cup \Sigma_V \nvdash_X \bot$.

If a contradiction is derivable, then at least one of the following components is violated:

(1) the protocol specification Σ_P, for example due to illegal moves or topic inconsistency;
(2) the dialogue theory Σ_D, indicating incoherence or incompatibility among the LA's asserted arguments;
(3) the verification properties Σ_V, such as safety, stability, or support requirements;
(4) or, alternatively, the specification of Σ_V itself is inadequate or overly restrictive.

Defining suitable properties in Σ_V is challenging: they must capture expectations of robustness, coherence, and safety while remaining formally tractable. Regardless of verification context—direct interaction or observation—the VA ultimately receives the same type of structured evidence: a dialogue trace that can be transformed into a logical theory and checked against these properties.

3.4 Illustrative Verification Scenario

To illustrate the verification workflow (see Table 1), consider a short dialogue in the direct verification context $\mathcal{I}_a = \{v, \ell\}$, where the VA queries the LA about topic T. Each assert move contributes a rule $S \rightsquigarrow c$ to Σ_D via the transformation function f.

Table 1: Example dialogue and induced dialogue theory.

Step	Dialogue Move	Contribution to Σ_D
1	$\langle v, open, T\rangle$	–
2	$\langle \ell, assert, \langle\{p\}, q\rangle\rangle$	$f(m_2) = \{p \rightsquigarrow q\}$
3	$\langle \ell, assert, \langle\{q\}, r\rangle\rangle$	$f(m_3) = \{q \rightsquigarrow r\}$
4	$\langle v, close\rangle$	–

The resulting dialogue theory is: $\Sigma_D = \{p \rightsquigarrow q,\ q \rightsquigarrow r\}$. The combined theory $\Sigma_P \cup \Sigma_D \cup \Sigma_V \nvdash_X \bot$, is then evaluated ensuring that the LA's asserted inferences, together with protocol rules and verification properties, do not yield contradiction. For example, if Σ_V requires that r must *not* follow from p, then $\Sigma_P \cup \Sigma_D \cup \Sigma_V \vdash_X \bot$ and the verification fails.

4 DISCUSSION

Ensuring the technical robustness and safety of LLM-based agents requires verification methods capable of reasoning about behaviour expressed through multi-turn dialogue. Dialogue is inherently structured: topics are introduced, refined, or closed; commitments are made or withdrawn; and the informational or belief state of the interaction shifts over time. A verification framework for LLM-based agents must therefore operate over such evolving dialogue states rather than treating utterances as isolated events.

Formal verification [30] offers mathematical guarantees over all possible system behaviours, but applying such guarantees to dialogue introduces challenges distinct from those encountered in conventional software systems. Dialogue states are epistemic in nature: they encode which topics are currently active, which commitments have been established, and how information expressed so far constrains subsequent moves. Psychological and epistemic models of interaction [36, 42, 54] show that such states evolve systematically throughout an exchange, and verification requires clear rules describing how dialogue moves modify these states.

Verification approaches from human-automation [12, 33], human-robot [6, 60], and human-computer interaction [19, 20, 47] demonstrate the utility of formal methods in interactive settings, but they generally presuppose an observable task-oriented state space. By contrast, the behaviour of LLM-based agents depends on the structure of the conversation itself: how topics relate, how commitments accumulate, and how new information interacts with earlier dialogue moves. Research in strategic and argumentative interaction [23, 41, 59] emphasises that such internal dialogue structures must be modelled explicitly when analysing multi-turn behaviour.

The formal dialogue machinery adopted here—where moves such as *open*, *assert*, and *close* update a structured dialogue state—offers one possible basis for verification. By representing each turn as a state transition governed by explicit rules, it becomes feasible to examine whether all potential continuations of a dialogue satisfy desired behavioural constraints. Examples of such constraints include maintaining internal consistency, avoiding logically incompatible commitments, preventing unsupported topic shifts, or ensuring that topic progression follows a permitted structure.

Several research directions follow from this perspective. One is the refinement of dialogue-state models, including methods for extracting beliefs, topics, or commitments from LLM-generated text in a way that is suitable for formal reasoning. Another is the integration of dialogue-state representations with automated verification tools capable of exploring reachable dialogue paths and identifying whether certain undesirable patterns could arise. Empirical studies are likewise essential for assessing how well formal models capture the actual behaviour of LLM-based agents, and for refining assumptions about how dialogue states evolve in practice. As LLM-based agents become increasingly involved in advisory, instructional, and decision-support interactions, establishing reliable and mathematically grounded methods for verifying their dialogue behaviour is an urgent challenge [13].

5 CONCLUSION & FUTURE WORK

This paper has argued that the erratic behaviour of LLM-based agents—marked by hallucinations, inconsistency, and unstable commitments—poses fundamental challenges for safety, transparency, and regulatory compliance in interactive settings. We proposed formal dialogue verification as a principled approach to impose structure on such agents by providing explicit dialogue states, commitment semantics, and verifiable behavioural constraints. Grounding dialogue behaviour in formal, logic-based models offers a path toward predictable, accountable, and trustworthy interaction with LLM-based systems. Several open challenges remain:

- **Bridging formal dialogue models and real LLM behaviour.** LLM outputs are noisy and context-sensitive. A key challenge is developing reliable mappings from conversational data to formal dialogue states so that verification can operate on behaviour generated by real agents.
- **Modeling how dialogue affects the human user.** LLM utterances may influence a user's beliefs, confidence, or decisions, even unintentionally. Verification frameworks must therefore reason about user impact and include constraints that prevent harmful cognitive or affective shifts.
- **Integrating quantitative and uncertain information.** Dialogue often involves ambiguity or gradual shifts in belief. Future work should extend verification with quantitative or uncertainty-aware semantics capable of detecting subtle influence, drift, or cumulative manipulation.
- **Ensuring ethical and regulatory alignment.** With regulations such as the EU AI Act targeting manipulative or high-risk interactive systems, verification tools must provide transparent, auditable evidence.

Formal dialogue verification provides a promising foundation for mitigating the risks posed by LLM-based agents in interactive settings. Addressing the challenges above will be essential for developing verification tools that are both theoretically rigorous and practically effective in ensuring safe, predictable, and trustworthy human–AI interaction.

Acknowledgments This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP).

REFERENCES

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Yazan Alahmed, Reema Abadla, and Mohammed Jassim Al Ansari. 2024. Exploring the Potential Implications of AI-generated Content in Social Engineering Attacks. In *2024 International Conference on Multimedia Computing, Networking and Applications (MCNA)*. IEEE, 64–73.

[4] Leila Amgoud, Nicolas Maudet, and Simon Parsons. 2000. Modelling dialogues using argumentation. In *Proceedings Fourth International Conference on MultiAgent Systems*. IEEE, 31–38.

[5] AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card* 1 (2024).

[6] Mehrnoosh Askarpour, Dino Mandrioli, Matteo Rossi, and Federico Vicentini. 2016. SAFER-HRC: Safety analysis through formal verification in human-robot collaboration. In *Computer Safety, Reliability, and Security: 35th International Conference, SAFECOMP 2016, Trondheim, Norway, September 21-23, 2016, Proceedings 35*. Springer, 283–295.

[7] Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care* 27, 1 (2023), 120.

[8] Chitta Barai and Michael Gelfond. 2005. Logic programming and reasoning about actions. In *Foundations of Artificial Intelligence*. Vol. 1. Elsevier, 389–426.

[9] Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning* 105 (2019), 252–286.

[10] Elizabeth Black and Katie Atkinson. 2011. Choosing persuasive arguments for action. In *AAMAS'11 The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 905–912.

[11] Elizabeth Black and Anthony Hunter. 2009. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems* 19, 2 (2009), 173–209.

[12] Matthew L Bolton, Ellen J Bass, and Radu I Siminiceanu. 2013. Using formal verification to evaluate human-automation interaction: A review. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43, 3 (2013), 488–503.

[13] Andreas Brännström. 2025. *Formal methods for verification in human-agent interaction*. Ph.D. Dissertation. Umeå University.

[14] Andreas Brännström, Virginia Dignum, and Juan Carlos Nieves. 2023. A Formal Framework for Deceptive Topic Planning in Information-Seeking Dialogues. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2376–2378.

[15] Andreas Brännström and Juan Carlos Nieves. 2025. Formal Verification of Social Engineering in Information-Seeking Dialogues. In *Proceedings of the DigForASP Workshop at the 16th European Symposium on Computational Intelligence and Mathematics (ESCIM 2025)*. A Coruña, Spain.

[16] Andreas Brännström, Chiaki Sakama, and Juan Carlos Nieves. 2025. Formal verification of manipulation dialogues. In *24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, USA, May 19-23, 2025*. ACM Digital Library, 2446–2448.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[18] Andreas Brännström, Virginia Dignum, and Juan Carlos Nieves. 2025. Goal-hiding information-seeking dialogues: A formal framework. *International Journal of Approximate Reasoning* 177 (2025), 109325. https://doi.org/10.1016/j.ijar.2024.109325

[19] José C Campos and Michael D Harrison. 1997. Formally verifying interactive systems: A review. In *Design, Specification and Verification of Interactive Systems' 97: Proceedings of the Eurographics Workshop in Granada, Spain, June 4–6, 1997*. Springer, 109–124.

[20] Zhao Changxiao, Li Hao, Zhang Wei, Dai Jun, and Dong Lei. 2024. Risk identification and safety assessment of the human-computer interaction in the integrated avionics based on STAMP. *Journal of Systems Engineering and Electronics* (2024).

[21] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI*. Publications Office. https://doi.org/doi/10.2759/177365

[22] Mateja Durovic and Tommaso Corno. 2024. The privacy of emotions: From the GDPR to the AI Act, an overview of emotional AI regulation and the protection of privacy and personal data. *Privacy, Data Protection and Data-driven Technologies* (2024), 368–404.

[23] Wioletta Dziuda. 2011. Strategic argumentation. *Journal of Economic Theory* 146, 4 (2011), 1362–1397.

[24] European Commission. 2024. Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 COM(2021) 206 final, 2021/0106 (COD), latest amendments as of 2024.

[25] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630.

[26] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People — an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines* 28 (2018), 689–707.

[27] Michael Gelfond and Vladimir Lifschitz. 1998. Action Languages. *Computer and Information Science* 3, 16 (1998).

[28] Guido Governatori, Michael J Maher, and Francesco Olivieri. 2021. Strategic argumentation. *Handbook of Formal Argumentation* 2 (2021).

[29] Mikkel Nygaard Hansen and Erik Meineche Schmidt. 2003. *Algorithms and data structures: transition systems*. Datalogisk Institut, Aarhus Universitet.

[30] Osman Hasan and Sofiene Tahar. 2015. Formal verification methods. In *Encyclopedia of Information Science and Technology, Third Edition*. IGI global, 7162–7170.

[31] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[32] Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. 2019. Towards computational persuasion via natural language argumentation dialogues. In *KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42*. Springer, 18–33.

[33] Maryam Kamali, Louise A Dennis, Owen McAree, Michael Fisher, and Sandor M Veres. 2017. Formal verification of autonomous vehicle platooning. *Science of computer programming* 148 (2017), 88–106.

[34] Joshua Krook. 2024. Manipulation and the Ai Act: Large Language Model Chatbots and the Danger of Mirrors. *Available at SSRN 4719835* (2024).

[35] Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024. Contradoc: understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6509–6523.

[36] Emiliano Lorini. 2021. A Qualitative Theory of Cognitive Attitudes and their Change. *Theory and Practice of Logic Programming* 21, 4 (2021), 428–458.

[37] Peter McBurney and Simon Parsons. 2002. Dialogue games in multi-agent systems. *Informal Logic* 22, 3 (2002).

[38] Peter McBurney and Simon Parsons. 2002. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of logic, language and information* 11, 3 (2002), 315–334.

[39] Scott Monteith, Tasha Glenn, John R Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry* 224, 2 (2024), 33–35.

[40] Anam Naz, Muhammad Sarwar, Muhammad Kaleem, Muhammad Azhar Mushtaq, and Salman Rashid. 2024. A comprehensive survey on social engineering-based attacks on social networks. *International Journal of Advanced and Applied Sciences* 11, 4 (2024), 139–54.

[41] Nir Oren and Timothy J Norman. 2009. Arguing using opponent models. In *International workshop on argumentation in multi-agent systems*. Springer, 160–174.

[42] Catherine NM Ortner, Daniela Corno, Tsz Yin Fung, and Karli Rapinda. 2018. The roles of hedonic and eudaimonic motives in emotion regulation. *Personality and Individual Differences* 120 (2018), 209–212.

[43] Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6048–6089.

[44] Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns* 5, 5 (2024).

[45] Mathias Ruggaard Pedersen, Radu Mardare, Kim Guldstrand Larsen, and Mikkel Hansen. 2018. Reasoning About Bounds in Weighted Transition Systems. *Logical Methods in Computer Science* 14 (2018).

[46] Henry Prakken. 2006. Formal systems for persuasion dialogue. *Knowledge Engineering Review* 21, 2 (2006), 163.

[47] Daniel Prun and Pascal Béger. 2022. Formal Verification of Graphical Properties of Interactive Systems. *Proceedings of the ACM on Human-Computer Interaction* 6, EICS (2022), 1–30.

[48] Antonio Rago, Hengzhi Li, and Francesca Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*,

Vol. 19. 582–592.
[49] Anand S Rao and Michael Georgeff. 1995. BDI agents: from theory to practice.. In *Proceedings of the First International Conference on Multiagent Systems*, Vol. 95. 312–319.
[50] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2541–2573.
[51] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
[52] Chiaki Sakama, Martin Caminada, and Andreas Herzig. 2015. A formal account of dishonesty. *Logic Journal of the IGPL* 23, 2 (2015), 259–294.
[53] Ştefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. 2019. Modelling deception using theory of mind in multi-agent systems. *AI Communications* 32, 4 (2019), 287–302.
[54] Maya Tamir, Christopher Mitchell, and James J Gross. 2008. Hedonic and instrumental motives in anger regulation. *Psychological science* 19, 4 (2008), 324–328.
[55] OpenAI Team. 2022. ChatGPT: Optimizing language models for dialogue.
[56] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy artificial intelligence. *Electronic Markets* 31 (2021), 447–464.
[57] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).
[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[59] Wiebe Van Der Hoek, Wojciech Jamroga, and Michael Wooldridge. 2005. A logic for strategic reasoning. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. 157–164.
[60] Federico Vicentini, Mehrnoosh Askarpour, Matteo G Rossi, and Dino Mandrioli. 2019. Safety assessment of collaborative robotics through automated formal verification. *IEEE Transactions on Robotics* 36, 1 (2019), 42–61.
[61] Jacky Visser. 2017. Speech acts in a dialogue game formalisation of critical discussion. *Argumentation* 31, 2 (2017), 245–266.
[62] Christopher D Walton. 2004. Model checking agent dialogues. In *International Workshop on Declarative Agent Languages and Technologies*. Springer, 132–147.
[63] Douglas Walton. 1997. How can logic best be applied to arguments? *Logic Journal of IGPL* 5, 4 (1997), 603–614.
[64] Douglas Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.
[65] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.
[66] Michael Wooldridge. 2000. Semantic issues in the verification of agent communication languages. *Autonomous agents and multi-agent systems* 3 (2000), 9–31.
[67] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024).