A Trust Extension of L-DINF for Explainable Multi-Agent Decision Making

Stefania Costantini[1,†], Valentina Pitoni[1,*,†]

[1]Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila,Italy

Abstract

This paper presents a trust-aware extension of L-DINF, an epistemic logic for modelling cognitive and cooperative agents. The extension treats trust not merely as an external score, but as an explicit semantic and normative condition regulating agency. It introduces a trust predicate and an ordered trust function, allowing trust to constrain intention adoption, action feasibility, preference-guided choice, and delegation. Differentiated trust thresholds for autonomous execution and lending, together with a three-valued decision policy, capture the intermediate region between unrestricted autonomy and prohibition. The resulting framework strengthens L-DINF as a modular and explainable formalism for trust-sensitive decision making in open multi-agent environments.

Keywords

Epistemic Logic, Trust, Multi-Agent Systems, Modal Logic

1. Introduction

The epistemic logic L-DINF [1, 2, 3, 4] has been introduced to model cognitive agents and Multi-Agent Systems (MAS). It is able to capture interaction among agents and group dynamics. It has been usefully applied to complement ASP (Answer Set Programming [5, 6, 7]) applications [8, 9] to provide user-tailored personalized behaviour. A prototype implementation of the logic already exists [10]. Since open environments involve potentially heterogeneous agents with varying capabilities and reliability, any formalism that models their interactions comprehensively must explicitly account for trust, which has not been accounted for in L-DINF nor in related work.

Trust is, in fact, a central requirement in autonomous and cooperative multi-agent systems, especially in domains where agents must act under uncertainty, coordinate with others, and make decisions whose consequences may be significant. In such settings, trust should not be treated merely as an informal notion or as an external score with no direct impact on reasoning. Rather, it should be represented as a principled component of deliberation, capable of affecting whether an agent may adopt an intention, execute an action, or transfer responsibility through delegation. This is particularly important in environments where autonomy must remain explainable, regulated, and sensitive to borderline cases that cannot be adequately captured by a simple binary distinction between permission and prohibition.

L-DINF provides a suitable formal basis for addressing this problem. As an epistemic logic for modelling cognitive and cooperative agents, L-DINF supports explicit reasoning about beliefs, intentions, preferences, action feasibility, and group-level interaction. Its expressive power makes it possible to represent agents not merely as executors of actions, but as cognitive entities whose practical behaviour depends on structured internal states and formally specified interaction constraints. This is precisely the kind of setting in which trust can be integrated in a non-superficial way, namely as a condition that shapes the space of admissible cognitive and practical outcomes.

The motivation for extending L-DINF with trust is twofold. First, trust directly affects action regulation: an agent may be able to perform an action, and may even prefer to perform it, while still lacking the level of trust required for autonomous execution. Second, trust affects cooperation: when direct

CILC 2026: The 41st Italian Conference on Computational Logic, June 23–25, 2026, Ferrara, Italy

*Corresponding author.

†These authors contributed equally.

stefania.costantini@univaq.it (S. Costantini); valentina.pitoni@univaq.it (V. Pitoni)

0000-0002-5686-6124 (S. Costantini); 0000-0002-4245-4073 (V. Pitoni)

execution is not justified, delegation may still be a reasonable and formally admissible alternative. A trust-aware logical framework should therefore be able to distinguish among at least three different outcomes, namely full autonomous execution, delegation, and inhibition. This finer-grained perspective is essential for modelling the grey zone in which trust is neither high enough to justify direct action nor low enough to justify a complete block.

The trust extension proposed for L-DINF addresses this need by introducing trust as a first-class logical component of the framework. More specifically, trust is incorporated into the language and semantics through an explicit trust predicate and an ordered trust function, allowing trust assessments to constrain intention adoption, action feasibility, preference-guided choice, and lending conditions. In this way, trust becomes part of the semantic regulation of agency rather than a post hoc annotation attached to agents or actions. The resulting framework strengthens L-DINF as a modular and explainable logic for regulated autonomy, where decisions about acting, delegating, or blocking are grounded in formally specified trust thresholds. The proposal aims to show that trust should be treated not merely as an external score or policy-layer annotation, but as a semantic admissibility condition on agency inside a cognitively structured logical framework. More fundamentally, the proposal suggests that trust should be seen as a semantic mechanism regulating autonomy itself. In this perspective, agency becomes graded rather than binary, with delegation emerging as an explicit intermediate mode between unrestricted autonomy and complete inhibition. This perspective is also compatible with recent trends in agentic AI and trustworthy autonomous systems, where escalation, human oversight, and controlled delegation are increasingly regarded as essential mechanisms for safe autonomy.

Importantly, the framework also distinguishes between trust evaluation and policy regulation. While one component assesses how reliable an agent is considered to be, another determines the trust thresholds required for intention adoption, autonomous execution, delegation, or inhibition. This separation strengthens the explainability of the framework, since practical outcomes depend not only on an assessed trust value, but also on explicitly specified normative criteria governing admissibility. In this respect, the framework is especially suited to borderline situations in which an agent may be able and willing to act, yet not sufficiently trusted for autonomous execution, making delegation the most appropriate outcome.

Beyond its immediate practical relevance, this extension also has methodological value. In many formal approaches, trust is handled at the level of implementation, policies, or external scoring mechanisms, while the underlying logic remains unchanged. In the present proposal, by contrast, trust is internalized into the formal framework itself. This means that the distinction between acting autonomously, acting through delegation, and refraining from action is not merely the result of an implementation choice, but follows from explicit semantic conditions. As a consequence, the framework is not only suitable for regulating action, but also for explaining why a certain practical outcome is justified in a given state of the system.

A further reason for adopting L-DINF as the host logic is that its architecture already contains the core dimensions on which trust should operate. The logic combines static and dynamic reasoning, supports preference-sensitive deliberation, and includes explicit mechanisms for group organization and inter-group lending. These features make it possible to formalize trust not simply as a descriptive property, but as a normative constraint on what an agent may do, what a group may authorize, and under which conditions delegation is justified. At the same time, the proposal remains modular. In the present version, trust is introduced as a regulator of practical agency, while the dynamic treatment of mental actions is inherited unchanged from the base L-DINF framework. This keeps the extension targeted and conceptually controlled, while preserving compatibility with the original architecture and its epistemic foundations.

This paper is therefore concerned with the formalization of a trust-aware extension of L-DINF. The main objective is to show how trust can regulate cognitive and practical agency within a unified epistemic framework, preserving modularity and explainability while providing a rigorous account of regulated autonomous action in multi-agent environments. From the viewpoint of computational logic, the contribution is not just the addition of a trust predicate, but the integration of trust into the semantic conditions governing intention, feasibility, delegation, and action selection.

More specifically, the extension makes it possible to represent situations in which agents have the same technical capability and role-based admissibility, while receiving different admissible outcomes because they are assigned different trust levels. Thus, an agent may be explicitly represented as capable of performing an action but not sufficiently trusted to execute it autonomously, so that delegation or inhibition becomes the justified outcome. This trust-sensitive distinction is not explicitly represented in the original L-DINF vocabulary, where practical feasibility and cooperation are not regulated by internal trust thresholds.

The proposal is assessed at two complementary levels: meta-theoretically, by identifying the fragment inherited unchanged from the base framework, analysing the coherence of the trust-dependent decision policy, and exhibiting a trust-sensitive outcome discrimination property; and operationally, through an illustrative cybersecurity scenario.

The paper is organized as follows: Section 2 reviews related work; Section 3 presents the trust-aware extension of L-DINF; Section 4 provides a formal validation and an illustrative example; and Section 5 summarizes our findings and outlines future research directions.

2. Related Work

A first relevant line of research is represented by classical BDI approaches. The foundational work of Rao and Georgeff established the importance of modelling rational agents in terms of beliefs, desires, and intentions [11], while subsequent developments such as AgentSpeak/Jason provided an operational framework for programming BDI agents [12]. These approaches remain central to the study of intentional agency, but they do not natively provide a trust-sensitive semantics in which trust explicitly constrains intention adoption, action feasibility, or delegation. In this respect, the present proposal is not intended to replace BDI-style reasoning, but rather to complement it with a richer formal treatment of regulated autonomy.

A second important line of work concerns the dynamic epistemic foundations on which L-DINF builds. In particular, the studies by Balbiani, Fernández-Duque, and Lorini on belief dynamics for resource-bounded agents and on the dynamics of epistemic attitudes, together with Lorini's work on reasoning about cognitive attitudes in a qualitative setting, provide a formal basis for representing structured cognitive change in resource-sensitive settings [13, 14, 15]. These contributions are especially relevant because they move beyond static informational models and treat cognition as a dynamic process involving belief update and attitude revision. However, they do not directly address the role of trust as a semantic regulator of practical reasoning. The trust extension of L-DINF continues this line of research by integrating trust into the conditions governing action and delegation.

A third and more immediate line of related work is the development of L-DINF itself. The original framework introduced a formal epistemic logic for group dynamics of agents [3, 1], thereby providing a basis for modelling collective agency, structured interaction, and dynamic reorganization. This framework was later extended with preference management, strengthening its ability to represent preference-guided action selection and cognitively grounded practical reasoning [16]. These earlier contributions are essential for the present work because they already identify the main semantic dimensions on which trust can naturally operate, namely beliefs, intentions, preferences, feasibility, and cooperative lending. The current proposal departs from them by making trust explicit in the language and semantics, and by using it to regulate when autonomous execution is allowed, when delegation is preferable, and when action must be blocked.

A further relevant perspective comes from work on trust modelling and trust-aware systems. The reference ontology of trust proposed by Amaral et al. offers an important conceptual clarification of trust relations, their structure, and the entities involved in trust assessment [17]. More broadly, recent studies on trust and security mechanisms, as well as surveys on trust-based collaborative intrusion detection, show that trust has become a crucial factor in distributed and security-sensitive decision making [18, 19]. Nevertheless, these approaches are generally concerned with trust evaluation, trust management, or application-level trust policies, rather than with embedding trust directly into the

semantics of a cognitive logic. By contrast, the trust extension of L-DINF aims to formalize how trust shapes the internal space of cognitively and normatively admissible actions. None of these approaches, to our knowledge, integrates trust as a condition affecting the semantic admissibility of actions inside a cognitive logical framework with explicit beliefs, preferences, and delegation.

Finally, the cognitive orientation of L-DINF is also compatible with broader perspectives on reasoning about other agents' mental states. In particular, the possibility of modelling group-based belief sharing and cognitively structured interaction resonates with ideas commonly associated with Theory of Mind [20]. This aspect is relevant because trust is not merely a scalar assessment; it also reflects how agents evaluate one another as potential collaborators, delegates, or decision-makers. The trust-aware extension of L-DINF should therefore be understood as part of a broader effort to model socially regulated and explainable agency in a formally rigorous way.

Overall, the existing literature provides important building blocks for the present work; BDI offers a foundational model of intentional agency, dynamic epistemic logic offers tools for modelling cognitive change, prior L-DINF research provides a rich formal setting for cooperative reasoning, and trust-oriented studies clarify the conceptual and operational relevance of trust. The contribution of this paper lies in bringing these strands together within a single logical framework in which trust is treated as an explicit semantic condition on agency. More generally, the framework opens the possibility of studying socially regulated cognition, where trust affects not only physical actions but also information acceptance, belief revision, and reasoning about other agents.

3. L-DINF Framework

L-DINF is a logic composed of a static component and a dynamic component [16]. The former, called *L-INF*, is a logic of explicit beliefs and background knowledge. The latter extends the static component with dynamic operators that express the consequences of agents' mental actions. In this section, we briefly recall the main ingredients of the framework and then introduce the trust-sensitive additions.

3.1. Syntax

Let $Atm = \{p, q, \ldots\}$ be a countable set of atomic propositions. Let Atm_A be the set of physical actions that agents can perform, including active sensing actions. Let Agt be the set of agents and Grp the set of groups of agents. Moreover, let $Res = \{r_1, \ldots, r_\ell\}$ be the set of all resources.

The language of *L-DINF*, denoted by $\mathcal{L}_{\textit{L-DINF}}$, is defined by the following grammar:

$$\begin{array}{lll} \varphi, \psi & ::= & p \mid \neg\varphi \mid \varphi \wedge \psi \mid B_i\varphi \mid K_i\varphi \mid \\ & & do^P_G(\phi_A) \mid do_G(\phi_A) \mid can\_do_G(\phi_A) \mid \\ & & intend_G(\phi_A) \mid exec_G(\alpha) \mid \\ & & pref\_do_i(\phi_A, d) \mid pref\_do_G(i, \phi_A) \mid [G : \alpha]\,\varphi \mid \\ & & Cl(\phi_A, \phi'_A) \mid fCl_i(\phi_A, \phi'_A) \mid lend_G(i, H, \phi_A) \mid trust_K(i, \tau) \\ \alpha & ::= & +\varphi \mid \vdash\!(\varphi,\psi) \mid \cap(\varphi,\psi) \mid \downarrow\!(\varphi, \psi) \mid \dashv\!(\varphi, \psi) \end{array}$$

where $p \in Atm$, $\phi_A, \phi'_A \in Atm_A$, $i \in Agt$, $G, H \in Grp$, $d \in \mathbb{N}$, and τ belongs to the finite ordered set

$$Levels = \{very\_low,\ low,\ medium,\ high,\ very\_high\},$$

with

$$very\_low < low < medium < high < very\_high.$$

For simplicity, whenever $G = \{i\}$ we often write the singleton agent as a subscript in place of the singleton group. Thus, for instance, we may write $exec_i(\phi_A)$ instead of $exec_{\{i\}}(\phi_A)$.

The formula $intend_i(\phi_A)$ indicates the intention of agent i to perform the physical action ϕ_A, in the sense of the BDI agent model [11]. As in the original L-DINF framework, we do not provide a

full formalization of BDI, and therefore intentions are treated at an abstract level. We assume that $intend_G(\phi_A)$ holds whenever all agents in group G intend to perform ϕ_A.

The formulas $do_i(\phi_A)$ and $do_G(\phi_A)$ indicate, respectively, the actual execution of a physical action by an individual agent or by a group. As in the original framework, these predicates are not axiomatized directly, since they are intended to be implemented through semantic attachment to the external environment. The formula $do_i^P(\phi_A)$ records that an action has been performed in the past.

The expression $can\_do_i(\phi_A)$ represents an enabling condition for the action ϕ_A, while $pref\_do_i(\phi_A, d)$ indicates the degree d of preference or willingness of agent i to perform ϕ_A. At the group level, $pref\_do_G(i, \phi_A)$ indicates that agent i is one of the most preferred executors of ϕ_A within group G.

The formula $Cl(\phi_A, \phi'_A)$ denotes the equivalence of two physical actions in the practical context at hand. Intuitively, equivalent actions may require similar resources, achieve equivalent results, or be substitutable for one another. The formula $fCl_i(\phi_A, \phi'_A)$ indicates that ϕ_A is the most convenient or most preferred action for agent i among those equivalent to ϕ'_A.

The formula $lend_G(i, H, \phi_A)$, where G and H are disjoint groups and $i \in H$, states that group H can lend agent i to group G for the execution of ϕ_A, provided that agent i is able and authorized to perform it.

The trust predicate $trust_K(i, \tau)$ is the main addition of the present work. Here, K denotes the evaluator, i the evaluated agent, and τ the trust level assigned to i. Trust is therefore made explicit in the language and can be used to constrain intention adoption, execution feasibility, preference-guided choice, and delegation.

The language of mental actions is denoted by $\mathcal{L}_{\mathsf{ACT}}$. The static fragment *L-INF* consists of all formulas of $\mathcal{L}_{\textit{L-DINF}}$ having no subformula of the form $[G{:}\alpha]\varphi$.

We consider agents as partitioned into groups: each agent always belongs to a unique group in Grp. Any agent i may perform a physical action $join_A(i, j)$ to join the group of agent j. The special case $join_A(i, i)$ denotes the action by which i leaves its current group and forms the singleton group $\{i\}$.

Mental actions. As in the original L-DINF framework [21, 13], the language includes mental actions for perception, inference, conjunction formation, belief removal, and reasoning from working memory. In the illustrative scenario below, only perception $+\varphi$ and inference $\downarrow(\varphi, \psi)$ are used. Since their semantics is inherited unchanged, we refer the reader to the original framework for the complete formal treatment.

Recall that $Res = \{r_1, \ldots, r_\ell\}$ is the set of resources. We write $r_j{:}n_j$ to denote an amount of n_j units of resource r_j, and we let

$$Amounts = \{(r_1{:}n_1, \ldots, r_\ell{:}n_\ell) \mid n_1, \ldots, n_\ell \in \mathbb{N}\}.$$

Given $\bar{r} = (r_1{:}n_1, \ldots, r_\ell{:}n_\ell)$ and $\bar{t} = (r_1{:}m_1, \ldots, r_\ell{:}m_\ell)$ in $Amounts$, we write $\bar{r} \leq \bar{t}$ iff $n_j \leq m_j$ for each j, and define $sum(\bar{r}) = \sum_{j=1}^{\ell} n_j$. If $\bar{t} \leq \bar{r}$, then $\bar{r} - \bar{t}$ is defined componentwise.

The role of the trust predicate should be understood in connection with the practical operators already present in the language. In the original framework, formulas such as $intend_i(\phi_A)$, $can\_do_i(\phi_A)$, and $lend_G(i, H, \phi_A)$ describe different aspects of practical agency, but they do not yet distinguish whether the corresponding action is justified from the viewpoint of trust. By introducing $trust_K(i, \tau)$ into the language, we make it possible to state explicitly that the same action may be intended, executable, or delegable only under suitable trust conditions. In this way, the trust-aware extension does not replace the practical layer of L-DINF, but refines it by adding a further semantic dimension.

3.2. Semantics

Many relevant aspects of an agent's behaviour are specified at the model level. This choice keeps the framework modular and allows different practical settings to instantiate executability, budgets, costs, preferences, equivalence classes, lending conditions, and trust assessments in different ways.

An *L-INF* model M is composed of two parts: a *core* part $\mathcal{C}_M$ and a collection of *packages* $\mathcal{P}_M$.

Definition 3.1. *The* core *part $\mathcal{C}_M$ of a model M is a tuple $(W, N, \mathcal{R}, V, S)$, where:*

- W *is a set of worlds;*
- $\mathcal{R} = \{R_i\}_{i \in Agt}$ *is a collection of equivalence relations on W;*
- $N : Agt \times W \to 2^{2^W}$ *is a neighbourhood function such that, for every $i \in Agt$ and $w, v \in W$:*
  (C1) if $X \in N(i, w)$, then $X \subseteq \{u \in W \mid wR_iu\}$;
  (C2) if wR_iv, then $N(i, w) = N(i, v)$;
- $V : W \to 2^{Atm}$ *is a valuation function;*
- $S : W \to 2^{\{do_G(\phi_A), do_i^P(\phi_A) | \phi_A \in Atm_A,\ i \in Agt,\ G \in Grp\}}$ *is a valuation function for executed physical actions.*

For simplicity, let $R_i(w) = \{v \in W \mid wR_iv\}$. The set $R_i(w)$ identifies the situations that agent i considers possible at world w. In cognitive terms, it can be viewed as the set of situations that agent i can retrieve from long-term memory and reason about. While $R_i(w)$ concerns background knowledge, $N(i, w)$ represents the set of facts explicitly believed by i at w, that is, the contents of the agent's working memory.

Constraint (C1) imposes that an agent can explicitly believe only facts compatible with its epistemic state. Constraint (C2) states that if two worlds are indistinguishable with respect to background knowledge, then they cannot differ with respect to explicit beliefs.

Definition 3.2. *Given a core model $\mathcal{C}_M = (W, N, \mathcal{R}, V, S)$, the packages $\mathcal{P}_M$ are:*

- $E : Agt \times W \to 2^{\mathcal{L}_{\mathsf{ACT}}}$*, an executability function for mental actions;*
- $B_1 : Agt \times W \to \mathbb{N}$ *and* $C_1 : Agt \times \mathcal{L}_{\mathsf{ACT}} \times W \to \mathbb{N}$*, respectively the budget and cost functions for mental actions;*
- $A : Agt \times W \to 2^{Atm_A}$*, an executability function for physical actions;*
- $B_2 : Agt \times W \to Amounts$ *and* $C_2 : Agt \times Atm_A \times W \to Amounts$*, respectively the budget and cost functions for physical actions;*
- $H : Agt \times W \to 2^{Atm_A}$*, an enabling function describing which physical actions are allowed by the agent's role;*
- $P : Agt \times W \times Atm_A \to \mathbb{N}$*, a preference function on physical actions;*
- $Q : Atm_A \times W \to 2^{Atm_A}$*, a function describing equivalence classes of physical actions;*
- $F : Agt \times W \times Atm_A \to Atm_A$*, a selector function choosing, for each agent and world, one preferred action in the equivalence class of a given action;*
- $L : Agt \times Grp \times Grp \times Atm_A \times W \to \{\mathsf{true}, \mathsf{false}\}$*, a lending function;*
- $T : Grp \times Agt \times W \to Levels$*, a trust function;*
- $\Theta : Grp \times Atm_A \times W \to Levels^5$*, a policy-threshold assignment function such that, for every policy authority $\Pi \in Grp$, physical action $\phi_A \in Atm_A$, and world $w \in W$,*

$$\Theta(\Pi, \phi_A, w) = \big(\tau_I^\Pi(\phi_A, w), \tau_C^\Pi(\phi_A, w), \tau_L^\Pi(\phi_A, w), \tau_B^\Pi(\phi_A, w), \tau_A^\Pi(\phi_A, w)\big),$$

with

$$\tau_B^\Pi(\phi_A, w) < \tau_A^\Pi(\phi_A, w).$$

Each of these functions is assumed to be invariant under epistemic equivalence in the appropriate sense, namely if wR_iv then the corresponding package values coincide at w and v.

Let us briefly describe the intended role of these packages. For an agent i, $E(i, w)$ is the set of mental actions executable at world w, while $B_1(i, w)$ and $C_1(i, \alpha, w)$ describe, respectively, its mental-action budget and the cost of executing α. These components regulate the cognitive side of agency, that is, the extent to which agents can acquire, manipulate, and revise explicit beliefs.

Similarly, $A(i, w)$ is the set of physical actions executable by i at w, while $B_2(i, w)$ and $C_2(i, \phi_A, w)$ describe the corresponding resource budget and action cost. The function $H(i, w)$ captures role-based admissibility, that is, which physical actions are normatively available to an agent in virtue of its current role within the system. This distinction between bare executability and role-based admissibility is important because an agent may be technically able to perform an action without being authorised to do so.

The preference function P assigns to each physical action a degree of willingness or desirability. This allows the framework to represent not only what an agent can do, but also what it would prefer to do when several admissible alternatives are available. The function Q associates each physical action with its equivalence class, thereby making it possible to group together actions that are interchangeable from a practical point of view. The selector function F then chooses one preferred representative of that class, thus linking equivalence and preference in a formally explicit way.

Finally, L specifies the conditions under which one group may lend an agent to another group for the execution of a task, whereas $T(K, i, w)$ gives the trust level assigned by evaluator K to agent i at world w. In the trust-aware extension, this last component plays a special role: unlike the other packages, it is not introduced merely to enrich the descriptive power of the model, but to constrain the admissibility of practical outcomes. In this sense, trust functions as a bridge between the descriptive representation of the system and the normative regulation of action.

3.2.1. Truth Conditions

Given a formula φ, let

$$\|\varphi\|^M_{i,w} = \{v \in W \mid wR_iv \text{ and } M, v \models \varphi\},$$

whenever $M, v \models \varphi$ is well-defined.

The truth conditions for the core static language are:

1. $M, w \models p$ iff $p \in V(w)$;
2. $M, w \models Cl(\phi_A, \phi'_A)$ iff $\phi'_A \in Q(\phi_A, w)$;
3. $M, w \models trust_K(i, \tau)$ iff $T(K, i, w) = \tau$;
4. $M, w \models exec_G(\alpha)$ iff $\exists i \in G$ such that $\alpha \in E(i, w)$;
5. $M, w \models fCl_i(\phi_A, \phi'_A)$ iff $\phi_A = F(i, w, \phi'_A)$;
6. $M, w \models \neg\varphi$ iff $M, w \not\models \varphi$;
7. $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$;
8. $M, w \models B_i\varphi$ iff $\|\varphi\|^M_{i,w} \in N(i, w)$;
9. $M, w \models K_i\varphi$ iff $M, v \models \varphi$ for every $v \in R_i(w)$;
10. $M, w \models \varphi$, for φ of the form $do_G(\phi_A)$ or $do^P_i(\phi_A)$, iff $\varphi \in S(w)$.

Trust-sensitive practical clauses. In the trust-extended setting, not every intended or preferred action is practically admissible. Intention adoption, feasibility, preference-sensitive action selection, lending, and autonomous execution are filtered by minimum trust requirements.

Besides the trust evaluator K, which assigns trust values through the function $T(K, i, w)$, we introduce a possibly distinct external policy authority $\Pi \in Grp$, responsible for fixing the trust thresholds governing the admissibility of physical actions. We assume that the model contains a threshold-assignment function Θ such that, for every physical action $\phi_A \in Atm_A$ and world $w \in W$,

$$\Theta(\Pi, \phi_A, w) = \big(\tau^\Pi_I(\phi_A, w), \tau^\Pi_C(\phi_A, w), \tau^\Pi_L(\phi_A, w), \tau^\Pi_B(\phi_A, w), \tau^\Pi_A(\phi_A, w)\big),$$

with

$$\tau^\Pi_B(\phi_A, w) < \tau^\Pi_A(\phi_A, w).$$

Here, $\tau^\Pi_I(\phi_A, w)$ is the threshold required for intention adoption, $\tau^\Pi_C(\phi_A, w)$ the threshold required for individual execution feasibility, $\tau^\Pi_L(\phi_A, w)$ the threshold required for delegation or lending, $\tau^\Pi_B(\phi_A, w)$

the blocking threshold, and $\tau_A^{\Pi}(\phi_A, w)$ the autonomy threshold. This makes it possible to treat thresholds as exogenous policy parameters, possibly sensitive not only to the action itself but also to external institutional, regulatory, or contextual factors encoded in the world w. At the same time, the framework keeps trust evaluation and policy specification conceptually distinct, which improves explainability: practical outcomes are grounded not only in assessed trust values but also in explicit admissibility criteria.

The trust-sensitive semantic clauses are then given as follows:

- $M, w \models intend_i(\phi_A) \iff \phi_A \in A(i,w) \ \wedge \ T(K,i,w) \geq \tau_I^{\Pi}(\phi_A, w)$;
- $M, w \models pref\_do_i(\phi_A, d) \iff \phi_A \in A(i,w) \cap H(i,w) \ \wedge \ P(i,w,\phi_A) = d \ \wedge \ T(K,i,w) \geq \tau_C^{\Pi}(\phi_A, w)$;
- $M, w \models pref\_do_G(i, \phi_A) \iff M, w \models pref\_do_i(\phi_A, d)$ for some $d \in \mathbb{N}$ such that $d = \max\{P(j,w,\phi_A) \mid j \in G, \ \phi_A \in A(j,w) \cap H(j,w), \ T(K,j,w) \geq \tau_C^{\Pi}(\phi_A, w)\}$;
- $M, w \models can\_do_i(\phi_A) \iff \phi_A \in A(i,w) \cap H(i,w) \ \wedge \ T(K,i,w) \geq \tau_C^{\Pi}(\phi_A, w)$;
- $M, w \models can\_do_G(\phi_A) \iff \exists i \in G \ M, w \models can\_do_i(\phi_A)$;
- $M, w \models lend_G(i, H, \phi_A) \iff (G \cap H = \emptyset) \wedge i \in H \wedge \phi_A \in A(i,w) \cap H(i,w) \wedge T(G,i,w) \geq \tau_L^{\Pi}(\phi_A, w) \ \wedge \ F(i,w,\phi_A) = \phi_A$.

To make the operational effect of trust explicit, we define the following policy for critical actions:

$$decision_{K,\Pi}(i, \phi_A, w) = \begin{cases} allow & \text{if } T(K,i,w) \geq \tau_A^{\Pi}(\phi_A, w), \\ delegate & \text{if } \tau_B^{\Pi}(\phi_A, w) < T(K,i,w) < \tau_A^{\Pi}(\phi_A, w), \\ block & \text{if } T(K,i,w) \leq \tau_B^{\Pi}(\phi_A, w). \end{cases}$$

The introduction of this policy makes the operational meaning of trust more transparent. The threshold $\tau_A^{\Pi}(\phi_A, w)$ identifies the level above which the system recognises the agent as sufficiently reliable for autonomous execution. By contrast, $\tau_B^{\Pi}(\phi_A, w)$ marks the point below which the risk associated with the action is considered too high to permit execution. The interval between these two thresholds corresponds to a genuinely intermediate region: the agent is not trusted enough to act alone, but not so untrusted as to justify a complete prohibition. This is precisely the region in which delegation becomes the most appropriate logical outcome. As a consequence, the framework does not collapse trust into a binary yes/no condition, but uses it to articulate different modes of regulated agency. In this sense, the policy induces a structured transition between autonomy, delegation, and inhibition rather than a flat trust/no-trust dichotomy.

This policy is especially relevant for cybersecurity and digital forensics. A binary allow/block choice is often too coarse: actions in the grey zone should not always be rejected outright, but neither should they be executed autonomously. By making delegation a first-class logical outcome, L-DINF provides an explicit and explainable treatment of borderline trust cases.

Moreover, the roles of K and Π are formally distinct. The evaluator K is responsible for monitoring agent i and assigning a trust value on the basis of the available evidence, whereas the policy authority Π determines the normative trust thresholds required for the admissibility of a given action. The simpler action-based notation is recovered as a special case by assuming that, for every world $w \in W$,

$$\tau_X^{\Pi}(\phi_A, w) = \tau_X(\phi_A), \qquad \text{for each } X \in \{I, C, L, B, A\}.$$

3.2.2. Inherited Dynamic Component

The dynamic semantics of mental actions is inherited unchanged from the original L-DINF framework [3, 1, 16]. In particular, formulas of the form $[G{:}\alpha]\varphi$ describe the result of executing a mental action α by a group G, through the corresponding update of explicit beliefs and cognitive resources as defined in the base logic.

The present contribution does not modify such dynamic clauses. Trust is introduced only at the level of practical agency, where it constrains intention adoption, preference-guided action selection, physical-action feasibility, lending, and the resulting policy outcome. Accordingly, the complete update semantics for mental actions is not repeated here; the reader is referred to the original L-DINF definitions and results. Physical-action resource consumption is likewise inherited from the base framework and remains unaffected by the trust-aware extension. A trust-sensitive treatment of mental actions is left for future work.

3.2.3. Axiomatization

A sound and complete axiomatization for the trust-free epistemic and dynamic components of $L\text{-}DINF$ is already available in the literature. In particular, the original framework and its treatment of group dynamics are developed in [3, 1], while the preference-aware extension adopted as the immediate basis of the present proposal is presented in [16]. In the present work, we inherit without modification the axioms governing knowledge, explicit belief, and the dynamics of mental actions from the base framework, and we restrict attention to the additional principles required by the trust-sensitive practical layer.

Since the semantic treatment introduced above makes trust thresholds depend on a policy authority Π, on the action ϕ_A, and possibly on the world w, whereas axioms are stated at the object-language level, we formulate the proof-theoretic counterpart under the following action-based special case:

$$\tau_X^{\Pi}(\phi_A, w) = \tau_X^{\Pi}(\phi_A) \qquad \text{for every } X \in \{I, C, L, B, A\}.$$

Thus, thresholds are taken to be fixed by the policy authority Π as a function of the action only. This is exactly the simplified notation already identified in the semantic section as a special case of the more general threshold assignment function Θ.

For every evaluator $E \in Grp$, agent $i \in Agt$, physical action $\phi_A \in Atm_A$, and threshold type $X \in \{I, C, L\}$, let

$$\mathsf{Trust}_{\geq X}^{E,\Pi}(i, \phi_A) \;:=\; \bigvee_{\tau \in Levels,\ \tau \geq \tau_X^{\Pi}(\phi_A)} trust_E(i, \tau).$$

Intuitively, $\mathsf{Trust}_{\geq X}^{E,\Pi}(i, \phi_A)$ says that evaluator E assigns to agent i a trust level at least as high as the threshold required by policy authority Π for action ϕ_A.

The trust-sensitive axioms are then the following:

(AxT_1) $\quad intend_i(\phi_A) \to \mathsf{Trust}_{\geq I}^{K,\Pi}(i, \phi_A),$

(AxT_2) $\quad intend_G(\phi_A) \leftrightarrow \bigwedge_{i \in G} intend_i(\phi_A),$

(AxT_3) $\quad pref\_do_i(\phi_A, d) \to \mathsf{Trust}_{\geq C}^{K,\Pi}(i, \phi_A),$

(AxT_4) $\quad pref\_do_G(i, \phi_A) \to \mathsf{Trust}_{\geq C}^{K,\Pi}(i, \phi_A),$

(AxT_5) $\quad can\_do_i(\phi_A) \to \mathsf{Trust}_{\geq C}^{K,\Pi}(i, \phi_A),$

(AxT_6) $\quad can\_do_G(\phi_A) \leftrightarrow \bigvee_{i \in G} can\_do_i(\phi_A),$

(AxT_7) $\quad lend_G(i, H, \phi_A) \to (G \cap H = \emptyset) \wedge (i \in H) \wedge can\_do_i(\phi_A) \wedge fCl_i(\phi_A, \phi_A) \wedge \mathsf{Trust}_{\geq L}^{G,\Pi}(i, \phi_A).$

These principles should be understood as necessary trust constraints on the derivability of practical formulas. In particular, the axioms do not attempt to internalize the full semantic content of executability, role-admissibility, preference selection, or environmental attachment, all of which remain partly model-dependent in the original $L\text{-}DINF$ framework. For this reason, formulas of the form $do_G(\phi_A)$ are still left outside the axiomatic treatment, exactly as in the base logic.

Observe also that the lending axiom is intentionally stated as an implication rather than as a biconditional. This choice matches the semantic clause more faithfully: lending depends not only on trust, group disjointness, and feasibility, but also on model-dependent conditions such as the preferred representative selected for the relevant action class. Hence, the axiom expresses a necessary trust-sensitive condition for lending, without overstating it as a complete syntactic characterization.

4. Validation

The trust-aware extension proposed in this paper is assessed at two complementary levels. On the one hand, it is analysed from a meta-theoretical viewpoint by relating its unchanged fragment to the original L-DINF framework, for which an axiomatization, a canonical model construction, and a strong completeness result are already available. On the other hand, its operational meaning is illustrated by showing that the trust-sensitive clauses capture the intended behaviour in a representative borderline scenario. These two analyses address different goals: the former identifies the formal continuity with the base logic, while the latter shows the explanatory adequacy of the trust-sensitive layer in practice.

4.1. Meta-theoretical validation

The present proposal extends L-DINF by introducing an explicit trust predicate, an ordered trust function, differentiated trust thresholds for intention adoption, autonomous execution, and lending, together with a three-valued policy distinguishing autonomous execution, delegation, and inhibition. At the same time, the dynamic semantics of mental actions is inherited unchanged from the original framework. This modular design makes it possible to relate the trust-aware extension to the already established meta-theoretical properties of L-DINF.

In particular, the original version of L-DINF was provided with an axiomatization, a canonical model construction, and a proof of strong completeness [3, 1, 16]. Since the present extension refines the interpretation of some practical formulas already occurring in L-DINF, the preservation statement below is deliberately restricted to the inherited fragment whose semantic clauses remain unchanged.

Let $\mathcal{L}_{\mathrm{unch}}$ denote the fragment of L-DINF containing only formulas whose semantic clauses are inherited unchanged in the present extension. In particular, $\mathcal{L}_{\mathrm{unch}}$ excludes formulas containing $trust_K(i, \tau)$, trust-dependent policy expressions, and the trust-filtered practical formulas $intend$, can_do, $pref\_do$, and $lend$.

Lemma 4.1 (Conservativity on the unchanged fragment)**.** *Let φ be a formula of $\mathcal{L}_{\mathrm{unch}}$. Then φ is valid in the trust-aware class of models if and only if it is valid in the original class of L-DINF models.*

Proof sketch. Every original L-DINF model M can be expanded into a trust-aware model M^T by adding a trust function and a threshold-assignment component satisfying the required ordering assumptions. Since every semantic clause occurring in $\mathcal{L}_{\mathrm{unch}}$ is inherited unchanged from the base framework, the truth value of any formula of this fragment is preserved under such an expansion. Conversely, restricting a trust-aware model to the components relevant for $\mathcal{L}_{\mathrm{unch}}$ does not affect the interpretation of formulas in that fragment. Hence, the extension introduces no new validities in the inherited unchanged sublanguage. □

The above observation yields the corresponding inherited completeness result.

Theorem 4.1 (Inherited strong completeness on the unchanged fragment)**.** *Let $\Gamma \cup \{\varphi\}$ be a set of formulas of $\mathcal{L}_{\mathrm{unch}}$. If $\Gamma \models \varphi$ with respect to the trust-aware class of models, then $\Gamma \vdash_{L\text{-}DINF} \varphi$.*

Proof sketch. By Lemma 4.1, semantic consequence over trust-aware models coincides, on $\mathcal{L}_{\mathrm{unch}}$, with semantic consequence over the original class of L-DINF models. Strong completeness for the inherited framework follows from the canonical-model treatment developed for the base logic [3, 1, 16]. Therefore, every semantic consequence preserved on the unchanged fragment remains derivable in the corresponding original proof system. □

A second useful property concerns the internal coherence of the trust-based decision policy.

Lemma 4.2 (Decision partition under externally fixed thresholds)**.** *For every evaluator $K \in Grp$, policy authority $\Pi \in Grp$, agent $i \in Agt$, physical action $\phi_A \in Atm_A$, and world $w \in W$, if $\tau_B^\Pi(\phi_A, w) < \tau_A^\Pi(\phi_A, w)$, then exactly one of the following outcomes holds:*

$$decision_{K,\Pi}(i, \phi_A, w) = allow,\ decision_{K,\Pi}(i, \phi_A, w) = delegate,\ decision_{K,\Pi}(i, \phi_A, w) = block.$$

Justification. Let $t = T(K, i, w)$. Since trust values are drawn from the ordered scale $Levels$, and since $\tau_B^\Pi(\phi_A, w) < \tau_A^\Pi(\phi_A, w)$, exactly one of the following mutually exclusive cases must obtain:

$$t \geq \tau_A^\Pi(\phi_A, w), \qquad \tau_B^\Pi(\phi_A, w) < t < \tau_A^\Pi(\phi_A, w), \qquad t \leq \tau_B^\Pi(\phi_A, w).$$

By definition of the policy function $decision_{K,\Pi}$, these three cases correspond, respectively, to the outcomes $allow$, $delegate$, and $block$. Hence the decision policy induces a partition of the trust scale into three jointly exhaustive and pairwise disjoint regions. □

Property 4.1 (Trust-sensitive outcome discrimination)**.** *There exist two trust-aware models M_1 and M_2, a world w, an agent i, and a physical action ϕ_A such that the models coincide on the core model, on the non-trust practical packages, and on Θ, while they differ only in $T(K, i, w)$. Moreover, $can\_do_i(\phi_A)$ holds in both models, but*

$$decision^{M_1}_{K,\Pi}(i, \phi_A, w) = allow \qquad \textit{and} \qquad decision^{M_2}_{K,\Pi}(i, \phi_A, w) = delegate.$$

Hence, the trust-aware layer can distinguish admissible outcomes without changing technical executability or role-based admissibility.

Proof sketch. Let the models coincide except for $T^{M_1}(K, i, w) = very\_high$ and $T^{M_2}(K, i, w) = medium$, and assume $\phi_A \in A(i, w) \cap H(i, w)$, $\tau_C^\Pi(\phi_A, w) = medium$, $\tau_B^\Pi(\phi_A, w) = low$, and $\tau_A^\Pi(\phi_A, w) = very\_high$. Both trust values satisfy the feasibility threshold, so $can\_do_i(\phi_A)$ holds in both models. The first value satisfies the autonomy threshold, while the second lies strictly between the blocking and autonomy thresholds. The two stated decision outcomes follow directly from the definition of $decision_{K,\Pi}$. □

Taken together, these observations do not yet amount to a full expressivity or strong completeness result for the trust-enriched language. Obtaining such stronger results would require an explicit proof-theoretic treatment of trust formulas, threshold-sensitive practical clauses, and policy expressions, together with a suitable canonical extension of the model construction. Nevertheless, Property 4.1 identifies a concrete additional modelling capability of the proposed layer: agents that are indistinguishable with respect to technical executability and the non-trust practical structure may receive different admissible outcomes because of their trust assessment. This is precisely the regulated-autonomy distinction that the extension is intended to make explicit.

4.2. Illustrative validation

We now reconsider the cybersecurity scenario as an instance of operational validation. The example is meant to show that the trust-sensitive semantics captures an explainable intermediate behaviour in which an agent may be able and willing to perform an action, while still lacking the level of trust required for autonomous execution under externally specified policy thresholds.

Assume a system composed of a sensor s, a junior analyst a, and a senior responder r. Let p denote the alert generated by the sensor, let q denote the derived conclusion that the host is compromised, and let ϕ_A denote the physical action $isolate\_host(h_1)$. Suppose further that the junior analyst belongs to the blue team G_B and the senior responder belongs to the response group G_R.

At the epistemic level, the framework explains how the system moves from a perceived signal to an explicit practical conclusion. After perception and inference, one obtains:

$$[\{s\} : +p]\, B_s p \qquad \text{and} \qquad [\{s\} :\downarrow (p, q)]\, B_s q.$$

Thus, the practical response is not triggered opaquely, but arises from an explicit sequence of cognitive steps.

Assume now that K is the trust evaluator and Π is the policy authority responsible for fixing the trust thresholds for critical response actions. Suppose that, at world w, the following trust values and thresholds hold:

$$T(K, a, w) = \mathit{medium}, \qquad T(K, r, w) = \mathit{very\_high},$$
$$\tau_I^{\Pi}(\phi_A, w) = \mathit{medium}, \qquad \tau_C^{\Pi}(\phi_A, w) = \mathit{medium}, \qquad \tau_L^{\Pi}(\phi_A, w) = \mathit{high},$$
$$\tau_B^{\Pi}(\phi_A, w) = \mathit{low}, \qquad \tau_A^{\Pi}(\phi_A, w) = \mathit{very\_high},$$

with

$$\tau_B^{\Pi}(\phi_A, w) < \tau_A^{\Pi}(\phi_A, w).$$

Since the trust of a is sufficient for intention adoption, one has:

$$M, w \models \mathit{intend}_a(\phi_A).$$

Moreover, since a is technically executable and normatively enabled for ϕ_A, and since

$$T(K, a, w) = \mathit{medium} \geq \tau_C^{\Pi}(\phi_A, w) = \mathit{medium},$$

it follows that

$$M, w \models \mathit{can\_do}_a(\phi_A).$$

This shows that the agent is not excluded from practical feasibility: the junior analyst may form the intention and is, in principle, able to perform the action.

However, the same agent is not sufficiently trusted for autonomous execution under the policy fixed by Π, because

$$T(K, a, w) = \mathit{medium} < \tau_A^{\Pi}(\phi_A, w) = \mathit{very\_high},$$

while at the same time

$$T(K, a, w) = \mathit{medium} > \tau_B^{\Pi}(\phi_A, w) = \mathit{low}.$$

Therefore,

$$\mathit{decision}_{K,\Pi}(a, \phi_A, w) = \mathit{delegate}.$$

This is the crucial point: practical feasibility and autonomous authorization do not coincide. The analyst may be able to act, but the institutional policy does not allow autonomous execution for an action of this criticality.

Assume moreover that the senior responder r is normatively enabled, sufficiently trusted, and that among the actions equivalent to ϕ_A, the preferred one is ϕ'_A, namely the less disruptive action $\mathit{restrict\_network}(h_1)$. Then:

$$M, w \models fCl_r(\phi'_A, \phi_A).$$

Thus, even if the original practical goal is to isolate the host, the framework may select a preferred equivalent action within the same practical class.

Suppose also that the trust of the borrowing group G_B in the senior responder is high enough to satisfy the lending condition, namely

$$T(G_B, r, w) \geq \tau_L^{\Pi}(\phi'_A, w).$$

Then:

$$M, w \models \mathit{lend}_{G_B}(r, G_R, \phi'_A).$$

Since

$$T(K, r, w) = very\_high \geq \tau_A^{\Pi}(\phi'_A, w),$$

the senior responder lies above the autonomy threshold for the delegated action, and hence

$$decision_{K,\Pi}(r, \phi'_A, w) = allow.$$

The scenario therefore illustrates the intended operational behaviour of the framework. The junior analyst is trusted enough to recognize the situation, adopt the relevant intention, and count as a feasible executor, but not enough to perform the action autonomously. The policy authority Π sets the normative thresholds for critical actions, while the evaluator K assigns trust values to agents. The result is an explainable three-step structure: cognitive recognition of the incident, intermediate policy outcome for the junior analyst ($delegate$), and autonomous execution by a more trusted responder. In this way, the framework captures the grey zone between unrestricted autonomy and outright prohibition without collapsing trust into a merely binary condition. Note also that the outcome is sensitive to policy variation: for instance, raising the blocking threshold from low to $medium$ would move the junior analyst from the delegation region to the blocking region, without changing executability or preference conditions.

5. Conclusion and Future Work

This paper has introduced a trust-aware extension of L-DINF in which trust is treated as an explicit logical component of multi-agent decision making. By integrating trust into the language and semantics, the framework provides a principled account of how trust constrains intention, feasibility, preference-guided choice, and delegation. This makes it possible to distinguish clearly among autonomous execution, delegation, and inhibition.

The proposed extension strengthens L-DINF as a logic of regulated autonomy and explainable agency. Its main contribution is to show that trust can be formalized not only as an evaluative notion but also as a normative and semantic condition on action. As a result, the framework makes explicit the reasons why an agent may act, must delegate, or must refrain from acting, while preserving the inherited fragment of L-DINF whose semantic clauses remain unchanged. More broadly, the paper argues that trust can be embedded into the semantic admissibility conditions of a cognitive logical framework rather than being left to an external implementation layer. Therefore, this work suggests that trust can be fruitfully integrated into the logical analysis of autonomous systems, especially in domains where accountability, explainability, and regulated agency are essential.

A natural direction for future work is to extend trust to *mental actions*, so that trust may regulate not only what agents do, but also how they accept information, infer conclusions, and revise beliefs. A further meta-theoretical direction is to characterize the exact expressive gain of the trust-aware language with respect to classical L-DINF. Property 4.1 provides an initial step in this direction by showing that different trust assessments can induce different admissibility outcomes even when the non-trust practical structure is kept fixed. Another important direction is to connect the present semantic layer with other trust-evaluation mechanisms. In the current framework, trust assessment is intentionally kept exogenous so as to preserve modularity and allow integration with different trust estimation methods. Since trust evaluation remains external to the logic, heterogeneous trust-estimation mechanisms can be, in fact, incorporated without modifying the semantic framework itself.

Beyond the specific extension presented here, we believe to have contributed to a broader view of regulated autonomy in which trust acts as a semantic mechanism governing agency rather than as an external score attached to agents. In this perspective, autonomy becomes graded rather than binary, with delegation emerging as an explicit intermediate mode between unrestricted autonomy and complete inhibition. In fact, the framework supports a notion of regulated autonomy in which, in a future perspective, responsibility can be transferred to more trusted agents instead of reducing decisions to a simple allow/block dichotomy.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. Costantini, V. Pitoni, Memory management in resource-bounded agents, in: M. Alviano, G. Greco, F. Scarcello (Eds.), AI*IA 2019 - Advances in Artificial Intelligence - XVIIIth International Conference of the Italian Association for Artificial Intelligence, 2019, Proceedings, volume 11946 of *LNCS*, Springer, 2019, pp. 46–58.

[2] S. Costantini, V. Pitoni, Towards a logic of "inferable" for self-aware transparent logical agents, in: Proceedings of the Italian Workshop on Explainable Artificial Intelligence (XAI.it@AI*IA 2020), volume 2742 of *CEUR Workshop Proceedings*, CEUR-WS, 2020, pp. 68–79.

[3] S. Costantini, A. Formisano, V. Pitoni, An epistemic logic for formalizing group dynamics of agents, Interaction Studies 23 (2022) 391 – 426. doi:10.1075/is.22019.cos.

[4] S. Costantini, A. Formisano, V. Pitoni, An epistemic logic for modular development of multi-agent systems, in: Lecture Notes in Computer Science, volume 13190 of *Lecture Notes in Artificial Intelligence*, Springer, 2022, pp. 72–91.

[5] M. Gelfond, Answer sets, in: Handbook of Knowledge Representation, Elsevier, 2007.

[6] V. Lifschitz, Answer set programming, volume 3, Springer Cham, 2019.

[7] G. Brewka, T. Eiter, M. T. (eds.), Answer set programming: Special issue, AI Magazine 37 (2016).

[8] A. Vozna, A. Monaldini, S. Costantini, V. Pitoni, D. Pado, An asp-based solution to the medical appointment scheduling problem, in: Proceedings of the 41st International Conference on Logic Programming (ICLP 2025), volume 439 of *Electronic Proceedings in Theoretical Computer Science*, Open Publishing Association, 2026, pp. 367–382. doi:10.4204/EPTCS.439.25.

[9] A. Vozna, A. Monaldini, S. Costantini, D. Pado, V. Pitoni, Scalable and ethical medical scheduling: An asp-based framework with patient-centered experimental validation, in: D. Azzolini, S. Batsakis, M. A. B. Santana, P. Bruno, L. A. Cecchi, J. Fandinno, M. Hecher, M. Maratea, J. F. Morales, B. Muñiz, E. Papadakis, L. Robaldo, L. Serafini, F. L. Scudo, I. Tachmazidis, A. Tarzariol, M. Vallati, A. Wyner (Eds.), Joint Proceedings of the Workshops and Doctoral Consortium of the 41st International Conference on Logic Programming (ICLP-WS-DC 2025) co-located with 41st International Conference on Logic Programming (ICLP 2025), Rende, Italy, September 9-13, 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, p. 8. URL: https://ceur-ws.org/Vol-4117/short_rcra_5.pdf.

[10] S. Costantini, L. De Lauretis, V. Pitoni, Bridging the gap: An executable simulator for the l-dinf epistemic logic of cooperative agents, in: Proceedings of the 27th Workshop "From Objects to Agents", June 15–17, 2026, Salerno, Italy, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[11] A. S. Rao, M. P. Georgeff, Modeling rational agents within a BDI-architecture, in: R. Fikes, E. Sandewall (Eds.), Proceedings of Knowledge Representation and Reasoning (KR&R-91), Morgan Kaufmann Publishers: San Mateo, CA, 1991, pp. 473–484.

[12] R. H. Bordini, J. F. Hübner, BDI agent programming in agentspeak using jason, in: International workshop on computational logic in multi-agent systems, Springer, 2005, pp. 143–164.

[13] P. Balbiani, D. F. Duque, E. Lorini, A logical theory of belief dynamics for resource-bounded agents, in: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016, ACM, 2016, pp. 644–652.

[14] P. Balbiani, D. Fernández-Duque, E. Lorini, The dynamics of epistemic attitudes in resource-bounded agents, Stud Logica 107 (2019) 457–488. URL: https://doi.org/10.1007/s11225-018-9798-4. doi:10.1007/S11225-018-9798-4.

[15] E. Lorini, Reasoning about cognitive attitudes in a qualitative setting, in: F. Calimeri, N. Leone, M. Manna (Eds.), Logics in Artificial Intelligence - 16th European Conference, JELIA 2019, Rende, Italy, May 7-11, 2019, Proceedings, volume 11468 of *Lecture Notes in Computer Science*,

Springer, 2019, pp. 726–743. URL: https://doi.org/10.1007/978-3-030-19570-0_47. doi:10.1007/978-3-030-19570-0_47.

[16] S. Costantini, A. Formisano, V. Pitoni, Preference management in epistemic logic L-DINF, in: Proc. of CILC 2023, the 38th Italian Conference on Computational Logic, CEUR Workshop Proceedings, volume 3428, 2023.

[17] G. C. M. Amaral, T. P. Sales, G. Guizzardi, D. Porello, Towards a reference ontology of trust, in: H. Panetto, C. Debruyne, M. Hepp, D. Lewis, C. A. Ardagna, R. Meersman (Eds.), On the Move to Meaningful Internet Systems: OTM 2019 Conferences - Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21-25, 2019, Proceedings, volume 11877 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 3–21. doi:10.1007/978-3-030-33246-4_1.

[18] G. Jethava, U. P. Rao, Exploring security and trust mechanisms in online social networks: An extensive review, Computers & Security (2024) 103790.

[19] W. Li, W. Meng, L. F. Kwok, Surveying trust-based collaborative intrusion detection: state-of-the-art, challenges and future directions, IEEE Communications Surveys & Tutorials 24 (2021) 280–305.

[20] A. Goldman, Theory of mind, in: E. Margolis, R. Samuels, S. P. Stich (Eds.), The Oxford Handbook of Philosophy of Cognitive Science, Oxford University Press, 2012.

[21] S. Costantini, A. Formisano, V. Pitoni, An epistemic logic for modular development of multi-agent systems, in: N. Alechina, M. Baldoni, B. Logan (Eds.), Proc. of EMAS'21, Revised Selected papers, volume 13190 of *LNCS*, Springer, 2021, pp. 72–91.