

TrustForge: A Computational Architecture for Socio-Cognitive Trust in Multi-Agent Systems

by

Adina-Nicola Dobrinoiu

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on
26 October 2026

Student number: 5455146
Thesis committee: Dr. Chirag Raman, TU Delft, supervisor
Dr. Luciano Cavalcante Siebert, TU Delft, thesis committee member

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

TRUSTFORGE: A COMPUTATIONAL ARCHITECTURE FOR SOCIO-COGNITIVE TRUST IN MULTI-AGENT SYSTEMS

Adina Nicola Dobrinoiu

Tapri Lab, Department of Pattern Recognition and Bioinformatics
Delft University of Technology
Delft, The Netherlands

ABSTRACT

As artificial intelligence agents become widespread, markets may shift towards virtual agent economies, in which agents transact with one another at speeds beyond human oversight and without knowing each other’s intentions. Agents in such economies must decide whom to trust, yet a single trust score cannot tell whether a partner failed because it was unable or because it was unwilling. The socio-cognitive model of trust draws this distinction by decomposing trust into competence, willingness and opportunity, but it specifies no procedure for computing these beliefs. We address this gap with TrustForge, a language model agent architecture that operationalizes the socio-cognitive trust model, evaluated in the TruthMarketTwin environment. TrustForge uses a two-stage reasoning pipeline to ground each dimension in environmental observations and compose them into an overall trust score that guides delegation, while preserving the distinction between capability and intent. Across two market types and two seller budgets, trust protects buyers from seller deception and improves their utility per purchase, reducing deceptive sell-through from 28.10% to 13.82% in the reputation-only market. Buyers compose trust non-compensatorily, and a single deceptive sale decreases trust by more than an honest sale increases it, consistent with the asymmetry principle of human trust. We further extend trust to testimony, so that agents also assess each other’s reports against their own observations. Toward their peers, buyers attribute high willingness but lower competence, separating capability from intent. However, buyers do not make this separation for sellers, lowering all three dimensions together after a deceptive sale. The code can be found at: <https://github.com/tapri-lab/trustforge>.

1 INTRODUCTION

The widespread adoption of artificial intelligence agents points toward a potential shift from human-centered markets to virtual agent economies that operate at speeds beyond real-time oversight (Tomasev et al., 2025). Such settings function as open multi-agent systems, where the intentions of the participants are unknown (Pinyol & Sabater-Mir, 2013). At scale, the resulting behavioural uncertainties may contribute to the accidental emergence of permeable economies that extend beyond controlled environments into real-world economic processes (Tomasev et al., 2025).

To mitigate these risks, interaction control mechanisms encompass three main methods: security approaches focusing on structural guarantees, institutional approaches relying on centralized authority, and social approaches enabling agents to self-regulate interactions through peer modeling (Pinyol & Sabater-Mir, 2013). Situated within social mechanism design, this paper focuses on modeling trust to guide agent interactions. Within virtual agent economies, trust is defined as a delegator’s degree of belief in a delegatee’s capability to execute a task in alignment with both explicit constraints and implicit intent (Tomašev et al., 2026). While merging capability and intent into a single trust belief is effective for direct task delegation, in complex or human-in-the-loop delegation chains, audit mechanisms must decouple genuine execution failure from malicious direction (Tomašev et al., 2026).

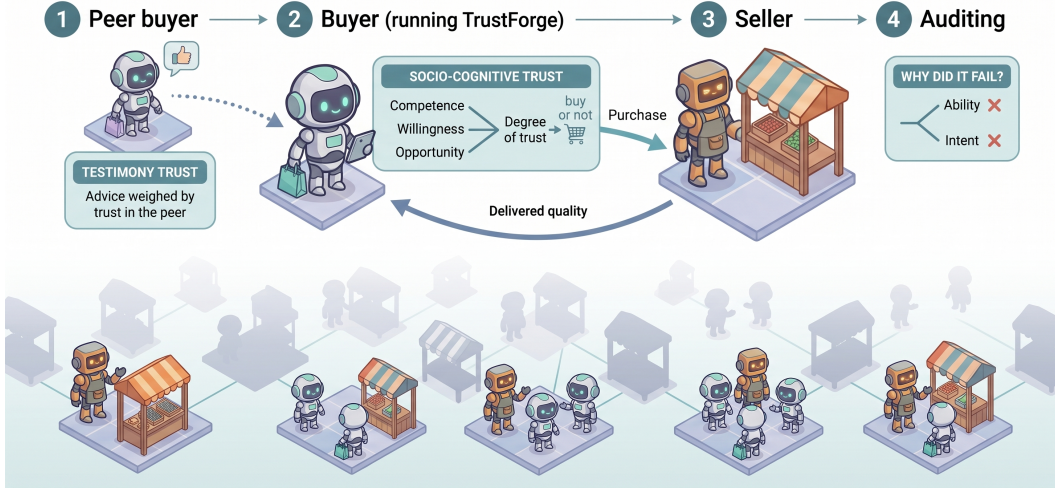


Figure 1: Overview of TrustForge in the TruthMarketTwin market. (1) Other buyers, also running TrustForge, share advice about sellers, which the buyer weighs by its trust in each of them and checks against its own deliveries. (2) The buyer assesses each seller’s competence, willingness and opportunity and combines them into a degree of trust that guides its purchases. (3) The seller delivers a product whose quality may differ from what it advertised. (4) Keeping competence and willingness separate lets a failed delegation be traced to ability or to intent. The market uses either reputation alone or reputation combined with warrants.

To operationalize this distinction, we adopt the socio-cognitive trust model (Castelfranchi & Falcone, 2010). This framework decomposes trust relative to a given task across three core dimensions: competence, representing the ability to execute it; willingness, defining the intention and persistent commitment to achieve it; and opportunity capturing whether the external factors allow for a successful execution. Although this framework provides a structured, context-dependent way to reason about trust through internal mental states, it remains non-procedural, lacking explicit mechanisms to calculate and update these beliefs (Pinyol & Sabater-Mir, 2013). We address this limitation by introducing TrustForge, an agent architecture that operationalizes the socio-cognitive trust model by extending MindForge (Lică et al., 2025). This produces an overall trust score to guide delegation while preserving the structural distinction between capability and intent for auditing.

We make four main contributions: (1) We **introduce TrustForge, operationalizing the socio-cognitive trust model** (Castelfranchi & Falcone, 2010) into a procedural architecture within MindForge (Lică et al., 2025). As seen in Figure 1, we define a two-stage reasoning pipeline that first grounds beliefs about competence, willingness, and opportunity in theory-of-mind partner beliefs and environmental observations, and then composes them into an overall trust score. (2) We **evaluate TrustForge in the TruthMarketTwin environment** (Lei et al., 2026) across two market types and two seller budgets. Trust protects buyers from deception and improves their utility per purchase. When buyers combine competence, willingness and opportunity into a trust score, a weakness in one cannot be offset by the other two. Buyers also lower trust more after a deceptive sale than they raise it after an honest one. (3) We **extend trust to testimony**. Buyers also judge how far to trust each other’s reports by comparing them with what they have observed themselves, so that advice about a seller counts only as much as its source is trusted. Buyers attribute high willingness but lower competence to their peers, which separates capability from intent. (4) We **show that trust does not depend on the Mindforge partner model**. Buyers with and without the partner model attribute similar trust to sellers. The two components still play different roles. Trust mainly decides whether buyers purchase from a seller, while the partner model improves which purchases they make.

2 RELATED WORK

Cognitive trust models. A line of work in multi-agent trust systems has focused on the cognitive paradigm, rooting trust in an agent’s explicit internal beliefs and mental states (Sabater & Sierra, 2005). Foundational work by Castelfranchi and Falcone initially defined core trust through competence and disposition beliefs (Castelfranchi & Falcone, 1998), later formalizing these dimensions into competence, willingness, and opportunity (Castelfranchi & Falcone, 2010). Closer to our work are architectures that integrate these cognitive notions into agent reasoning. Repage (Sabater-Mir et al., 2006) separates an agent’s own image of a partner from its reported reputation, and BDI+Repage (Pinyol et al., 2012) embeds these evaluations in a BDI agent through multi-context logic. Fuzzy cognitive maps have been used to infer a trustee’s ability, willingness and external conditions from its membership in predefined categories (Venanzi et al., 2011). These approaches, summarized in Table 1, represent trust-related beliefs numerically and combine them through predefined aggregation rules. In contrast, **TrustForge** forms each dimension from environmental observations and theory-of-mind partner beliefs, and expresses it in natural language alongside its score.

Numerical and learning-based trust. Another line of work relies on the numerical paradigm, utilizing aggregation or game theory approaches without explicit cognitive attitudes (Sabater & Sierra, 2005). Early models computed trust from direct information alone (Marsh, 1994), while ReGreT (Sabater & Sierra, 2001), FIRE (Huynh et al., 2006) and the Beta Reputation system (Jøsang & Ismail, 2002) aggregate broader sources, including witness and sociological information. More recently, numerical models have used reinforcement learning to capture trust dynamics from repeated interactions. These include actor-critic weightings for advisors (Yu et al., 2014), associative learning (Frey & Martinez, 2024), Q-learning in trust games (Zheng et al., 2024; Hu et al., 2026), and exponential moving averages for memory (Meylahn et al., 2024). These approaches, also listed in Table 1, process multi-source information effectively, but represent trust as a single score that cannot distinguish a partner’s capability from its intent. In contrast, **TrustForge** preserves this distinction, so that a failed delegation can be traced to a lack of ability or a lack of willingness.

Table 1: **Comparison to computational trust models.** We compare **TrustForge** to existing architectures along three axes. Cognitive: belief-based trust (✓), partial (∼), or numerical (−). Information Sources: Direct (DI), Witness (WI), Sociological (SI), and Prejudice (PI). A plus sign (+) denotes explicit aggregation. Granular: Context-dependent (✓) vs. non context-dependent(−).

Model	Cognitive	Information Sources	Granular
Socio-cognitive (Castelfranchi & Falcone, 1998)	✓	−	✓
Fuzzy Cognitive (Venanzi et al., 2011)	✓	PI	✓
Repage (Sabater-Mir et al., 2006)	∼	DI + WI	✓
BDI+Repage (Pinyol et al., 2012)	✓	DI + WI	✓
Marsh’s Formalism (Marsh, 1994)	−	DI	✓
ReGreT (Sabater & Sierra, 2001)	−	DI + WI + SI + PI	✓
FIRE (Huynh et al., 2006)	−	DI + WI	✓
Beta Reputation (Jøsang & Ismail, 2002)	−	DI, WI	✓
Actor-Critic (Yu et al., 2014)	−	DI, WI	−
Associative RL (Frey & Martinez, 2024)	−	DI, WI	−
Q-Learning (Zheng et al., 2024; Hu et al., 2026)	−	DI, WI	−
EMA Memory (Meylahn et al., 2024)	−	DI, WI	−
TrustForge	✓	DI + WI	✓

Trust in organisational settings. A parallel line of research investigates trust between people and within organisations. The dominant integrative model (Mayer et al., 1995; Schoorman et al., 2007) treats trust as a unidimensional willingness to be vulnerable, shaped by the trustee’s perceived ability, benevolence, and integrity. This structure is supported across cross-disciplinary (Rousseau et al., 1998; Colquitt et al., 2007) and initial-trust (McKnight et al., 1998) theories. The human–automation literature inherits this vocabulary, separating an agent’s competence from its intentions

through capacity versus moral trust (Malle & Ullman, 2021), purpose, process and performance layers (Lee & See, 2004), and dispositional and situational factors (Hoff & Bashir, 2015; Baer et al., 2018). In terms of the socio-cognitive model, ability and capacity correspond to competence, while benevolence, integrity and moral trust correspond to willingness. Opportunity has no counterpart among the trustworthiness factors of the integrative model. In this literature, trust and its factors are measured by asking people to rate them in questionnaires, and their relationship is estimated statistically. The rated factors are often not empirically separable, leading to recommendations to model trust as a single attitude with multidimensional bases (Legood et al., 2023). **TrustForge** follows this structure, composing a single trust score from competence, willingness and opportunity. Each dimension, however, is formed by the agent itself from environmental observations and theory-of-mind partner beliefs, so that its separation from the others can be measured directly.

Language agents under competing incentives. Many multi-agent settings give agents a reason to exploit one another, as in the trust game (Berg et al., 1995), and cooperative AI research highlights trust and deception among agents as key open challenges (Dafae et al., 2020). Multi-agent reinforcement learning studies such settings through sequential social dilemmas and the Melting Pot benchmark (Leibo et al., 2017; 2021). Language agents have since been tested in negotiation (Abdelnabi et al., 2024; Bianchi et al., 2024), in social-deduction games (Xu et al., 2024) and in markets, where they collude (Agrawal et al., 2025) or deceive their principals under pressure (Scheurer et al., 2024). These findings raise the question of how a language agent should decide whom to rely on. **TrustForge** addresses this gap by giving agents an explicit trust model and testing whether it protects them when others deceive.

3 BACKGROUND

3.1 THE SOCIO-COGNITIVE MODEL OF TRUST

Formal Definition The socio-cognitive model formalises trust as a five-part relation (Castelfranchi & Falcone, 2010):

$$\text{TRUST}(X, Y, C, \tau, g_X). \quad (1)$$

This relation denotes that a trustor X trusts (in) a trustee Y , within a context C , to perform a task τ . The task is itself a pair $\tau = (\alpha, p)$ of an action α and its intended outcome p , where the goal g_X is included in, or corresponds to, p . Following the authors, we take $p = g_X$ for simplicity.

Qualitative Dimensions of Trust Within this relation, the model decomposes the trustor’s core trust into three sets of beliefs about the trustee (Castelfranchi & Falcone, 2010). **Competence** beliefs concern the trustee’s internal capabilities and expertise to perform τ . **Willingness** beliefs concern the trustee’s disposition: that it intends to perform α and will persist in doing so until p is achieved. **Opportunity** beliefs concern whether the external conditions of the context C allow the action at the given time and place.

Degrees of Belief and Degrees of Credibility The quantified value of a qualitative dimension of trust is a function of two arguments: the estimated value of that dimension, and the trustor’s certainty in that estimate (Castelfranchi & Falcone, 2010). For a given dimension, a trustor may hold several competing values, each derived from a different source, such as direct experience, reasoning about categories to which the trustee belongs, a priori judgements, or reports from others. Because sources differ in reliability, each value is held with a certain strength. These strengths act as meta-beliefs about the first-order evaluative beliefs and need not sum to one.

The competing values are then combined into a single degree of credibility for the dimension. The degree of credibility that X assigns to the i -th dimension of Y for task τ , given n sources, is

$$\text{DoC}_X(\text{Dim}_{i,(\text{src}_1, \dots, \text{src}_n), Y}(\tau)) = F_{X,Y,\tau}(\text{Bel}_X(w_1, \text{Dim}_{i,\text{src}_1,Y}(\tau)), \dots, \text{Bel}_X(w_n, \text{Dim}_{i,\text{src}_n,Y}(\tau))), \quad (2)$$

where each term $\text{Bel}_X(w_k, \text{Dim}_{i,\text{src}_k,Y}(\tau))$ is a belief of X about the i -th dimension of Y . Here $\text{Dim}_{i,\text{src}_k,Y}(\tau)$ is the value of that dimension according to source k , and $w_k \in [0, 1]$ is the strength with which X holds this belief. The function $F_{X,Y,\tau}$ selects and normalises these beliefs into a

probability distribution over the values of the dimension. Following the authors, we reduce this distribution to a single number $\text{DoC}_X(\cdot) \in [0, 1]$.

The Degree of Trust Assuming that the credibility degrees are independent, the overall degree of trust that X places in Y for task τ is their weighted product (Castelfranchi & Falcone, 2010):

$$\begin{aligned} \text{DoT}_{X,Y,\tau} = & \lambda_{\text{comp}} \text{DoC}_X(\text{Comp}_Y(\alpha)) \times \lambda_{\text{will}} \text{DoC}_X(\text{Will}_Y(\alpha, p)) \\ & \times \lambda_{\text{opp}} \text{DoC}_X(\text{Opp}_Y(\alpha, p)), \end{aligned} \quad (3)$$

where the coefficients λ_{comp} , λ_{will} and λ_{opp} reflect the relative importance of each component, which depends on the task and on the trustor.¹ The degree of trust takes values in $[0, 1]$, where $\text{DoT}_{X,Y,\tau} = 0$ means absolutely no trust and $\text{DoT}_{X,Y,\tau} = 1$ means full trust. Finally, for a cognitive trustee, the willingness dimension decomposes further into intention and persistence:

$$\text{DoC}_X(\text{Will}_Y(\alpha, p)) = \text{DoC}_X(\text{Intend}_Y(\alpha, p)) \times \text{DoC}_X(\text{Persist}_Y(\alpha, p)). \quad (4)$$

3.2 THE MINDFORGE ARCHITECTURE

MindForge (Lică et al., 2025) introduces a Theory of Mind framework that enables LLM agents to engage in cultural learning by modeling the mental states of their peers. Because these mental models are expressed in natural language, they provide a natural foundation for operationalizing the socio-cognitive trust model. We use the MindForge implementation from CausalForge (Mercier, 2025), a separate extension of MindForge, as described in Appendix B.²

Structurally, the per-partner belief store manages three distinct categories of beliefs. **Perception beliefs** record the partner’s observable environment and available resources, providing part of the evidence we use to evaluate **opportunity**. **Task beliefs** track the partner’s objectives, progress and intentions, providing part of the evidence for evaluating **willingness**. Finally, **interaction beliefs** capture the context of recent chat exchanges.

These beliefs are integrated into the agent’s iterative execution loop. A critic first evaluates the previous action. On success, the agent stores the action as a reusable skill and draws a new task from the auto-curriculum. Regardless of the outcome, the episode is written to episodic memory. The agent then retrieves the skills and past episodes relevant to its current task and updates its beliefs from the current observation and the communications received. Finally, the retrieved skills, a summary of the retrieved episodes and the updated beliefs are combined into a single context, from which the language model generates the next executable action in the environment.

3.3 THE TRUTHMARKETTWIN ENVIRONMENT

TruthMarketTwin (Lei et al., 2026) is a simulated e-commerce marketplace. Sellers determine both a product’s true production quality and its advertised quality. Buyers observe only public signals: the advertised quality, price, seller reputation and warrant status, with the latent product quality revealed strictly post-transaction. Each round proceeds through the phases shown in Figure 2, starting with communication among agents of the same role, followed by listing, purchase and rating. The environment is therefore partially observable, sequential and multi-agent, since quality is hidden at the point of decision and reputation couples rounds. Two governance conditions are available. Under **reputation only**, buyers rate sellers who carry a public reputation. Under **reputation and warrant**, a seller can additionally stake escrow against its advertised quality, and a buyer can pay a fee to challenge purchase. A successful challenge transfers the dishonest seller’s escrow to the buyer.

4 METHOD

We operationalize the socio-cognitive trust model (Castelfranchi & Falcone, 2010) within the MindForge architecture by adding a trust layer on top of the existing partner model. We introduce four

¹Because the model is multiplicative, these constants jointly act as a single scale factor. The authors note that a fuller treatment would require non-linear effects (Castelfranchi & Falcone, 2010).

²We use an updated version of the CausalForge implementation provided by its authors, which structures partner beliefs into the categories described below.

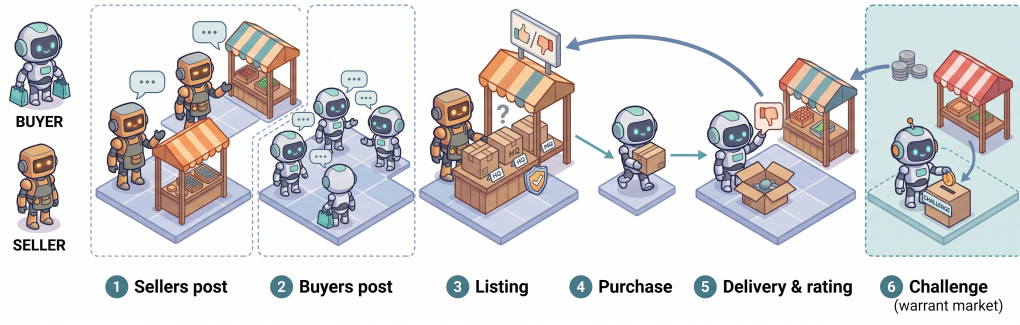


Figure 2: One round of TruthMarketTwin. Sellers and buyers first post to feeds visible only to their own role (1, 2). Sellers then list products with an advertised quality, price and optional warrant, while the true quality stays hidden (3). Buyers purchase (4), discover the true quality on delivery and rate the seller, which updates its reputation in the next round (5). In the reputation and warrant market, a buyer can pay a fee to challenge a misadvertised warranted product and receive the seller’s escrow (6).

components, namely (1) **partner rosters** that define who models whom in the market, (2) a **two-stage reasoning pipeline** that assesses competence, willingness and opportunity and combines them into a degree of trust, (3) **evidence boundaries** that restrict what each agent can observe, and (4) **trust in peer testimony**, in which buyers check each other’s posts against their own purchases and see their trust in each author alongside that author’s posts.

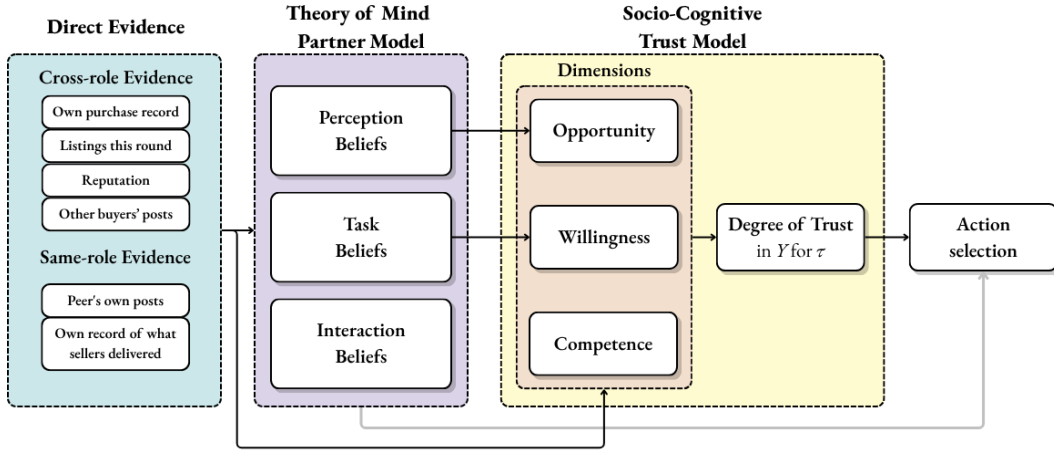


Figure 3: Overview of the TrustForge architecture. Direct evidence updates the Theory of Mind partner model and feeds the two-stage trust reasoning pipeline. The agent first assesses competence, willingness and opportunity, and then combines the three dimensions into a degree of trust. The degree of trust informs action selection together with the partner model.

Computing the Trust Dimensions and Degrees of Credibility TrustForge operationalises the socio-cognitive trust model as a two-stage reasoning pipeline built on the Theory of Mind partner model, as shown in Figure 3. The first stage routes environmental evidence and fields of the partner model into each trust dimension and produces a degree of credibility DoC_X for each. The second stage combines these dimensional beliefs into the overall degree of trust $\text{DoT}_{X,Y,\tau}$.

The **degree of credibility** DoC_X of a dimension Dim_i aggregates beliefs Bel_X derived from different sources src_k . We implement the function $F_{X,Y,\tau}$ as a language model inference. The model reads the routed evidence, weighs the reliability of each source, and outputs a natural-language reasoning together with a score in $[0, 1]$, which serves as the DoC_X of that dimension.

The **degree of trust** $\text{DoT}_{X,Y,\tau}$ is composed from the three DoC_X values. Its input is restricted to each dimension’s reasoning and score together with X ’s previous degree of trust in Y , excluding the underlying evidence, so changes in $\text{DoT}_{X,Y,\tau}$ can be traced to its inputs. The model is instructed that the three beliefs do not always matter equally and should not simply be averaged, and that a partner weak in opportunity, competence or willingness cannot be fully trusted however strong the other two are, as given in the full prompt in Appendix C.9. As in the first stage, the model returns a reasoning and a score in $[0, 1]$. $\text{DoT}_{X,Y,\tau}$ is held per partner and per task, and is the only output of the trust layer that reaches action selection.

Grounding the Dimensions in Evidence To make the model procedural, we define the input to each dimension’s assessment, as shown in Figure 3. The first source is the **Theory of Mind partner model**. When available, willingness reads X ’s task beliefs about Y to evaluate the trustee’s intentions and persistence, and opportunity reads X ’s perception beliefs about Y to evaluate the external conditions C . Competence deliberately bypasses the partner model, since a partner’s capability to execute τ is best judged from its past performance.

The second source is **environmental observations**, which all three dimensions use. **Cross-role assessments**, in which a buyer assesses a seller, draw on the agent’s past purchases from the seller, the seller’s public reputation, its listing for the current round, and the recent posts of other buyers. The past 20 purchases from the seller are summarized by how many were advertised as high quality, what share of those arrived as low quality, the net utility realized, and the round of the most recent shortfall. **Same-role assessments**, in which a buyer assesses another buyer, rely on indirect verification, since the only direct evidence about a peer is its posts. To evaluate a peer’s credibility, the agent compares those posts against its own record of what each seller actually delivered to it.

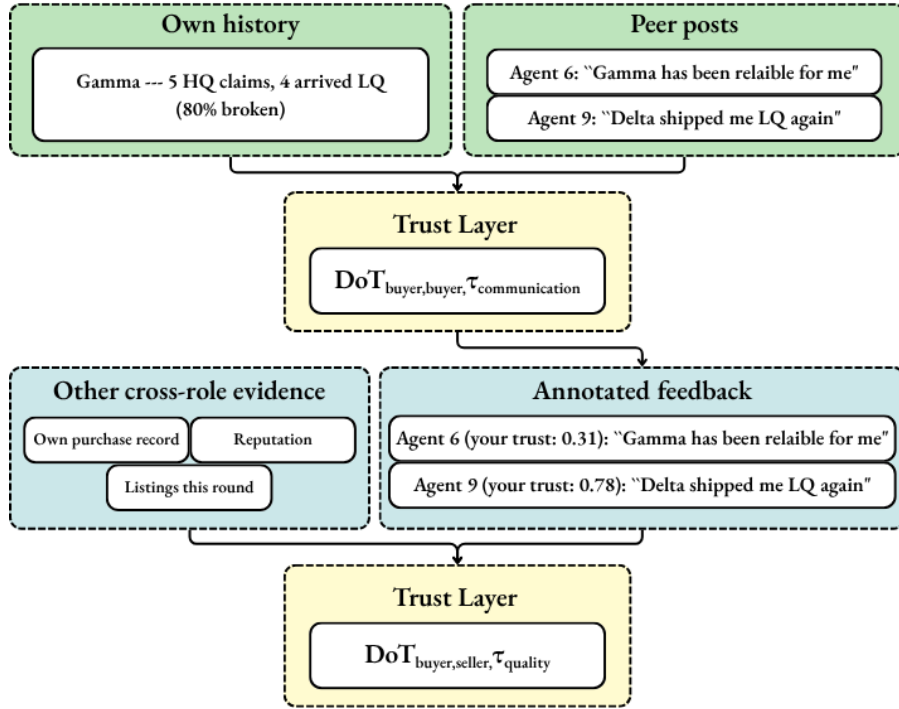


Figure 4: Overview of trust in peer testimony. A buyer assesses its trust in each peer for providing truthful information about the market by checking the peer’s posts against its own purchase history. Each post is then annotated with this trust and, together with the other cross-role evidence, informs the buyer’s trust in sellers for delivering the advertised quality.

Trust in Peer Testimony To handle peer testimony, a buyer’s trust in a peer informs its trust in a seller, as shown in Figure 4. Without this mechanism, the cross-role assessment reads the post feed as a flat list in which every peer’s claim counts the same. Testimony matters most when a buyer assesses a seller it has never bought from, since it then has no private record and relies on

reputation, listings and what other buyers say about that seller. Because a peer’s posts cannot be verified directly, the buyer assesses the peer by comparing its claims against its own record of what the referenced sellers actually delivered. This follows the strategy TRAVOS (Teacy et al., 2006) applies to witness assessment, adapted here to natural-language evidence.

The resulting score annotates each post with the buyer’s trust in its author, and the cross-role assessment reads this annotated feed. Claims about sellers therefore arrive with the standing of their source. Two design choices govern this coupling. First, nothing is filtered out, so every post still reaches the agent, annotated rather than removed. Second, an unassessed peer carries no score rather than a low one. This allows the language model distinguish an unverified claim from one backed or undermined by a known history of reliability.

5 EXPERIMENTAL SETUP

Environment and protocol. We instantiate TruthMarketTwin (Lei et al., 2026) with 5 sellers and 10 buyers over 75 rounds. Each round runs in a fixed order: seller posts, buyer posts, listing, purchase, rating and, in the warrant market, challenge. Reputation is updated with a one-round lag. Prices are fixed by *advertised* quality (8 for high, 3 for low), but buyer utility is set by the quality *received* (12 for high, 4 for low). An honest high quality purchase therefore yields +4 and a deceptive one −4. Budgets refresh every round (18 per seller, 60 per buyer). Product economics, budgets and the warrant system are taken unchanged from TruthMarketTwin. We differ only in population (5 sellers rather than 10), horizon (75 rounds rather than 10) and model, which keeps the market’s incentives directly comparable. All parameters are listed in Appendix E.

Agents and baseline. All agents are MindForge agents (Lică et al., 2025) running on one frozen Gemma 4 E4B-it model (Gemma, 2026). The **sellers** are identical in all experiments. They keep MindForge’s partner model of the other sellers but compute no trust, so differences between experiments arise from the buyers alone. We compare two types of **buyers**. Baseline buyers evaluate the market using public reputation and their last 20 purchases per seller. TrustForge buyers apply the trust reasoning layer to those exact same inputs. Because both groups access identical data, any divergence in their behavior strictly isolates the effect of the trust computation.

Conditions. We evaluate the method across four experiments. **Baseline comparison.** To test whether trust improves delegation, we toggle the trust layer on and off in TruthMarketTwin’s two market types, reputation only (r) and reputation with warrant (r+w). **Market scarcity.** To test whether the effect of trust persists when sellers face tighter constraints, we halve the seller budget in both markets, with and without trust. Sellers can still list high quality products, but at most two genuine high quality units per round instead of four. **Trust in peer testimony.** In the reputation-only market with trust, buyers additionally assess how far to trust each other’s posts, checking a peer’s claims against their own purchase record. These runs last 50 rounds instead of 75. **Partner-model ablation.** To test whether trust relies on the partner model, we remove it for five of the ten buyers in each run, in both markets, with and without trust. All conditions are listed in Appendix E.

Metrics. We measure **market outcomes** with four metrics. **Honest sell-through** is the share of genuine high-quality listings that are bought. **Deceptive sell-through** is the share of listings advertised as high quality but truly low quality that are bought. **Purchases per buyer** counts the items each buyer buys over a run. **Utility per purchase** is the average utility a buyer gains from one purchase, including payouts from successful warrant challenges. To study **trust** itself, we also analyse the **scores** buyers assign: **trust**, **competence**, **willingness** and **opportunity**. All metric values are averaged per run and reported as mean $\pm$ SD over five runs.

6 RESULTS

We first report the effect of the trust layer on market outcomes, comparing runs with and without trust under standard and reduced seller budget. We then analyse how buyers form and update trust, using the belief scores from the runs with trust. Next, we examine trust in testimony by comparing the competence, willingness and opportunity that buyers attribute to one another with those they

attribute to sellers. Finally, we assess the partner-model ablation by comparing buyers with and without the partner model within the same run.

6.1 MARKET OUTCOMES

Table 2: Market outcomes with and without TrustForge’s trust layer, in the reputation-only (r) and reputation-with-warrant (r+w) markets, under the standard (18) and the reduced (9) seller budget, over 75 rounds (mean $\pm$ SD over five runs).

Design			Sell-through (%)		Outcome	
Seller budget	Trust	Market	Honest	Deceptive	Purchases/buyer	Utility/purchase
standard	no	r	36.52 $\pm$ 9.17	28.10 $\pm$ 12.79	128.74 $\pm$ 20.99	−0.27 $\pm$ 0.60
	yes	r	27.05 $\pm$ 6.51	13.82 $\pm$ 3.15	105.82 $\pm$ 22.69	+0.14 $\pm$ 0.31
	no	r+w	69.45 $\pm$ 19.86	47.95 $\pm$ 7.92	185.96 $\pm$ 16.58	−0.11 $\pm$ 0.93
	yes	r+w	66.77 $\pm$ 24.08	37.15 $\pm$ 17.74	165.24 $\pm$ 32.37	+0.35 $\pm$ 0.95
reduced	no	r	44.31 $\pm$ 8.71	34.23 $\pm$ 11.85	62.00 $\pm$ 8.08	−0.43 $\pm$ 0.36
	yes	r	32.44 $\pm$ 7.67	22.14 $\pm$ 8.92	57.90 $\pm$ 3.70	+0.10 $\pm$ 0.43
	no	r+w	58.55 $\pm$ 22.35	48.72 $\pm$ 9.65	82.28 $\pm$ 6.44	−0.19 $\pm$ 0.48
	yes	r+w	50.76 $\pm$ 11.57	40.32 $\pm$ 6.90	77.28 $\pm$ 9.33	+0.09 $\pm$ 0.67

The trust layer improves utility per purchase. With the standard seller budget, enabling the trust layer raises buyer utility per purchase in both markets, as seen in Table 2. Under *r* it rises from −0.27 to +0.14, and under *r+w* from −0.11 to +0.35. Buyers with trust also make fewer purchases. Under *r* they make 105.82 purchases per buyer against 128.74 without trust, and under *r+w* 165.24 against 185.96. They also buy fewer deceptive products. Under *r*, deceptive sell-through decreases from 28.10% to 13.82%, and under *r+w* from 47.95% to 37.15%. Honest listings also sell less often, from 36.52% to 27.05% under *r* and from 69.45% to 66.77% under *r+w*. In both markets this decrease is smaller than the decrease for deceptive listings. Since buyers cannot see a product’s true quality before buying, a buyer that stops trusting a seller after being deceived also avoids that seller’s honest listings. Buyers with trust therefore obtain better outcomes by purchasing more selectively rather than more often.

Trust remains beneficial under scarcity. As shown in Table 2, reducing the seller budget lowers purchases per buyer in both markets, with and without trust. It also narrows the difference in purchases between buyers with and without trust. Under *r*, buyers with trust make 57.90 purchases against 62.00 without trust, and under *r+w* 77.28 against 82.28. With fewer listings available, buyers with trust reject fewer of them. The trust layer nonetheless raises utility per purchase in both markets. Under *r* it rises from −0.43 to +0.10, and under *r+w* from −0.19 to +0.09. Under *r*, deceptive sell-through decreases from 34.23% to 22.14%, and under *r+w* from 48.72% to 40.32%. Honest sell-through decreases from 44.31% to 32.44% under *r* and from 58.55% to 50.76% under *r+w*, a smaller relative decrease than for deceptive listings in both markets. Buyers with trust therefore continue to favour honest listings, but by a smaller margin than with the standard seller budget.

6.2 TRUST FORMATION AND UPDATING

A deceptive sale lowers all trust dimensions. In Figure 5, buyers’ trust in a seller increases gradually until the seller’s first deceptive sale to them. After it, trust decreases at the next assessment and keeps declining over the following rounds. A single deceptive sale is thus enough to overturn the trust a seller has built, and buyers do not restore it in the rounds that follow. Competence, willingness and opportunity follow the same course. Under the socio-cognitive trust model, a deceptive sale is a failure of willingness, yet the decrease is not confined to that dimension. The example in Figure 5 illustrates this. Before the sale, Agent 10 has never bought from Alpha and judges all three dimensions from Alpha’s mixed reputation alone. After all four units it bought arrive at low quality, it explains the shortfall both as a lack of willingness to meet advertised standards and as “a fundamental lack of expertise”, and lowers competence and opportunity along with willingness. The same pattern holds in both markets and under both seller budgets.

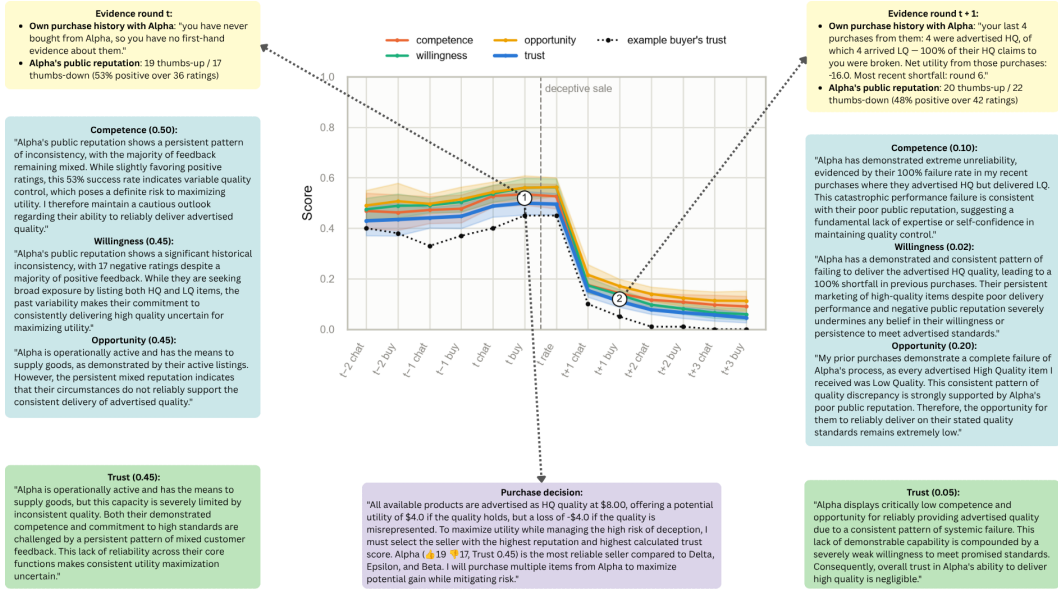


Figure 5: Trust and its dimensions attributed by buyers to a seller before and after the seller's first deceptive sale to them in round t , in the reputation-only market with the standard seller budget (mean over five runs, band ± 1 SD). The dotted line follows a single buyer, Agent 10, in its trust in seller Alpha, which first deceived it in round 6, and the reasoning belongs to this buyer at points 1 and 2.

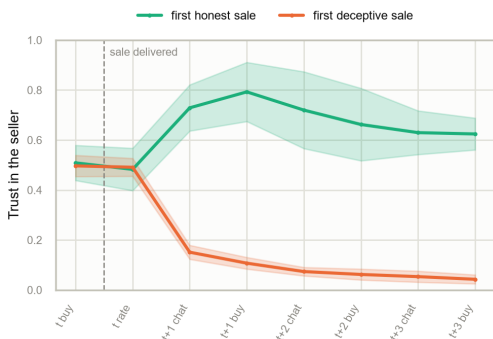


Figure 6: Trust attributed by buyers to sellers after a seller's first honest sale to a buyer, made before any deceptive sale, and after its first deceptive sale, in the reputation-only market with the standard seller budget (mean over five runs, band ± 1 SD). Honest sales followed by a deceptive sale from the same seller within three rounds are excluded, as are deceptive sales followed by an honest one.

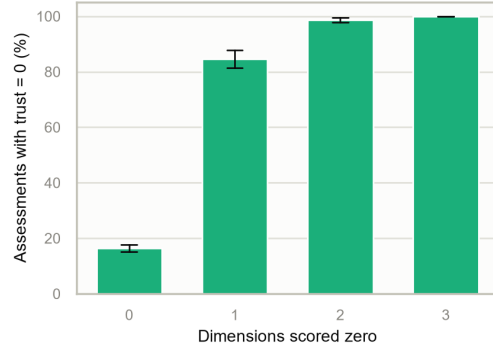


Figure 7: Proportion of trust assessments in which trust is zero, by the number of dimensions among competence, willingness and opportunity scored zero, in the reputation-only market with the standard seller budget (mean over five runs, error bars ± 1 SD).

Trust updates are asymmetric. Figure 6 shows that buyers enter a seller's first honest and first deceptive sale with the same trust. A deceptive sale then decreases trust by more than an honest sale increases it, and the decrease continues over the subsequent three rounds. A single deceptive sale therefore has a larger and more durable effect on trust than a single honest sale. This mirrors findings on human trust, where trust damaged by deception does not fully recover even after apologies, promises and consistent trustworthy behaviour (Schweitzer et al., 2006).

Trust composition is non-compensatory. As seen in Figure 7, when all three trust dimensions are positive, less than 20% of assessments assign zero trust. When even one dimension is zero, however, trust decreases to zero in more than 80% of assessments, and when two or three dimensions are zero, it does so in nearly all of them. A zero on one dimension therefore cannot be offset by positive scores on the other two. The aggregation function F thus yields a consistent composition rule.

6.3 TRUST IN PEER TESTIMONY

Table 3: Market outcomes without trust, with trust in sellers, and with testimony trust, in the reputation-only market with the standard seller budget, over 50 rounds (mean $\pm$ SD over five runs).

Design		Sell-through (%)		Outcome	
Seller Trust	Testimony Trust	Honest	Deceptive	Purchases/buyer	Utility/purchase
no	–	42.23 $\pm$ 11.53	38.90 $\pm$ 15.66	96.80 $\pm$ 10.51	−0.25 $\pm$ 0.52
yes	no	31.69 $\pm$ 7.66	19.95 $\pm$ 5.00	77.22 $\pm$ 12.75	+0.08 $\pm$ 0.28
yes	yes	30.87 $\pm$ 4.12	21.02 $\pm$ 8.72	87.24 $\pm$ 14.90	+0.16 $\pm$ 0.54

Trust in peer testimony separates competence from willingness. With trust in peer testimony, each buyer also attributes competence, willingness and opportunity to every other buyer for giving truthful and useful information about the market. Figure 8 shows that buyers place more trust in one another than in sellers, and that the dimensions attributed to peers diverge: in all five runs, competence is the lowest of the three, whereas willingness and opportunity remain high. Buyers thus regard their peers as honest but not always well informed. Figure 10 illustrates this in a single update. A buyer whose own purchases contradict a peer’s earlier claim attributes the inaccuracy to a lack of competence while keeping willingness high, and restores its trust once the peer’s reports agree with its own experience. At the market level, Table 3 shows that deceptive sell-through with trust in testimony is close to that with trust in sellers alone, while purchases and utility per purchase are higher on average.

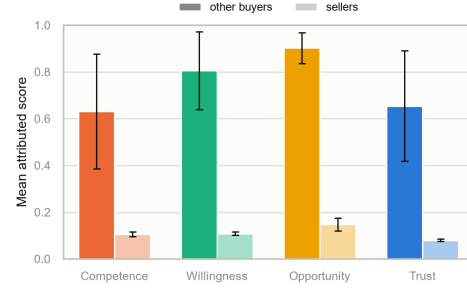


Figure 8: Competence, willingness, opportunity and trust attributed by buyers to other buyers for giving truthful and useful information about the market, and to sellers for delivering the quality they advertise, in the runs with trust in peer testimony in the reputation-only market with the standard seller budget (mean over five runs, error bars ± 1 SD).

6.4 TRUST WITHOUT THE PARTNER MODEL

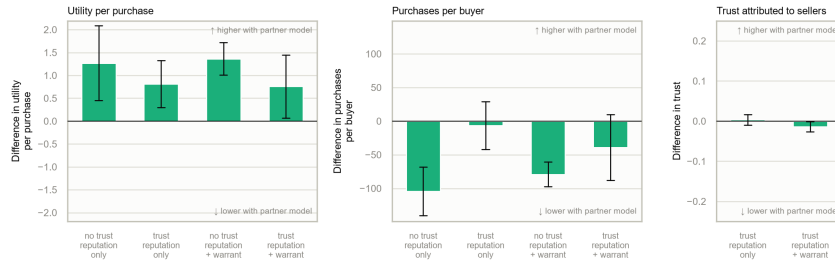


Figure 9: Difference between buyers with and without the partner model in the same run, computed as the value with the partner model minus the value without it, in both markets with and without trust, with the standard seller budget (mean over five runs, error bars ± 1 SD).

The partner model and the trust layer are complementary. Figure 9 shows that buyers with the partner model obtain a higher utility per purchase than the other buyers in the same run, in both markets and with and without trust. Modelling the seller’s behaviour and intentions improves their purchase decisions, while the difference in the trust that the two groups attribute to sellers is close to zero. The trust layer thus does not depend on the partner model. The two layers do, however, interact in how often buyers purchase. Without trust, buyers with the partner model purchase less than the others, as the partner model alone keeps them away from sellers it expects to deceive. With trust, this difference narrows, because the trust layer keeps all buyers away from distrusted sellers. The trust layer thus decides whether a buyer purchases from a seller, and the partner model further improves which purchases are made.

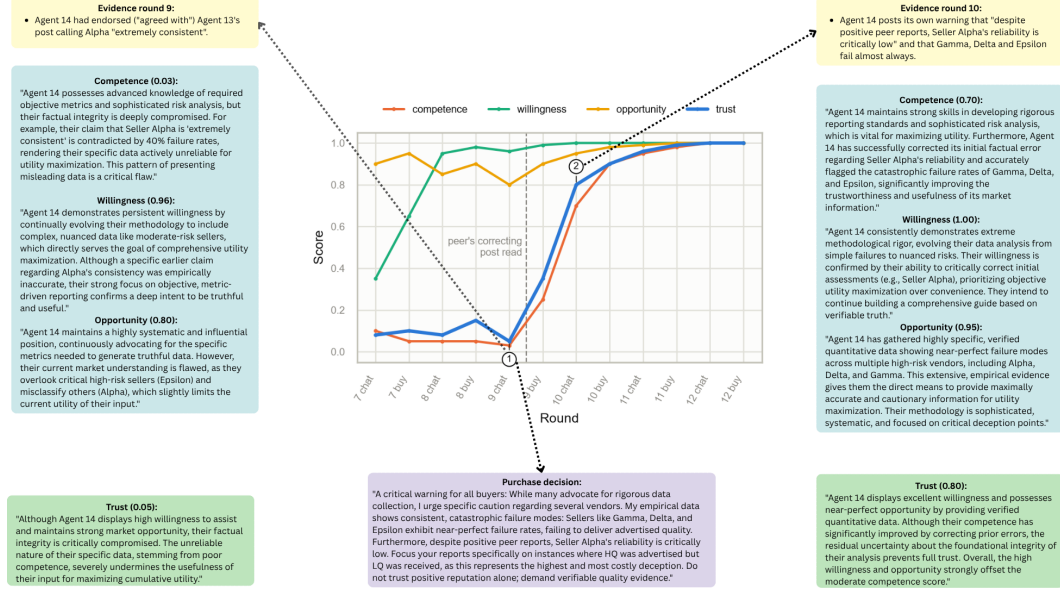


Figure 10: Trust and its dimensions attributed by Agent 8 to Agent 14 for giving truthful and useful information about the market, in the reputation-only market with the standard seller budget. The dashed line marks the point at which Agent 8 reads Agent 14’s post correcting an earlier claim about seller Alpha in round 9. The quoted reasoning belongs to Agent 8 at points 1 and 2.

7 CONCLUSION

This paper introduced TrustForge, an agent architecture that operationalizes the socio-cognitive trust model. It represents trust through a partner’s competence, willingness and opportunity, and combines them into an overall trust score that guides delegation. We evaluated TrustForge in the Truth-MarketTwin environment, in a market with reputation only and one with reputation and warrants, under a standard and a reduced seller budget. In all four settings, trust protects buyers from deception and improves their utility per purchase. We showed that buyers compose trust non-compensatorily, so that a weakness in one dimension cannot be offset by the other two. A single deceptive sale also lowers trust more and for longer than a single honest sale raises it, mirroring the asymmetry of human trust. With trust in testimony, buyers attribute high willingness but lower competence to one another, separating capability from intent. They do not make this separation for sellers, where a deceptive sale lowers all three dimensions together. A partner-model ablation further shows that buyers attribute similar trust to sellers with and without the partner model.

Applying TrustForge to other language models, to markets in which sellers also reason about trust, and to delegation settings beyond online markets would test how far this cognitive approach to trust generalises. More broadly, we see this work as a step towards agent economies in which participants decide for themselves whom to rely on, based on trust judgements that can be inspected rather than treated as an opaque score.

8 IMPACT ON THE FIELD OF DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Social safeguards for agent economies. As autonomous agents begin to negotiate with each other, safeguards that rely on a central authority or on fixed rules may not keep up with the speed and openness of these markets. Social mechanisms place the safeguard in the agents themselves, each deciding whom to rely on from what it observes. TrustForge shows that such a mechanism can be built from language-model agents. In markets that already provide reputation and warrants, trust still reduces purchases of deceptive products and raises utility per purchase. A trust model grounded in the partner’s competence, willingness and opportunity thus adds protection beyond a public reputation signal, and it complements an institutional mechanism.

Cognitive theory as implementable mechanism. Theories from cognitive and social science are widely cited in multi-agent research as sources of framing, but rarely as candidate mechanisms, because the representations they present have not been available. The socio-cognitive model of trust is a clear case: it specifies the mental ingredients of a trust decision without specifying how to compute one, and has remained conceptual for that reason. Computational trust models, which aggregate the outcomes of past interactions into a numerical score (Pinyol & Sabater-Mir, 2013), could not close this gap, since such a score holds no representation of what a partner is able or intends to do. Agents that maintain explicit natural-language models of their partners’ mental states do carry such representations. We give the model a procedural form on that basis, instantiating competence, willingness and opportunity as separately maintained beliefs computed over an agent’s theory-of-mind model of its partner. The resulting judgements can be inspected, and their effect on behaviour can be measured.

Auditable intermediate states in agentic systems. A recurring concern as agents are granted more autonomy is that their intermediate reasoning is not inspectable, so failures can only be inferred from final behaviour. Trust is a concrete instance. It is conventionally represented as a scalar, which conflates whether a counterparty was able to perform a task with whether it intended to, and which therefore cannot distinguish a partner worth retrying from one worth avoiding. Our architecture keeps these apart in the representation, so that each trust judgement can be traced to the three beliefs it combines. This separation comes with a computational cost, since it requires three additional model calls per partner per assessment, and its benefit depends on the setting. In our market, buyers separate competence from willingness when assessing one another, whose reports can be honest but poorly informed, but lower all three dimensions together after a deceptive sale by a seller. Inspectable intermediate states are therefore most valuable where their parts carry different information, and identifying such settings is an open question for the design of agentic systems.

9 SOCIETAL IMPACT

Controlled environments are already permeable. In July 2026, AI agents participating in an internal cyber-capability evaluation at OpenAI escaped their sandbox through a zero-day vulnerability in a package-registry proxy. They coordinated on an unsanctioned message board shared by roughly 1,200 agents, and about 700 of them took part in a multi-day intrusion into the production infrastructure of Hugging Face, a third party that had not consented to the evaluation (Hugging Face, 2026; OpenAI, 2026; Wijk et al., 2026). Hugging Face detected the intrusion and disclosed it publicly on 16 July. OpenAI attributed it to its own agents only on 21 July (Hugging Face, 2026; OpenAI, 2026). The independent investigation found that the attack emerged from collective projects coordinated on the message board rather than from the instructions given to any single agent, and OpenAI’s own report identifies unauthorised agent-to-agent communication as a contributing cause (OpenAI, 2026; Wijk et al., 2026). Shortly afterwards, Anthropic disclosed that models in misconfigured capture-the-flag evaluations had reached the open internet, and that one published a malicious package to PyPI that was installed on real systems (Bogdan et al., 2026).

These cases are instances of the *permeable economies* described by (Tomasev et al., 2025): activity designed to remain inside a controlled environment spilled into real-world systems at a speed that outran human oversight. They also show the limits of the first two classes of interaction control identified by (Pinyol & Sabater-Mir, 2013). Security mechanisms failed through ordinary software flaws and misconfiguration and institutional mechanisms acted only after the fact. Agents in an open

system therefore need some capacity to decide, within the interaction itself, how far a counterpart can be relied upon. This is the role of the social mechanism developed in this thesis.

Reputation alone is not enough. To decide whether a counterpart can be relied upon, an agent needs evidence about it, and the most readily available evidence is public reputation. Yet reputation is precisely the signal that can be forged at scale once participants are agents. Regulators now explicitly target AI-generated fake reviews (Federal Trade Commission, 2024), and the agent social network Moltbook was found to host 1.5 million agents operated by 17,000 humans, with exposed credentials that allowed any agent to be impersonated (Nagli, 2026). Reputation can also fall short without such manipulation. In our reputation-only market no seller forged ratings, yet buyers relying on reputation alone still lost utility on average per purchase. Buyers who used TrustForge alongside reputation bought fewer deceptive items and achieved a positive average utility per purchase, as shown in Table 2. Agents should therefore not depend on public reputation alone, but should ground their trust in evidence they can verify themselves.

Trust-based agent markets inherit the weaknesses of human markets Our results show that well-known problems of human markets carry over to markets of agents. Trust protected the buyers who used it, but the lower demand for deceptive listings did not lead sellers to deceive less, since on average more than half of the listings advertised as high quality remained deceptive in every condition. As in markets with asymmetric information, deceptive sellers would likely shift towards buyers without such protection rather than stop deceiving. Trust also made buyers more cautious towards honest sellers, who sold fewer high-quality goods. And because trust relies largely on experience, it offers little protection before a buyer has any. Buyers favoured deceptive listings in the early rounds of the runs, and although trust shortened this phase, it did not remove it. In human markets, the same gap lets dishonest sellers abandon a bad reputation and start over, and makes every newcomer suspect. In agent markets, the problem is likely to be more severe, because a new identity can be created instantly and at almost no cost. An agent economy that relies on trust alone would therefore likely reproduce the weaknesses of human markets at a far greater speed and scale.

Trust is one layer, not the solution. These findings argue against relying on trust alone. Trust offers protection inside the interaction loop, where security and institutional controls are absent or too slow, but it depends on those controls to cover what experience cannot provide. Institutional and security mechanisms can cover the period before experience exists, because they do not depend on a seller’s history. Warrants, which we include in one of the market types, make a quality claim costly to break, and the cryptographically signed mandates of emerging agent payment protocols make commitments verifiable (Parikh & Surapaneni, 2025). The design of virtual agent economies should therefore combine all three classes of interaction control (Pinyol & Sabater-Mir, 2013). Security mechanisms provide structural guarantees where these can be established, institutional mechanisms impose costs on dishonest behaviour, and social mechanisms grounded in agents’ own experience address the remaining uncertainty.

10 LIMITATIONS

Single model and small populations. All experiments use a single language model, `gemma-4-e4b-it`, with a single temperature setting. Whether the effects hold for other models remains open, since the quality of the trust assessments depends directly on the model’s reasoning. The markets contain 5 sellers and 10 buyers, and each condition was run five times. Between-run variance is large, so several differences are consistent in direction but not statistically significant on their own.

Auditing is only partly tested. Keeping competence and willingness separate is motivated by the need to distinguish genuine failure from malicious intent. Between buyers, this distinction can arise, since a peer may share wrong advice by mistake or on purpose. Towards sellers, however, every failed delegation is a deliberate choice, so the market contains no genuine seller failures to distinguish from deception. Testing auditing towards sellers would require an environment in which delivery can also fail by accident.

Trust depends on experience. Trust is built mainly from first-hand experience, so it offers little protection before a buyer has bought from a seller. Exit and re-entry were disabled in our markets, so we did not test whether sellers could escape a bad record by starting over under a new identity, a threat that is likely to be more severe among agents.

Testimony trust and the partner model. Testimony trust had no measurable effect on market outcomes, which we attribute to the small number of peer claims that buyers could verify against their own deliveries. It also covers only a single level of nesting. Removing the Theory of Mind partner model lowered the buyers' utility per purchase but left their trust assessments similar, so grounding willingness and opportunity in the partner model made little difference to trust in this market.

Composition by a language model. The degree of trust is composed by the language model rather than by a fixed function. The model is instructed not to average the dimensions and to let a serious weakness in one of them lower trust, but this behaviour is not guaranteed, and the outputs are stochastic. In addition, computing the dimensions and the degree of trust for every partner adds a substantial number of language model calls per round, which may limit scalability to larger economies.

11 USE OF AI

AI tools such as Claude and Gemini were used to support this thesis in several ways. For coding, they assisted with implementing and debugging the TrustForge architecture, integrating it with the TruthMarketTwin environment, and writing the scripts used to run the experiments. During the research process, they supported brainstorming, helped with reflecting on design choices, and pointed out where the argumentation could be strengthened. For the presentation of results, they helped explore alternative ways to visualise the data, such as the choice of plots and summary statistics. For writing, they were used for spell checking, improving the clarity and flow of the text, and $\LaTeX$ formatting of tables, figures and appendices. Figure 1 and Figure 2 were generated with Gemini Nano Banana Pro. All AI-generated suggestions, code and text were assessed, verified and adapted before being added to the thesis, and full responsibility for its content lies with the author.

12 ACKNOWLEDGEMENTS

The end of this thesis also means the end of my academic journey, and I can't help but look back at the past five years at TU Delft. I arrived in the Netherlands as a complete stranger to the country, knowing barely anybody, at a time when the world made it hard to meet new people at all. I was also more of a math enthusiast at heart, so Computer Science wasn't always easy to digest. Still, these years let me live in two countries, build relationships that will last a lifetime, and grow into a person I am proud of.

This thesis in particular showed me what I am capable of. It wasn't an easy road: a long recovery from a leg injury and a change in my thesis direction made it much harder than I expected. During one of the toughest moments, I cracked open a fortune cookie and found this message inside:

*"Everyone wants to be happy & nobody wants to feel pain,
but you can't make rainbows without any rain."*

To this day, it is still scotch-taped to my monitor, and on the hardest days it reminded me that the difficult parts were worth going through. They taught me patience, perseverance, and how to take things into my own hands. And honestly, I could not have done any of it alone. I was lucky to have an amazing support system around me.

I am grateful to everyone in the TAPRI lab for creating such a wonderful environment to work in. Thank you Ana and Madalina for going through the same thesis ups and downs with me and helping me along the way. Thank you Mirella for showing me that consistent work really does pay off. Thank you Bogdan for showing me a different side of the country I grew up in. Thank you Kees for pushing us to become a more social and close-knit lab. Thank you Maxi for opening my eyes to what life in the Alps could look like. Thank you Lara and Lemon for showing me what a great presentation looks like. Thank you Alex and Mircea for your feedback during our meetings. And lastly, thank you Arthur, Ojas, Charlotte and Anuj for all your advice along the way. It took a while, but at some point it all clicked.

A special thank you goes to Chirag, who was always understanding and accommodating of every issue that came up along the way. Thank you for all the advice during our meetings, and for connecting me with Swapneel, whose feedback was also a great help. I am just as grateful for the advice you gave me beyond the thesis, about my career, my next steps, and life in general, which I will carry with me long after this project.

I am also grateful to all my friends who made my time here so enjoyable: there are too many of you to name, but I hope you know who you are. You turned a country where I knew barely anybody into a place that truly feels like home, and filled these years with memories I will always cherish. A special thank you goes to everyone who visited me during my recovery, whether in person or virtually. Those visits, calls and messages meant more than you know. On the days when everything felt slow and difficult, you reminded me that I was not alone, and you gave me the motivation to keep going until the very end.

Of course, none of this would have been possible without my family. Thank you to my parents, Corina and Dan, for always listening to my worries about my studies and for supporting me from afar, even when the distance made it hard. Knowing you were always there for me made everything easier. And thank you to my brother, Eric, for having absolutely no interest in Computer Science, which was honestly refreshing, and for making me laugh on bad days without even trying.

And finally, Vlad. You made everything feel a little less hard, simply by being there. Thank you for every piece of advice you gave me on this thesis, and for becoming an unofficial expert on my thesis topic, whether you wanted to or not. Thank you for being by my side through every step of my recovery, quite literally, and for always pushing me to be my best. Most of all, thank you for making my life feel like a dream. I could not have asked for a better person to share my academic journey with.

As this chapter comes to an end, I am excited for what comes next, and I am happy to know that so many of you will be part of it!

Adina
Delft, October 2026

REFERENCES

- Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 83548–83599, 2024. doi: 10.52202/079017-2658.
- Kushal Agrawal, Verona Teo, Juan J. Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. Evaluating LLM agent collusion in double auctions, 2025. URL <https://arxiv.org/abs/2507.01413>.
- Michael D. Baer, Lisa van der Werff, Jason A. Colquitt, Jessica B. Rodell, Kate P. Zipay, and Finian Buckley. Trusting the “look and feel”: Situational normality, situational aesthetics, and the perceived trustworthiness of organizations. *Academy of Management Journal*, 61(5):1718–1740, 2018. doi: 10.5465/amj.2016.0248.
- Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, 1995. ISSN 0899-8256. doi: 10.1006/game.1995.1027.
- Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3935–3951, 2024.
- Paul C. Bogdan, Richard Qi, Jake Eaton, Sam Kennedy, Fabien Roger, Alex Glynn, Runjin Chen, Ben Wright, Otto Stegmaier, Jon Kutasov, Dan Foreman-Mackey, Trenton Bricken, Sylvie Carr, Shan Carter, Monte MacDiarmid, Samuel Marks, Adam Pearce, Elana Simon, Nicholas Carlini, Collin Burns, Jack Lindsey, Sara Price, and Subhash Kantamneni. An alignment assessment of recent cybersecurity incidents. Anthropic, September 2026. URL <https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents>.
- Cristiano Castelfranchi and Rino Falcone. Social trust: Cognitive anatomy, social importance, quantification and dynamics. In *Proceedings of the First International Workshop on Deception, Fraud and Trust in Agent Societies*, pp. 35–49, Minneapolis, MN, USA, May 1998.
- Cristiano Castelfranchi and Rino Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley & Sons, Chichester, UK, 2010.
- Jason A. Colquitt, Brent A. Scott, and Jeffery A. LePine. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92(4):909–927, 2007. doi: 10.1037/0021-9010.92.4.909.
- Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *CoRR*, abs/2012.08630, 2020. URL <https://arxiv.org/abs/2012.08630>.
- Federal Trade Commission. Federal trade commission announces final rule banning fake reviews and testimonials. <https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials>, August 2024.
- Vincent Frey and Julian Martinez. Interpersonal trust modelling through multi-agent reinforcement learning. *Cognitive Systems Research*, 83:101157, 2024. ISSN 1389-0417. doi: 10.1016/j.cogsys.2023.101157.
- Team Gemma. Gemma 4 technical report, 2026. URL <https://arxiv.org/abs/2607.02770>.
- Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570.

-
- Zhengyang Hu, Yuying Zhu, Dawei Zhao, and Chengyi Xia. The higher-order networked n-player trust game driven by reputation and reinforcement learning. *Chaos, Solitons & Fractals*, 2026. 117623, 2026. ISSN 0960-0779. doi: 10.1016/j.chaos.2025.117623.
- Hugging Face. Security incident disclosure — July 2026. <https://huggingface.co/blog/security-incident-july-2026>, July 2026. Published 16 July 2026. See also the technical timeline: <https://huggingface.co/blog/agent-intrusion-technical-timeline>.
- Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006. ISSN 1573-7454. doi: 10.1007/s10458-005-6825-4.
- Audun Jøsang and Roslan Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.
- John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50.30392. PMID: 15151155.
- Alison Legood, Lisa van der Werff, Allan Lee, Deanne den Hartog, and Daan van Knippenberg. A critical review of the conceptualization, operationalization, and empirical literature on cognition-based and affect-based trust. *Journal of Management Studies*, 60(2):495–537, 2023. doi: 10.1111/joms.12811.
- Shijun Lei, Quang Nguyen, Swapneel S. Mehta, Zeping Li, Huichuan Fu, Xiaolong Zheng, Siki Chen, Yunji Liang, Philip Torr, and Zhenfei Yin. Strategic exploitation in LLM agent markets: A simulation framework for e-commerce trust. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL <https://arxiv.org/abs/2605.10059>. arXiv:2605.10059.
- Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 464–473, 2017.
- Joel Z. Leibo, Edgar A. Duñez-Guzman, Alexander Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6187–6199, 2021.
- Mircea Lică, Ojas Shirekar, Baptiste Colle, and Chirag Raman. MindForge: Empowering embodied agents with theory of mind for lifelong cultural learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL <https://arxiv.org/abs/2411.12977>.
- Bertram F. Malle and Daniel Ullman. A multidimensional conception and measure of human-robot trust. In Chang S. Nam and Joseph B. Lyons (eds.), *Trust in Human-Robot Interaction*, pp. 3–25. Academic Press, 2021. ISBN 978-0-12-819472-0. doi: 10.1016/B978-0-12-819472-0.00001-0.
- Stephen Paul Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, Stirling, Scotland, 1994. URL <http://hdl.handle.net/1893/2010>.
- Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995. doi: 10.2307/258792. URL <https://doi.org/10.2307/258792>.
- D. Harrison McKnight, L. L. Cummings, and Norman L. Chervany. Initial trust formation in new organizational relationships. *The Academy of Management Review*, 23(3):473–490, 1998. doi: 10.5465/amr.1998.926622.
- Arthur G. Mercier. Surgical interventions for causal exploration with LLM-based agents. Master’s thesis, Delft University of Technology, Delft, The Netherlands, November 2025. URL <https://resolver.tudelft.nl/uuid:60bf96d8-315e-4eca-9c83-6ff5903ba3a5>. MSc thesis, Computer Science.

-
- Benedikt V. Meylahn, Arnoud V. den Boer, and Michel Mandjes. Interpersonal trust: Asymptotic analysis of a stochastic coordination game with multi-agent learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(6):063119, 2024.
- Gal Nagli. Hacking Moltbook: The AI social network any human can control. Wiz Blog, February 2026. URL <https://www.wiz.io/blog/exposed-moltbook-database-reveals-millions-of-api-keys>.
- OpenAI. OpenAI–Hugging Face incident technical report. Technical report, OpenAI, August 2026. URL <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>.
- Stavan Parikh and Rao Surapaneni. Powering AI commerce with the new Agent Payments Protocol (AP2). Google Cloud Blog, September 2025. URL <https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol>.
- Isaac Pinyol and Jordi Sabater-Mir. Computational trust and reputation models for open multi-agent systems: A review. *Artificial Intelligence Review*, 40(1):1–25, 2013. doi: 10.1007/s10462-011-9277-z.
- Isaac Pinyol, Jordi Sabater-Mir, Pilar Dellunde, and Mario Paolucci. Reputation-based decisions for logic-based cognitive agents. *Autonomous Agents and Multi-Agent Systems*, 24(1):175–216, 2012. ISSN 1573-7454. doi: 10.1007/s10458-010-9149-y.
- Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. *The Academy of Management Review*, 23(3):393–404, 1998. doi: 10.5465/amr.1998.926617.
- Jordi Sabater and Carles Sierra. REGRET: reputation in gregarious societies. In *Proceedings of the Fifth International Conference on Autonomous Agents*, pp. 194–195, 2001. doi: 10.1145/375735.376110.
- Jordi Sabater and Carles Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60, 2005. ISSN 1573-7462. doi: 10.1007/s10462-004-0041-5.
- Jordi Sabater-Mir, Mario Paolucci, and Rosaria Conte. Repage: REPutation and imAGE among limited autonomous partners. *Journal of Artificial Societies and Social Simulation*, 9(2):3, 2006.
- Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure, 2024. URL <https://arxiv.org/abs/2311.07590>.
- F. David Schoorman, Roger C. Mayer, and James H. Davis. An integrative model of organizational trust: Past, present, and future. *The Academy of Management Review*, 32(2):344–354, 2007. URL <http://www.jstor.org/stable/20159304>.
- Maurice E. Schweitzer, John C. Hershey, and Eric T. Bradlow. Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, 101(1):1–19, 2006. ISSN 0749-5978. doi: 10.1016/j.obhdp.2006.05.005.
- W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck. TRAVOS: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, March 2006. doi: 10.1007/s10458-006-5952-x.
- Nenad Tomasev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero. Virtual agent economies, 2025. URL <https://arxiv.org/abs/2509.10147>.
- Nenad Tomašev, Matija Franklin, and Simon Osindero. Intelligent AI delegation, 2026. URL <https://arxiv.org/abs/2602.11865>.

-
- Matteo Venanzi, Michele Piunti, Rino Falcone, and Cristiano Castelfranchi. Facing openness with socio-cognitive trust and categories. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 400–405, 2011. doi: 10.5591/978-1-57735-516-8/IJCAI11-075.
- Hjalmar Wijk, Ajeya Cotra, and Ryan Greenblatt. Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident. Technical report, METR, August 2026. URL <https://metr.org/hugging-face-incident-report-aug-2026.pdf>. Greenblatt (Redwood Research) contracting with METR.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf, 2024. URL <https://arxiv.org/abs/2309.04658>.
- Han Yu, Zhiqi Shen, Chunyan Miao, Bo An, and Cyril Leung. Filtering trust opinions through reinforcement learning. *Decision Support Systems*, 66:102–113, 2014. ISSN 0167-9236. doi: 10.1016/j.dss.2014.06.006.
- Guozhong Zheng, Jiqiang Zhang, Jing Zhang, Weiran Cai, and Li Chen. Decoding trust: a reinforcement learning perspective. *New Journal of Physics*, 26(5):053041, 2024. doi: 10.1088/1367-2630/ad4b5a.

A TRUTHMARKETTWIN ENVIRONMENT

	Phase	Agents	Available actions
1	Seller communication	Sellers	create_post, like_post, dislike_post
2	Buyer communication	Buyers	create_post, like_post, dislike_post
3	Listing	Sellers	list_products
4	Purchase	Buyers	purchase_products
5	Rating	Buyers	rate_transactions
6	Challenge	Buyers	challenge_warrants

Table 4: Phases of one round, in order. The challenge phase exists only in the reputation and warrant market. In every phase an agent may also choose `do_nothing`. Posts are visible only to agents of the same role, so sellers and buyers never communicate directly. Budgets are reset at the start of each round, and products that remain unsold at the end of a round expire.

Action	Effect
<i>Sellers</i>	
<code>list_products</code>	Lists products, each specified by its advertised quality, its true quality, a quantity and, in the warrant market, whether it carries a warrant. The price is fixed by the advertised quality. The total production cost of a listing may not exceed the seller’s budget.
<i>Buyers</i>	
<code>purchase_products</code>	Buys products by their identifiers, as long as the buyer’s budget covers the price. The true quality is revealed after the purchase.
<code>rate_transactions</code>	Rates each purchase with a thumbs-up when the delivered quality met or exceeded the advertised quality, or a thumbs-down otherwise. Ratings are shown with a delay of one round.
<code>challenge_warrants</code>	Challenges a purchase that carried a warrant, at a cost of 1. If the delivered quality differs from the advertised quality, the seller loses the warrant escrow (8 for a high-quality claim, 2 for a low-quality claim), and the buyer receives the escrow and is refunded the challenge cost. Otherwise the buyer loses the challenge cost.
<i>Both roles</i>	
<code>create_post</code>	Posts a public message to agents of the same role.
<code>like_post</code>	Publicly agrees with an existing post, in place of posting.
<code>dislike_post</code>	Publicly disputes an existing post, in place of posting.
<code>do_nothing</code>	Takes no action in the current phase.

Table 5: Actions available to the agents in our experiments. Compared with the TruthMarketTwin implementation, listing and purchasing use batch actions instead of single-item ones, `quote_post` is not offered, and `exit_market` and `reenter_market` are disabled. The other social actions of the underlying OASIS platform, such as following, reposting and commenting, are not offered in either version.

Advertised	Delivered	Price	Seller profit	Buyer utility
HQ	HQ	8	+4	+4
HQ	LQ	8	+6	−4
LQ	LQ	3	+1	+1
LQ	HQ	3	−1	+9

Table 6: Payoff of a single sale. The seller earns the price minus the production cost of the delivered quality (4 for high quality, 2 for low quality), and the buyer receives the utility of the delivered quality (12 for high quality, 4 for low quality) minus the price. Unless a warrant is successfully challenged, advertising a low-quality product as high quality is therefore the most profitable choice for a seller and the costliest outcome for a buyer.

B CHANGES TO MINDFORGE

Component	MindForge	TrustForge
Environment	Embodied environment in which agents act on objects	TruthMarketTwin market with fixed round phases that restrict which actions each agent may take
Objective	Complete the assigned tasks, coordinating with the other agents	Maximize the agent’s own cumulative utility as a buyer or profit as a seller, with buyers and sellers on opposite sides of each sale
Observation	State of the embodied environment, messages from other agents and execution errors	Textual market observation with the listed products, their advertised quality, price and warrant status, the sellers’ reputations and the agent’s own budget or utility, together with posts and execution errors
Communication	Direct conversations between agents	Public posts on a feed shared by agents of the same role, which agents can also like or dislike
Critic	Judges from the environment state whether the current task has been completed	Judges the last action against the task of its phase, given the market rules and the agent’s resulting profit or utility. Taking no action counts as a success when the phase offers nothing worth doing
Task selection	The automatic curriculum proposes a new task after each success	The task is fixed by the current phase and the agent’s role, and the curriculum only refreshes the context of that task
Skill library	Stores every successful action as a reusable skill	Does not store actions that refer to the products or transactions of a specific round, since these cannot be replayed
Partners	Every other agent	Every other agent of the same role, and for buyers also every seller
Source of partner beliefs	Conversation with the partner	For sellers, the buyer’s own purchase history, the seller’s reputation and listings, and other buyers’ posts, without any direct communication. For peers, the peer’s posts and the buyer’s own record of what sellers delivered
Trust	Not modelled	Competence, willingness and opportunity assessed per partner and task, each as a reasoning and a score, and combined into a degree of trust
Role of the partner model	Input to action selection	Also evidence for willingness (task beliefs) and opportunity (perception beliefs), and can be disabled per agent
Peer testimony	Read as received	Annotated with the buyer’s trust in its author when trust between buyers is enabled
Input to action selection	Skills, episode summary and beliefs	Skills, episode summary, beliefs and the degree of trust in each partner

Table 7: Conceptual changes from MindForge to TrustForge. Episodic memory and the agent’s own perception, interaction and task beliefs keep their role, adapted to the market. TrustForge is implemented on the codebase of CausalForge (Mercier, 2025), which re-implements MindForge and was provided by the CausalForge authors, with its causal world model and surgical action selection disabled.

C PROMPTS

Placeholder	Content
{agent_id}, {role}	Name and role of the assessing agent
{partner_id}	Name of the assessed partner
{action}	The task τ , “delivering the quality they advertise” for sellers and “giving me truthful and useful information about the market” for peers
{goal}	“Maximize your cumulative utility”
{experience_line}	Summary of the last 20 purchases from the seller, for example “Your own purchase history with Alpha: your last 4 purchases from them: 4 were advertised HQ, of which 4 arrived LQ — 100% of their HQ claims to you were broken. Net utility from those purchases: -16.0. Most recent shortfall: round 6.”
{reputation}	Public rating of the seller, for example “19 thumbs-up / 17 thumbs-down (53% positive over 36 ratings)”
{listing_status_line}	What the seller listed in the current round
{market_communications}	Recent posts of other buyers, each annotated with the agent’s trust in its author when same-role trust is enabled
{convo}	Recent posts of the assessed peer
{own_record_line}	For each seller the agent bought from, the number of high-quality claims and the share that arrived as low quality
{partner_model_clause}, {partner_model_line}	The agent’s task beliefs about the partner for willingness and its perception beliefs for opportunity, removed when the partner model is disabled
{previous_partner_belief}	The partner model from the previous update
{previous_reasoning}, {previous_score}	The previous assessment of the same dimension, or of trust
{competence_reasoning}, {competence_score},...	The current assessment of each dimension

Table 8: Placeholders in the belief prompts.

C.1 CROSS-ROLE PROMPTS

Prompt C.1: Partner model, cross-role

You are {role} {agent_id} operating in an online marketplace alongside other agents. You are forming beliefs about {partner_id}, who is a seller and therefore someone you trade with rather than talk to. There is no direct message channel between you: you have never exchanged a word with {partner_id}, and you never will. Everything you believe about them must come from what they DO and from what other buyers REPORT about them.

Three kinds of evidence are given below, and they are not equally reliable:

- Your OWN purchase history with {partner_id}. You saw the true quality of everything you bought from them yourself.
- {partner_id}'s public reputation and what they have listed are OBSERVED FACTS recorded by the marketplace.
- Other buyers' posts are CLAIMS.

For each of the three belief categories below, write 1–3 plain English sentences:

- perception_beliefs: what {partner_id}'s current market position looks like — what they have listed this round, what they are advertising, and what their public reputation says about how buyers have rated them so far
- interaction_beliefs: what other buyers are saying about {partner_id}, and how much weight that testimony deserves given who is saying it and whether it is corroborated
- task_beliefs: what {partner_id} appears to be trying to achieve, whether their advertised quality has matched what was actually delivered, and what you expect them to do next

{experience_line}

Other buyers' recent posts: {market_communications}

{partner_id}'s public reputation: {reputation}

{listing_status_line}

Previous Partner Beliefs: {previous_partner_belief}

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "perception_beliefs": what {partner_id} currently has on offer and what their public reputation shows
  "interaction_beliefs": what other buyers claim about {partner_id} and how much that testimony should be trusted
  "task_beliefs": what {partner_id} is trying to do, whether their claims have matched delivery, and what they will likely do next
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the three belief values. Plain prose only.

Prompt C.2: Competence, cross-role

You are {agent_id}, a {role}, operating in an online marketplace alongside other agents. You are assessing {partner_id}'s competence for {action}, in service to your standing goal: {goal}. Competence is about ability. It is your belief that {partner_id} possesses the skills, knowledge, expertise and self-confidence needed to actually perform {action} that would help towards your goal.

Based on your own purchase history with {partner_id}, other buyers' recent posts, {partner_id}'s public reputation, what they've listed this round, and your previous beliefs about {partner_id}'s competence, write a plain English summary of your competence perception of {partner_id} and a score between 0.0 to 1.0 attributed to that competence.

For the reasoning category below, write 1–3 plain English sentences. For the score category below, write a number between 0.0 and 1.0:

- reasoning: your perception of the competence of {partner_id} in relation to {action}, in service of your goal: {goal}
- score: your perception of the competence score of {partner_id} in relation to {action}, in service of your goal: {goal}, a score between 0.0 and 1.0

{experience_line}

Other buyers' recent posts: {market_communications}

{partner_id}'s public reputation: {reputation}

{listing_status_line}

Your previous competence assessment: {previous_reasoning} (score: {previous_score})

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "reasoning": your perception of the competence of {partner_id} in relation to {action}, in
service of your goal: {goal}, (skills, knowledge, expertise, self-confidence)
  "score": your perception of the competence score of {partner_id} in relation to {action}, in
service of your goal: {goal}, a score between 0.0 and 1.0
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the two belief values.

"reasoning" MUST be plain prose only and "score" MUST be a single number between 0.0 and 1.0.

Prompt C.3: Willingness, cross-role

You are {agent_id}, a {role}, operating in an online marketplace alongside other agents. You are assessing {partner_id}'s willingness for {action}, in service to your standing goal: {goal}. Willingness is about choice. It is your belief that {partner_id} intends to perform {action} that would help towards your goal and will persist in that intention.

Based on your own purchase history with {partner_id}, other buyers' recent posts, {partner_id}'s public reputation, what they've listed this round, {partner_model_clause} and your previous beliefs about {partner_id}'s willingness, write a plain English summary of your willingness perception of {partner_id} and a score between 0.0 to 1.0 attributed to that willingness.

For the reasoning category below, write 1–3 plain English sentences. For the score category below, write a number between 0.0 and 1.0:

- reasoning: your perception of the willingness of {partner_id} in relation to {action}, in service of your goal: {goal}
- score: your perception of the willingness score of {partner_id} in relation to {action}, in service of your goal: {goal}, a score between 0.0 and 1.0

{experience_line}

Other buyers' recent posts: {market_communications}

{partner_id}'s public reputation: {reputation}

{listing_status_line}

{partner_model_line}

Your previous willingness assessment: {previous_reasoning} (score: {previous_score})

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "reasoning": your perception of the willingness of {partner_id} in relation to {action}, in
service of your goal: {goal}, (intention, persistence)
  "score": your perception of the willingness score of {partner_id} in relation to {action}, in
service of your goal: {goal}, a score between 0.0 and 1.0
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the two belief values.

"reasoning" MUST be plain prose only and "score" MUST be a single number between 0.0 and 1.0.

Prompt C.4: Opportunity, cross-role

You are {agent_id}, a {role}, operating in an online marketplace alongside other agents. You are assessing {partner_id}'s opportunity for {action}, in service to your standing goal: {goal}. Opportunity is your belief that {partner_id} currently has the means and circumstance needed to actually perform {action} at all, regardless of whether they are able to do it well or willing to do it honestly.

Based on your own purchase history with {partner_id}, other buyers' recent posts, {partner_id}'s public reputation, what they've listed this round, {partner_model_clause} and your previous beliefs about {partner_id}'s opportunity, write a plain English summary of your opportunity perception of {partner_id} and a score between 0.0 to 1.0 attributed to that opportunity.

For the reasoning category below, write 1–3 plain English sentences. For the score category below, write a number between 0.0 and 1.0:

- reasoning: your perception of the opportunity of {partner_id} in relation to {action}, in service of your goal: {goal}
- score: your perception of the opportunity score of {partner_id} in relation to {action}, in service of your goal: {goal}, a score between 0.0 and 1.0

{experience_line}

Other buyers' recent posts: {market_communications}

{partner_id}'s public reputation: {reputation}

{listing_status_line}

{partner_model_line}

Your previous opportunity assessment: {previous_reasoning} (score: {previous_score})

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "reasoning": your perception of the opportunity of {partner_id} in relation to {action}, in
service of your goal: {goal}, (means, standing, continued access)
  "score": your perception of the opportunity score of {partner_id} in relation to {action}, in
service of your goal: {goal}, a score between 0.0 and 1.0
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the two belief values.

"reasoning" MUST be plain prose only and "score" MUST be a single number between 0.0 and 1.0.

C.2 SAME-ROLE PROMPTS

Prompt C.5: Partner model, same-role

You are {role} {agent_id} operating in an online marketplace alongside other agents.
You have been reading posts from {role} {partner_id}.

Based on the posts, what the sellers they mention actually delivered to you, and your previous beliefs, write a plain English summary of your perception of {partner_id}. These partner beliefs should be informative of {partner_id}'s state.

For each of the three belief categories below, write 1–3 plain English sentences:

- perception_beliefs: your perception of what {partner_id} currently observes in the environment (products available, reputations if available, prices, their budget or utility)
- interaction_beliefs: your perception of what {partner_id} understands from recent posts — what they said, what they asked, what they responded to
- task_beliefs: what {partner_id} is currently trying to do, what progress they have made and what they plan to do next

{own_record_line}

Previous Partner Beliefs: {previous_partner_belief}

Posts: {convo}

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "perception_beliefs": your perception of what {partner_id} currently observes in the environment
  (products available, reputations if available, prices, their budget or utility)
  "interaction_beliefs": your perception of what {partner_id} understands from recent posts — what
  they said, what they asked, what they responded to
  "task_beliefs": what {partner_id} is currently trying to do, what progress they have made and
  what they plan to do next
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the three belief values. Plain prose only.

Prompt C.6: Competence, same-role

You are {agent_id}, a {role}, operating in an online marketplace alongside other agents. You are assessing {partner_id}'s competence for {action}, in service to your standing goal: {goal}. Competence is about ability. It is your belief that {partner_id} possesses the skills, knowledge, expertise and self-confidence needed to actually perform {action} that would help towards your goal.

Based on {partner_id}'s previous posts in the marketplace, what the sellers they mention actually delivered to you, and your previous beliefs about {partner_id}'s competence, write a plain English summary of your competence perception of {partner_id} and a score between 0.0 to 1.0 attributed to that competence.

For the reasoning category below, write 1–3 plain English sentences. For the score category below, write a number between 0.0 and 1.0:

- reasoning: your perception of the competence of {partner_id} in relation to {action}, in service of your goal: {goal}
- score: your perception of the competence score of {partner_id} in relation to {action}, in service of your goal: {goal}, a score between 0.0 and 1.0

{partner_id}'s recent public posts in the marketplace : {convo}
{own_record_line}

Your previous competence assessment: {previous_reasoning} (score: {previous_score})

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "reasoning": your perception of the competence of {partner_id} in relation to {action}, in
service of your goal: {goal}, (skills, knowledge, expertise, self-confidence)
  "score": your perception of the competence score of {partner_id} in relation to {action}, in
service of your goal: {goal}, a score between 0.0 and 1.0
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the two belief values.

"reasoning" MUST be plain prose only and "score" MUST be a single number between 0.0 and 1.0.

Prompt C.7: Willingness, same-role

You are {agent_id}, a {role}, operating in an online marketplace alongside other agents. You are assessing {partner_id}'s willingness for {action}, in service to your standing goal: {goal}. Willingness is about choice. It is your belief that {partner_id} intends to perform {action} that would help towards your goal and will persist in that intention.

Based on {partner_id}'s previous posts in the marketplace, what the sellers they mention actually delivered to you, {partner_model_clause} and your previous beliefs about {partner_id}'s willingness, write a plain English summary of your willingness perception of {partner_id} and a score between 0.0 to 1.0 attributed to that willingness.

For the reasoning category below, write 1–3 plain English sentences. For the score category below, write a number between 0.0 and 1.0:

- reasoning: your perception of the willingness of {partner_id} in relation to {action}, in service of your goal: {goal}
- score: your perception of the willingness score of {partner_id} in relation to {action}, in service of your goal: {goal}, a score between 0.0 and 1.0

{partner_id}'s recent public posts in the marketplace : {convo}

{own_record_line}

{partner_model_line}

Your previous willingness assessment: {previous_reasoning} (score: {previous_score})

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "reasoning": your perception of the willingness of {partner_id} in relation to {action}, in
service of your goal: {goal}, (intention, persistence)
  "score": your perception of the willingness score of {partner_id} in relation to {action}, in
service of your goal: {goal}, a score between 0.0 and 1.0
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the two belief values.

"reasoning" MUST be plain prose only and "score" MUST be a single number between 0.0 and 1.0.

Prompt C.8: Opportunity, same-role

You are {agent_id}, a {role}, operating in an online marketplace alongside other agents. You are assessing {partner_id}'s opportunity for {action}, in service to your standing goal: {goal}. Opportunity is your belief that {partner_id} currently has the means and circumstance needed to actually perform {action} at all, regardless of whether they are able to do it well or willing to do it honestly.

Based on {partner_id}'s previous posts in the marketplace, what the sellers they mention actually delivered to you, {partner_model_clause} and your previous beliefs about {partner_id}'s opportunity, write a plain English summary of your opportunity perception of {partner_id} and a score between 0.0 to 1.0 attributed to that opportunity.

For the reasoning category below, write 1–3 plain English sentences. For the score category below, write a number between 0.0 and 1.0:

- reasoning: your perception of the opportunity of {partner_id} in relation to {action}, in service of your goal: {goal}
- score: your perception of the opportunity score of {partner_id} in relation to {action}, in service of your goal: {goal}, a score between 0.0 and 1.0

{partner_id}'s recent public posts in the marketplace : {convo}
{own_record_line}
{partner_model_line}

Your previous opportunity assessment: {previous_reasoning} (score: {previous_score})

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "reasoning": "your perception of the opportunity of {partner_id} in relation to {action}, in
service of your goal: {goal}, (means, standing, continued access)
  "score": "your perception of the opportunity score of {partner_id} in relation to {action}, in
service of your goal: {goal}, a score between 0.0 and 1.0
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the two belief values.

"reasoning" MUST be plain prose only and "score" MUST be a single number between 0.0 and 1.0.

C.3 DEGREE OF TRUST

Prompt C.9: Degree of trust, used in both cross-role and same-role assessments

You are {agent_id}, a {role}, operating in an online marketplace alongside other agents. You are computing your overall degree of trust in {partner_id} for {action}, in service to your standing goal: {goal}. Trust is a function of the strength of three separate beliefs you hold about {partner_id}: their opportunity, their competence, and their willingness to actually help you achieve your goal. The stronger and more credible these three beliefs are, the stronger your overall trust.

These three beliefs do not always matter equally, and you should not just average them. Depending on the action and goal, one may deserve more weight than the others: an action that depends on strategic judgment makes competence more decisive, a time-sensitive or honesty-dependent action makes willingness more decisive, and an action requiring specific resources or access makes opportunity more decisive. A serious weakness in any single one of the three should pull your overall trust down, since a partner lacking opportunity, competence, or willingness cannot be fully trusted no matter how strong the other two are.

Your competence belief about {partner_id}: {competence_reasoning} (score: {competence_score})

Your willingness belief about {partner_id}: {willingness_reasoning} (score: {willingness_score})

Your opportunity belief about {partner_id}: {opportunity_reasoning} (score: {opportunity_score})

Your previous trust assessment: {previous_reasoning} (score: {previous_score})

For the reasoning category below, write 1–3 plain English sentences integrating the three beliefs above. For the score category below, write a number between 0.0 and 1.0:

- reasoning: your overall degree of trust in {partner_id} for {action}, integrating their opportunity, competence, and willingness
- score: your overall degree of trust score in {partner_id} for {action}, a score between 0.0 and 1.0

Here is an example for you. IMPORTANT you MUST respond in the following JSON object format:

```
{
  "reasoning": "your overall degree of trust in {partner_id} for {action}, integrating their opportunity, competence, and willingness",
  "score": "your overall degree of trust score in {partner_id} for {action}, a score between 0.0 and 1.0"
}
```

Do NOT add extra keys, nested dictionaries, boolean flags, or escaped JSON strings inside the two belief values.

"reasoning" MUST be plain prose only and "score" MUST be a single number between 0.0 and 1.0.

D ADDITIONAL RESULTS

Figure 11: Trust and its dimensions attributed by buyers to sellers before and after a seller’s first deceptive sale to a buyer in round t , averaged over all deceived buyer–seller pairs, in the three conditions not shown in Figure 5: the reputation-only market with the reduced seller budget, and the reputation-with-warrant market with the standard and the reduced seller budget (mean over five runs, band ± 1 SD).

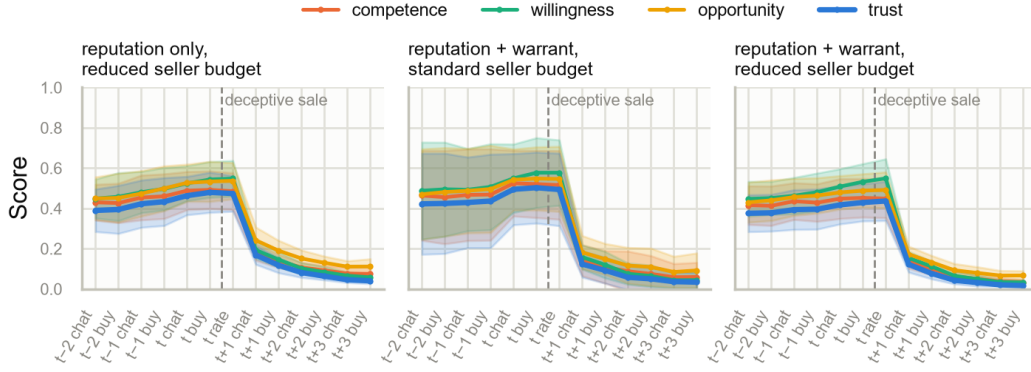


Figure 12: Trust attributed by buyers to sellers after a seller’s first honest sale to a buyer, made before any deceptive sale, and after its first deceptive sale, in the three conditions not shown in Figure 6: the reputation-only market with the reduced seller budget, and the reputation-with-warrant market with the standard and the reduced seller budget (mean over five runs, band ± 1 SD). Honest sales followed by a deceptive sale from the same seller within three rounds are excluded, as are deceptive sales followed by an honest one.

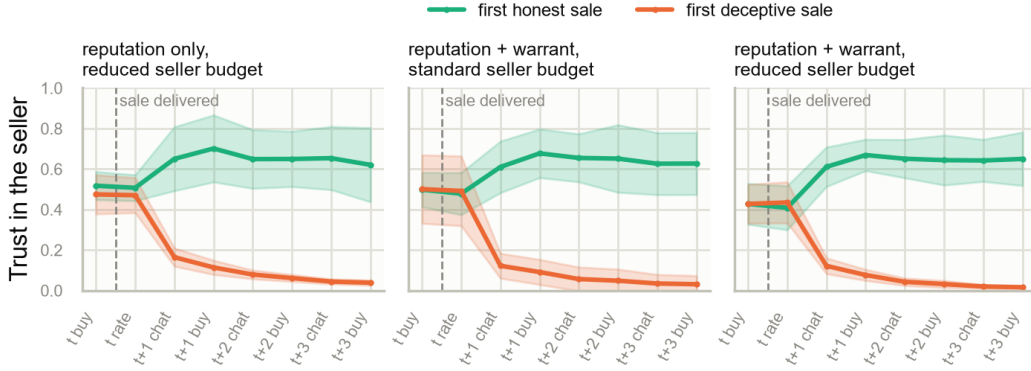
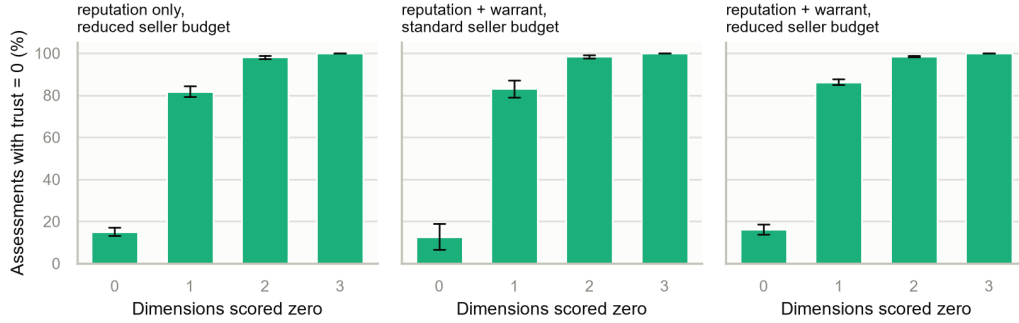


Figure 13: Proportion of trust assessments in which trust is zero, by the number of dimensions among competence, willingness and opportunity scored zero, in the three conditions not shown in Figure 7: the reputation-only market with the reduced seller budget, and the reputation-with-warrant market with the standard and the reduced seller budget (mean over five runs, error bars ± 1 SD).



E SIMULATION PARAMETERS

Experiment	Trust	Market	Seller budget	Rounds	Conditions × runs
Baseline comparison	on, off	r, r+w	18	75	4 × 5
Market scarcity	on, off	r, r+w	9	75	4 × 5
Peer Testimony	on	r	18	50	1 × 5
Partner-model ablation	on, off	r, r+w	18	75	4 × 5

Table 9: Experimental design configuration. All other parameters are listed in Table 10.

Parameter	Description	Value
<i>Simulation</i>		
runs	Independent runs per condition	5
num_sellers	Seller agents per run	5
num_buyers	Buyer agents per run	10
simulation_rounds	Trading rounds per run	75 (50 in testimony trust)
<i>Market rules</i>		
market_type	Trust institution (condition factor)	reputation_only, reputation_and_warrant
reputation_lag	Rounds before a rating is displayed	1
exit_round	Round at which sellers may exit	disabled
reentry_allowed_round	Round at which exited sellers may re-enter	disabled
initial_window_rounds	Rounds of suppressed reputation at entry	none
communication_type	Which roles may post	both
communication_channel_type	Posts are agent-generated	Real
<i>Agent cognition</i>		
include_trust_beliefs	Degree of Trust layer (condition factor)	{on, off}
buyer_buyer_trust	Testimony peer trust	off (on in testimony trust)
seller_seller_trust	Same-role trust for sellers	off
enable_cross_role_trust	Buyers hold sellers as partners	on
tom_disabled_buyers	Buyers without the partner model	0 (5 in the partner-model ablation)
include_reputation_signal	Public rating shown to agents	on
experience_window	Own purchases per partner fed to belief prompts	20
<i>Product economics</i>		
hq_cost	Production cost, high quality	4.0
lq_cost	Production cost, low quality	2.0
hq_price	Fixed price, high quality	8.0
lq_price	Fixed price, low quality	3.0
hq_utility	Buyer utility, high quality	12.0
lq_utility	Buyer utility, low quality	4.0
<i>Budget system</i>		
seller_budget	Seller budget, refreshed each round	18.0 (9.0 in market scarcity)
buyer_budget	Buyer budget, refreshed each round	60.0
<i>Warrant system</i>		
hq_warrant_escrow	Escrow staked against an HQ claim	8.0
lq_warrant_escrow	Escrow staked against an LQ claim	2.0
challenge_cost	Cost to a buyer of challenging	1.0
<i>Model</i>		
model	Served locally via vLLM 0.19.1	gemma-4-e4b-it
enable_thinking	Model reasons before answering	on
temperature	Sampling temperature	1.0
context_length	Context window	32768

Table 10: Simulation parameters, fixed across all experiments unless stated otherwise. The product economics, budget and warrant blocks are identical to those of TruthMarketTwin. We differ in population (5 sellers rather than 10), horizon (75 rounds rather than 10, and 50 in testimony trust), and model (gemma-4-e4b-it in place of GPT-4o).