Modèles et apprentissages en sciences humaines et sociales
MASHS 2026

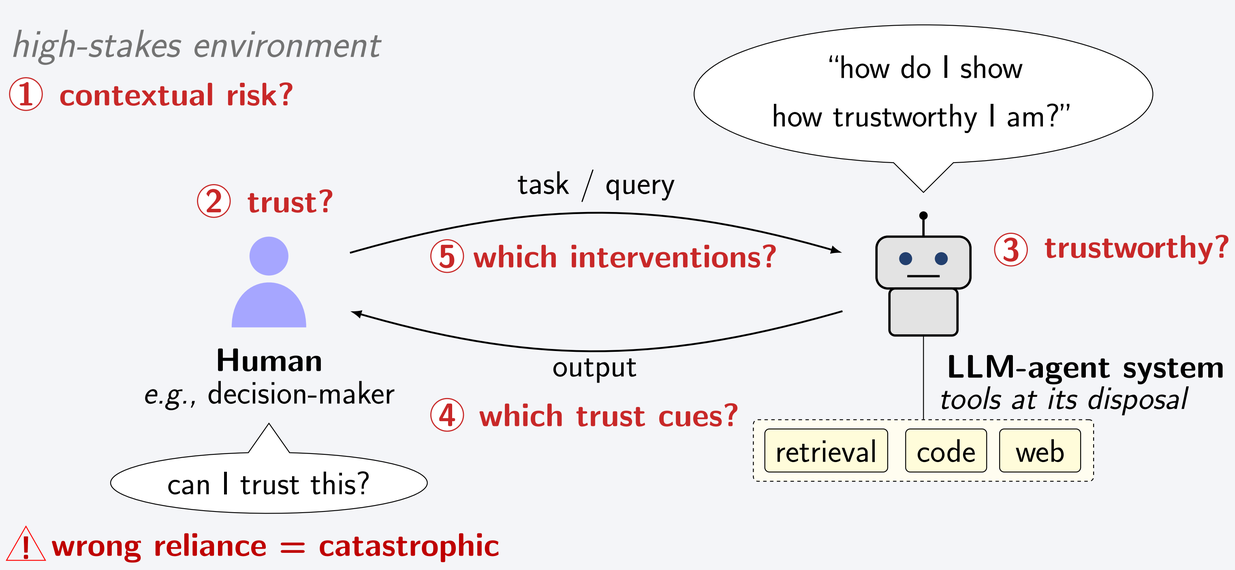
CRiDiT: Calibrating Human–AI Trust in High-Stakes Systems

Explanations are not calibration. Run-time trust management needs both sides measured, and cues and interventions decided at run time under contextual risk.

Yuntian Ding[1,*] Nicolas Herbaut[1,2] Camille Salinesi[1]

[1]Centre de Recherche en Informatique de Paris 1 (CRI), Université Paris 1 Panthéon-Sorbonne
[2]University of Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France
[*] yuntian.ding@univ-paris1.fr

The Run-Time Trust Problem



The challenges

A human relies on an LLM-agent system — the AI-infused system class we focus on here — which is an opaque model orchestrating a set of tools. In a high-stakes setting, where a wrong reliance decision can be catastrophic, **showing the system's trustworthiness level to the user is the core difficulty**: five things stay unclear in the interaction loop:

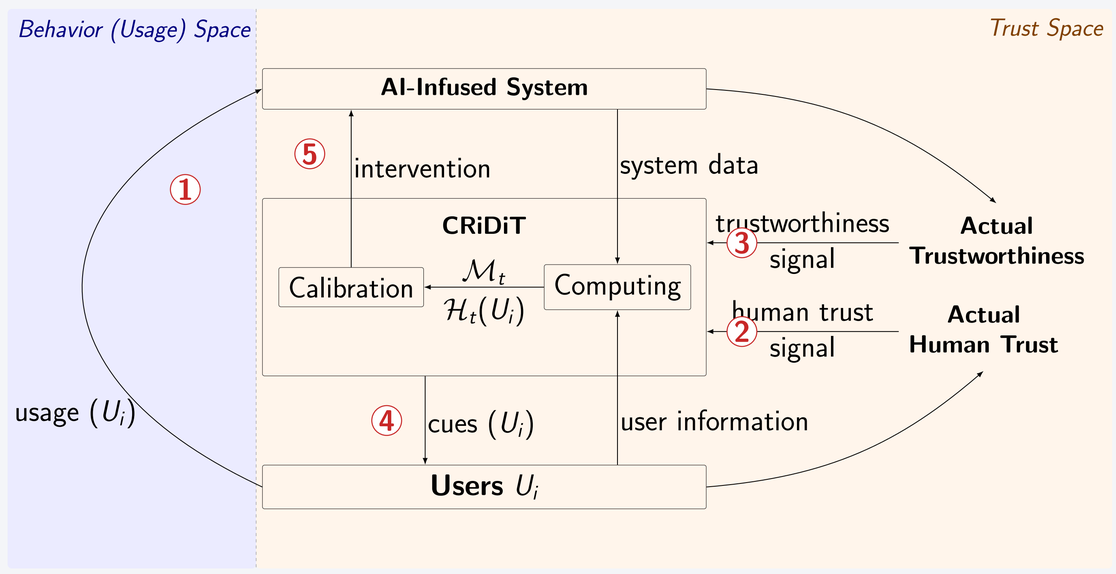
① **Contextual risk:** how much uncertainty or assurance exists in a high-stakes environment is not made explicit.
② **Human trust:** how much trust the human places in the system is hard to observe.
③ **System trustworthiness:** an LLM agent orchestrating tools gives fuzzy or absent evidence of how trustworthy it actually is.
④ **Trust communication:** it is unclear which trust cues to surface, matched to the user's *current* trust level.
⑤ **Trustworthy intervention:** it is unclear which interventions to configure, or who should take them, to raise trustworthiness in the current context.

1. The CRiDiT Model

CRiDiT [1] is our closed-loop architecture for the interaction loop above: four layers (input, trust computing, calibration, output). Each challenge maps onto a connection in the loop, and each connection yields an engineering requirement for an artifact that monitors and calibrates trust in an AI-infused system.

Challenge → requirement:

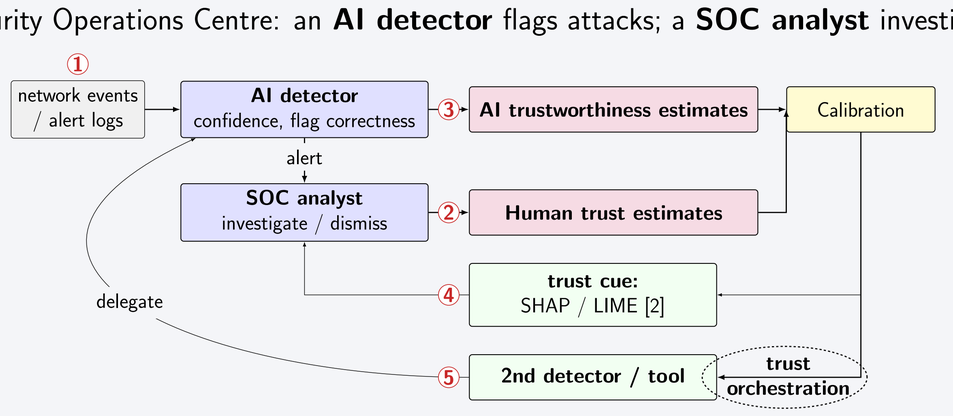
① contextual risk → **Context-awareness**: agents shall perceive the environment from contextual events (interaction conditions, risks) and use them in calibration.
② + ③ opaque trust & trustworthiness → **Dynamic trust management**: agents shall estimate, update & compare system trustworthiness and human trust in real time.
④ + ⑤ unclear cues & interventions → **Effective calibration policy**: agents shall map the trust gap $\Delta_t = \mathcal{H}_t - \mathcal{M}_t$ to decide on concrete cues / interventions when $|\Delta_t| > \tau$ (τ is the trigger threshold).



$\mathcal{M}_t$: *estimated trustworthiness* (machine side); $\mathcal{H}_t(U_i)$: individual *estimated human trust* (human side).

2. Use Cases in the Literature – Read through CRiDiT

In a Security Operations Centre: an **AI detector** flags attacks; a **SOC analyst** investigates or dismisses [2].



Other use cases

Domain	Machine side: System data *(in)*	Machine side: Intervention *(out)*	Human side: User information *(in)*	Human side: Trust cue *(out)*
Healthcare	diagnostic confidence, guideline / evidence match	specialist interventions (human-in-the-loop)	accept / override, clinician expertise	explainable-AI interface: feature attribution, e.g. SHAP [3]
Deepfake	agreement, detection logs	route to 2nd model / expert	accept / override, consults explanation	visual explainability: Grad-CAM heatmap [4]
Education	learner level, goal / prerequisite match	explanation adaptation	follows / rejects, consults explanation	N/A [5]

Mapped onto CRiDiT, the gaps surface: how human trust ② and trustworthiness ③ are estimated is rarely specified and validated for trust calibration, the trust cues ④ are highly reliant on the quality of explanations and the interventions ⑤ either stay fixed at design time, or depend on human intervention which is not always guaranteed in highly automated settings under contextual risk ①.

3. Open Research Directions

- **Cue → cognition → trust:** how a trust cue (verbal message, visual indicator) travels through *perception* → *comprehension* → *projection* → *decision*, and how this chain conditions effective calibration.
- **Machine-side trust intervention:**
  Two questions arise: *when* to delegate, and *how* to compare candidate components. From these gaps, we derive requirements for a **trust-delegation** artifact:
  - **Automation & Flexibility:** agents shall map user information to *tool-level* trustworthiness to decide *when* to delegate a task, rather than a hard-coded configuration.
  - **Decision consensus:** agents shall map tool-level trustworthiness to an *orchestration policy* (e.g., game theory) to reach a consensus on delegation.
  - **Security:** connections open attack surfaces → agents shall verify tool performance and detect malicious manipulation.
- **Validation by simulation:** simulating the closed loop:
  - **How:** Users with trust dynamics simulated using human-side data, against an agent orchestrating tools of estimated trustworthiness.
  - **Goal:** the gap $|\Delta_t|$ shrinks after a cue or intervention, and wrong-reliance decisions fall against a no-calibration baseline.

References

[1] Yuntian Ding, Nicolas Herbaut, and Camille Salinesi. "Toward Operational Trust Calibration in AI-Infused Systems: A Systematic Review of Computational Trust Models". In: (2025).
[2] Humberto Goncalves et al. "Explainable AI for Cloud Intrusion Detection: A User Study of SHAP and LIME in AWS GuardDuty". In: *Proceedings of the 5th IEEE International Conference on AI in Cybersecurity (ICAIC)*. IEEE, 2026.
[3] Proloy Karmakar et al. "Human–AI Interaction in Healthcare: A User-Centered Evaluation of Trust and Usability in Explainable AI Interfaces". In: *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*. IEEE, 2026. DOI: 10.1109/QPAIN69676.2026.11546108.
[4] Francesco Castro et al. "DeepDect: an explainable AI platform for face swapping and face generation DeepFake detection". In: *Multimedia Tools and Applications* (2026). DOI: 10.1007/s11042-026-21672-1.
[5] Mouadh Guesmi et al. "Interactive explanation with varying level of details in an explainable scientific literature recommender system". In: *International Journal of Human–Computer Interaction* 40.22 (2024), pp. 7248–7269.