Agent Transparency as a mechanism of Trust Improvement between Humans and Industry 4.0 Systems

Loïck Simon, Philippe Rauffet, Clément Guerin

Agent Transparency as a mechanism of Trust Improvement between Humans and Industry 4.0 Systems

Loïck SIMON*, Philippe RAUFFET*, Clément GUERIN*

*FHOOX Team, Lab-STICC UMR CNRS 6285, University of South Brittany
LOICK.SIMON.UNIV-UBS.FR, PHILIPPE.RAUFFET@UNIV-UBS.FR,
CLEMENT.GUERIN@UNIV-UBS.FR

Abstract. Industry 4.0 is witnessing the emergence of new hybrid teams, composed of human operators and intelligent, autonomous cyber-physical systems. Within these teams, new questions regarding cooperation arise, especially concerning human trust in the artificial agents able to provide recommendations based on data or processing capabilities that humans lack. This paper explores the issue of trust through two case studies, modeling the various layers of trust, examining their impact on the acceptance of recommendations, and investigating whether agent transparency affects trust.[1]

Keywords. Trust in autonomy, Agent Transparency, Industry 4.0

1 Introduction

Human-autonomy teams in Industry 4.0 involve the collaboration between human workers and autonomous systems, including robots and AI, to enhance efficiency, productivity, and competitiveness in industrial environments. These teams harness the complementary skills and strengths of both humans and machines to perform tasks and make decisions, resulting in better performance and customer satisfaction [1, 2, 3]. In this new era of manufacturing, human-autonomy teams are essential in shaping the future of work and industry competitiveness.

In many industrial contexts, these hybrid teams operate through a vertical cooperation structure, where a higher-level hierarchical agent oversees and holds authority, while the lower-level agent offers advice [4]. Specifically, artificial intelligence (AI) can analyze a situation and suggest a course of action to a human operator. The human can choose to trust and follow the AI's recommendation or consider and debate the AI's advice before making a decision.

1.1 Different layers of Human-Autonomy Trust

According to [5], in the relationship between humans and AI, the human acts as the trustor and the AI as the trustee, with the trust relationship heavily influenced by

[1] This paper is a modified version of [23] presented in the French-Speaking Conference CIGI adfa, p. 1, 2011.

the situation's characteristics (such as task goals and environmental constraints). Trust in AI is shaped by three interrelated layers [6, 7], as illustrated in Figure 1:

- ***Dispositional trust:*** This refers to the a priori trust in technology, influenced by personal traits such as age, gender, and general propensity to trust technology.
- ***Trust in signal:*** Also known as learned trust [6], this layer pertains to how humans perceive and interact with the accuracy of the AI's messages over time. It is affected by the perceived reliability of the AI both before and during interactions, as well as the explanations provided by the AI.
- ***Situational trust:*** This involves elements related to the environment that impact trust, such as the perceived risk of a situation and the evaluation of the consequences if we follow agent's advice.

Furthermore, there may be a moderating effect between trust in signal and situational trust [7]. Humans primarily base their decision to comply with the AI on trust in the signal. However, the perception of risk associated with the situation can influence this trust, altering the degree to which humans choose to comply with the AI.

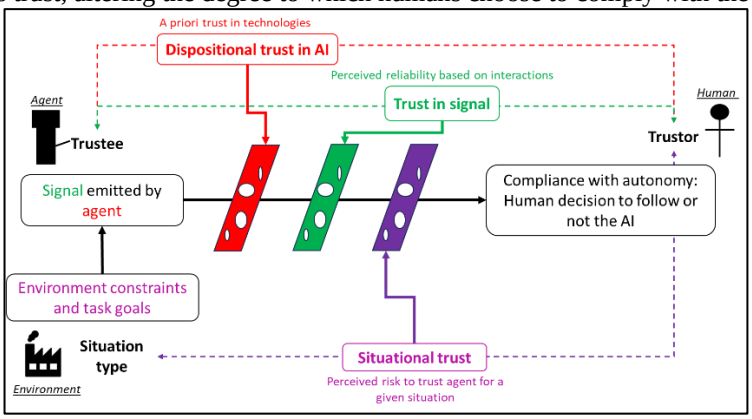

Fig. 1. 3 layers of trust, as explaining factors of human compliance with autonomous systems

1.2 Agent transparency to correct and calibrate trust

Recent research on Human-Autonomy Teaming [8, 9] indicates that ***transparency*** play a crucial role in the way a human intends to cooperate with an autonomous agent. By adjusting agent transparency, it is possible to correct misplaced trust—whether undue distrust or overconfidence—and to calibrate trust, aligning compliance rates more closely with system reliability [10]. As highlighted in recent literature reviews on transparency [11, 12], there are two primary approaches to describing and operationalizing transparency: the Situation Awareness-based Agent Transparency (SAT) model by Chen et al. [13], and Lyons's framework for transparency in human-robot interaction [14].

The SAT model defines three levels of agent transparency based on Situation Awareness theory [13]. Level 1 is dedicated to the communication of basic information about the current states and the actions of the agent. Level 2 deals with the rationales and the explanations underlying the agent's actions or decisions. That al-

lows the human to understand the information processing of the agents, as well as its work constraints and its rooms for actions. Finally, level 3 is more focused on the sharing of information related to estimated or projected outcomes, as well as the communication of the probability metrics associated with these projections and these estimations.

According to Lyons [14], transparency can also be considered in different models related to cooperative situations, divided into two main dimensions. The robot-TO-human (rTOh) transparency deals with what the agent communicates about itself to the human operator, including its goals (intentional model), the tasks it carries out (task model), its reasoning (analytical model), and its work situation (environmental model). Conversely, robot-OF-human (rOFh) transparency concerns what the agent communicates about the human operator to this human. The agent plays the role of a kind of a virtual "mirror" for the human partner. This rOFh transparency can encompass the cognitive and physical state of the human (operator model) as well as the organization of the activity between the human and the agent (the teamwork model).

1.3 Research questions

Drawing from the various concepts and models discussed above, this paper addresses three main research questions or hypotheses:

- ***R1:*** We aim to investigate how each of the three layers of trust [6] in autonomy influences and drives compliance with AI.
- ***R2:*** In line with the literature and the hypotheses proposed by Chancey et al. [7], we seek to explore how situational trust can moderate the impact of trust in signal on compliance with autonomy.

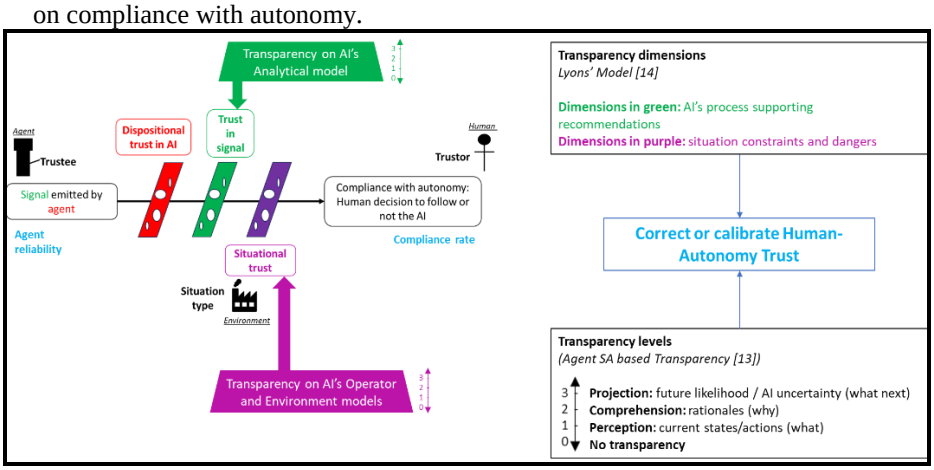

Fig. 2. agent transparency, as moderator of trust in signal and situational trust

- ***R3:*** As depicted in Figure 2, we assume compliance with and trust in AI can be adjusted and enhanced by varying agent transparency across different levels and dimensions. Specifically, transparency related to situational risk and operator workload (provided by Robot-OF-human models [14]) could affect situational

trust, while transparency related to the agent's processes generating recommendations (provided by Robot-To-Human models [14]) may influence trust in signal.

2 Method

2.1 Human-autonomy teaming and trust at different levels of activity in Industry 4.0

Industry 4.0 introduces new human-machine systems, comprising human operators and cyber-physical components, whose multifaceted interactions at operational, tactical, and strategic levels will aid in monitoring and managing manufacturing systems. These interactions can be examined through various frameworks. Schmidt [16] employs a functional-structural approach, identifying three types of cooperation:

- ***Augmentative cooperation*** boosts physical or intellectual teamwork by adding agents with identical skills to handle a workload that a single agent cannot manage.
- ***Confrontative cooperation*** improves solutions and reduces errors by comparing views between agents, requiring mutual monitoring and control, following Reagan's principle, "trust but verify."
- ***Integrative cooperation*** involves agents with complementary skills working together.

Additionally, Hoc's [17] functional approach highlights the resulting interferences between human and artificial agents, which must be managed at three levels:

- ***Cooperation in action:*** Involves managing operational activities in real-time and short-term, including creating (e.g., disagreement), detecting (e.g., redundancy), anticipating, and resolving interferences.
- ***Cooperation in planning:*** Encompasses tactical activities for managing common goals and plans, and dividing functions.
- ***Meta-cooperation:*** Supports the previous levels by developing a shared communication code and models for both self and partner.

2.2 Two case studies to explore transparency and trust

The research presented in this paper stems from two French national projects. First, the ***SEANATIC*** project, supported by ADEME (the French Agency for Environment and Sustainable Development), brought together the expertise of academic and industrial partners, including Lab-STICC UMR CNRS 6285, Azimut, IoT.bzh, Thalos, and Piriou. This project focused on developing a comprehensive set of solutions to collect, analyze, and present technical data to aid mechanical engineers and fleet managers in improving preventive maritime maintenance using new intelligent and predictive tools based on machine learning. Second, the ***HUMANISM*** project, funded by ANR (the French National Research Agency), was conducted by three academic partners: Lab-STICC UMR CNRS 6285, CReSTIC EA 3804, and LAMIH UMR CNRS 8201. The project aimed to model and develop new mechanisms for function allocation, dialogue, and HMI to facilitate human-machine cooperation in Industry 4.0, particularly

with intelligent robots and AI at operational, tactical, and strategic levels. Table 1 lists the characteristics of these two case studies, presented in sections 3 and 4, highlighting their differences in addressing the question of human-autonomy trust across various settings, such as different levels of activity management and agent accuracy.

Table 1. characteristics of the two use cases studied in two research projects

	Humanism project	**Seanatic project**
ACTIVITY IN INDUSTRY 4.0	Order picking at different stations in a production line	Maintenance operation planning in maritime domain
HUMAN-MACHINE COOPERATION TYPES [16]	*Confrontative cooperation*: agent proposes actions, human decides to accept or not *Integrative cooperation:* agent processes and shares information that a human could hardly manage	
LEVEL OF COOPERATION INTERFERENCES [17]	Cooperation in action, at operational level	Cooperation in planning, at tactical level
ABILITY AND ACCURACY OF AUTONOMOUS AGENT	Robot always myopic	Strong and reliable predictive model
	Systemic inaccuracy in recommendation	High accuracy (90%) in recommendation
KNOWLEDGE OF PARTICIPANTS	Accuracy unknown by participants	Accuracy known by participants

2.3 Evaluation of trust and compliance

Various questionnaires and metrics are proposed in the literature to evaluate the different layers of trust, as well as the compliance rate of human participants with AI recommendations. Table 2 details what we used in our experiments (cf. §3 and §4).

Table 2. Measures of trust and compliance

Trust layers	**Metrics, questionnaires and references**
COMPLIANCE	- Measured in each situation after the participant's decision (accept or reject agent's recommendation).
TRUST IN SIGNAL	- Measured at the beginning and after each situation using the IMOTRIS scale (French translation of [18]). - Measured using a trust signal item after each situation.
SITUATIONAL TRUST	- Perception of risk measured with risk perception scales (French translation of [19]). - Mental workload measured with ISA (French translation of [20]) after each situation.
DISPOSITIONAL TRUST	- Participant's affinity towards technology using the ATI scale (French translation of [21]). - Participant's propensity to trust technology using the PTT scale (French translation of [22]).

2.4 Data processing and analysis

Finally, we analyzed the relationships between the different layers of trust and agent transparency settings by computing logistic regressions and linear mixed models with the lme4 package in R studio. In all statistical models, we included in our models the dimensions of transparency and situation characteristics as interacting categorical predictors, and we included a random intercept for participants. We also used stepwise model selection using AIC to select the best model.

3 1st case study: HAT in operational activities of order picking

3.1 Experiment

In Humanism project, the experiment was designed to instantiate a cooperation between human operators and a collaborative robot (i.e. cobot), preparing different customer orders at their own workstation, sharing the same resources (the parts are stored in the warehouse, and supplied to the different stations through a common conveyor belt) (Figure 3). Regularly, there is a seeming interference between the human and the agent, taking the form of a stock-out of a shared resource (cf. Figure 3). The cobot asked some help from the participant, playing the role of the supervisor of the production facility, by taking some parts in the stock of another human operator and by transferring in the cobot stock, to solve this problem of stock-out. However, this request could be irrelevant, due to an informational myopia of the cobot, which is not able to verify if it can be supplied from the warehouse with these missing parts.

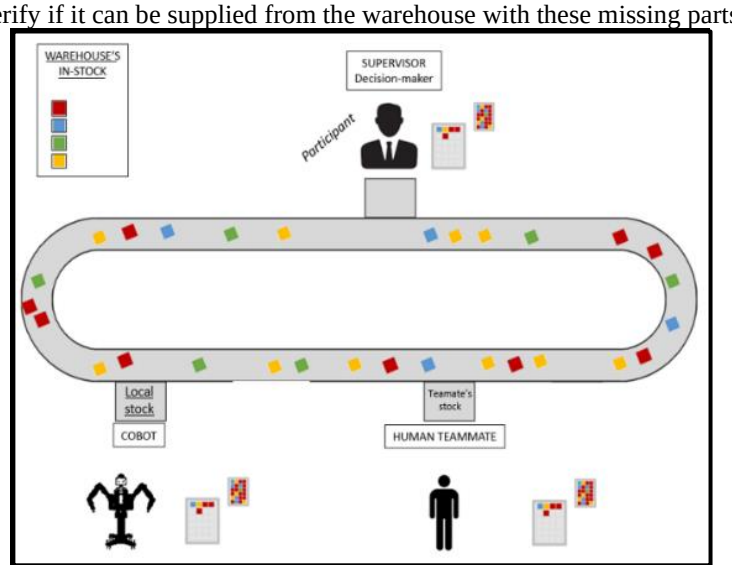

Fig. 3. Order Picking experiment

53 students specialized in industrial engineering (17 women, 36 men, mean age = 21.34 years, SD = 1.67) participated to this experiment. They were presented to repeated situations, in which the transparency varied in levels (cf. SAT framework

[13]), along two different dimensions [14], as explained in the table 3. This experiment aims to answer two of our research questions:

- ***R1.*** Which kind of trust really drive compliance with AI?
- ***R2.*** How situational trust moderates the effect of trust in signal upon compliance?

Table 3. Transparency conditions in Humanism

CONDITION	TRANSPARENCY LEVEL ON ANALYTICAL MODEL, RELATED TO TRUST IN SIGNAL	TRANSPARENCY LEVEL ON OPERATOR MODEL, RELATED TO SITUATIONAL TRUST
S1	**A_L2:** Cobot alerts on a potential problem of missing resources, but without transparency on its myopia. Participant could think cobot had considered the warehouse stock.	**O_L0*:*** there is no transparency on teammate taskload
S2		**O_L1+:** human teammate is less busy than cobot
S3		**O_L1-:** human teammate is busier than cobot, and cobot request can disturb human activity
S4	**A_L3:** Additionally, cobot specified it did not consider the warehouse's in-stock, being transparent on its myopia. Participants were certain about cobot limitations.	**O_L0:** described above
S5		**O_L1+:** described above
S6		**O_L1-:** described above

3.2 Results

R1. Relationship between compliance and trust layers

- ***Dispositional trust.*** We did not find any significant effect of participant's affinity towards technology (using the ATI scale) and participant's propensity to trust technology (using the PTT scale) upon compliance.
- ***Trust in signal.*** Compliance and trust in signal were significantly correlated with Mann-Whitney tests. Compared to non-compliant participants, participants following the cobot's proposal reported to better understand to cobot ($W=50846$, $p = .008$), to feel it as more reliable ($W=51383$, $p = .004$), as well as more trustworthy ($W=55450$, $p < .001$).
- ***Situational trust.*** Similarly, there was a significant relationship between compliance and risk perception. Indeed, participants accepting the cobot's proposal perceived lower risk of the situation than participants declining their help to the cobot ($W=23378$, $p=.038$).

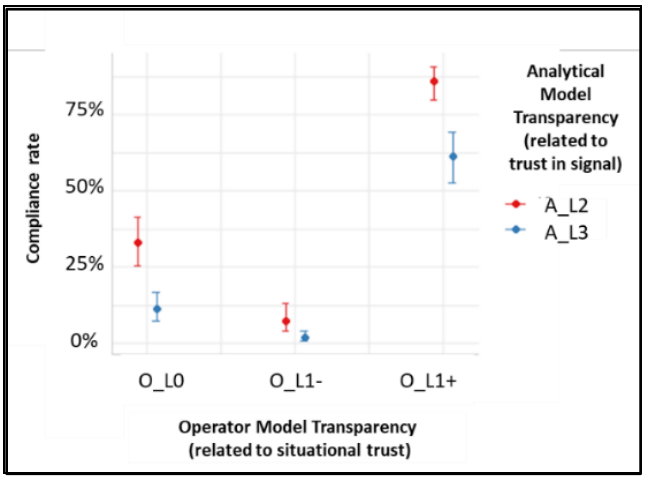

Fig. 4. Mixed effect of trust layers on compliance

R2. Moderation effect of situational trust on compliance

Compliance is significantly higher when cobot is non-transparent on the situation (condition O_L0, with no information on human teammate activity), compared to condition O_L1-, where cobot is transparent about the fact that the teammate is busier than itself (OR = 8.47, $p < .001$).

Moreover, situational trust was analyzed as interacting with trust in signal upon compliance. Indeed, participants complied less with a cobot having a low transparency on the analytical model and a high transparency on the negative state of the operator (A_L2 and O_L1-), than when they cooperated with a cobot very transparent on the analytical model and on the positive state of the teammate (A_L3 and O_L1+) (OR = 7.29, $p < .01$).

4 2nd case study: HAT in tactical planning activities

4.1 Experiment

In Seanatic project, a case study was designed, with a human and an AI cooperating in planning maintenance; AI can suggest advancing or postponing operations, and human decides to accept this modification, or decline it and keep the initial date from the CMMS tool (cf. Figure 5, left side). 39 participants (25 men, 14 women, mean age = 22.15 years, SD = 2.77), students in industrial engineering, were presented to repeated situations, where AI suggested different decisions (advance or postpone a maintenance operation) in different context (critical or non-critical part to change), with different level of agent transparency. In addition to default transparency of AI on analytical and environmental models in every situation, as illustrated in Figure 5 (right), information was added, either on system reliability (in condition **Rel**, displaying information supporting ***trust in signal***), or on environment risk on equipment and

logistics (in condition ***Risk***, displaying information related to ***situational trust***). A last condition (***Rel&Risk***) combined both previous settings. In this experiment, we aim to answer two of our main research questions:

- ***R1.*** Which kind of trust does really drive compliance with AI?
- ***R3.*** How may transparency settings play upon the different layers of trust?

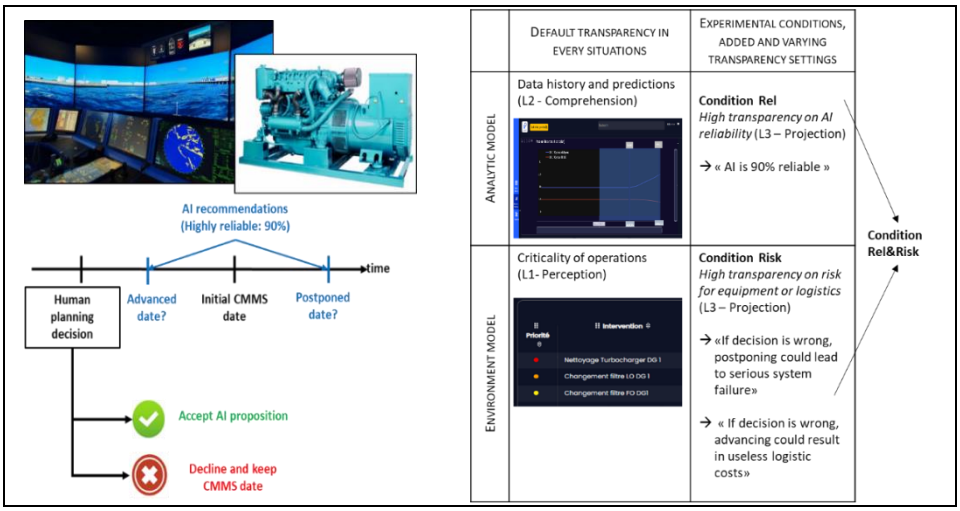

Fig. 5. Maintenance planning experiment (left) and transparency conditions in Seanatic (right)

4.2 Results

R1. Relationship between compliance and trust layers

- ***Dispositional trust.*** We did not find any significant effect of participant's affinity towards technology (using the ATI scale) and participant's propensity to trust technology (using the PTT scale) upon compliance.
- ***Trust in signal.*** Mann-Whitney analyses showed that trust in signal and compliance are significantly associated, with higher trust for participants complying with the predictive maintenance tool ($W = 3857.5$, $p < .001$).
- ***Situational trust.*** Similarly, compliance was found to be correlated with risk perception. Participants complying with predictive maintenance suggestions significantly reported the perception of a lower risk ($W=13911$, $p < .001$).

R3. Effect of transparency on trust in signal and situational trust

As depicted in Figure 6, when AI was transparent only on "Reliability" (“Rel” condition), there was an increase in trust in signal compared to situations where AI communicated risks (respectively for "Rel & Risk": $OR = 0.39$, $p < .05$; and for "Risk": $OR = 0.26$, $p = .001$). Moreover, for risk perception, when AI was transparent only on Reliability (“Rel” condition), there was a decrease in risk perception compared to situations where AI was transparent about risks (respectively for "Rel&Risk": $OR = 4.73$, $p < .001$; and for "Risk": $OR = 6.02$, $p < .001$). Finally, the

criticality of maintenance operations did not influence trust in the signal. On the contrary, we observed that when the criticality of the proposal is "Moderate," the perception of risk was higher compared to a "High" criticality (OR = 3.21, $p < .005$).

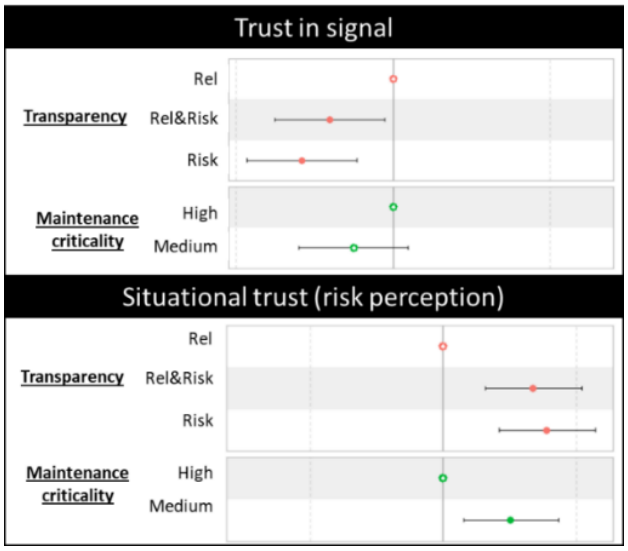

Fig. 6. Effect of transparency on different trust layers

5 Discussion

These two case studies have explored the relationships between agent transparency, trust, and compliance, within human-autonomy teams (HAT) involved in activities representative of Industry 4.0. We were thus able to study these relationships both in operational activities (cooperation in doing) and tactical activities (cooperation in planning), as well as with intelligent systems that are more or less reliable (very myopic and uncertain for Humanism, very reliable for Seanatic).

These different experiments showed some convergence and reproducibility of the results, particularly regarding question ***R1***. Contrary to what we might have expected from the results of our two experiments, dispositional trust does not significantly affect compliance. We argue here that dispositional trust is not directly link to behavioural measure because it is more influenced by factors relatives to the situation. Our following results are congruent with this explanation. The other two layers of trust are strongly correlated with compliance: compliance increases with a higher trust in signal and a lower risk perception (risk perception being linked with situational trust).

Moreover, the findings related to question ***R2*** and presented in Humanism case study were therefore congruent with the assumptions of Chancey et al. [7], posing that risk perception can be considered as a moderator of the human trust in autonomous agent's reliability, and that moderating effect impacts human compliance with a robot.

Finally, considering the question ***R3***, we have found that playing on agent transparency levels and dimensions can significantly modify trust in signal and situational trust, and subsequently compliance with intelligent systems (cf. Figure 6). This find-

ing corroborates the different research works mentioned in §1.2, and open new perspectives, in design or operation stages, to correct or better calibrate trust in autonomy, to avoid distrust, mistrust and complacency effect.

6 Conclusion

This research work investigated the question of trust in human-autonomy teams (HAT) within Industry 4.0 situations. This paper articulates the different conceptual works on trust in autonomy, in terms of behavioural performance (compliance with AI) and trust layers (dispositional trust, situational trust and trust is signal). Moreover, based on two case studies, it provides insights on how each trust layer can play upon compliance and how these layers are interrelated. Finally, it demonstrates how trust and compliance can be modified and manipulated by varying agent transparence. It opens perspectives for the design and operational control of HAT by incorporating a bidirectional dialog between the operator and the autonomous agent.

7 Acknowledgements

The authors would like to acknowledge the support of the Humanism project, funded by ANR (the French National Research Agency), and the Seanatic project, funded by ADEME (the French Agency for Environment and Sustainable Development) for the two studies presented in this paper.

8 References

1. Rauffet, P. (2022). Tools and methods for Human-Autonomy Teaming: Contributions to cognitive state monitoring and system adaptation. Habilitation à Diriger des Recherches. *University of South Brittany*
2. Rauffet, P., Guerin, C., Chauvin, C., & Martin, E. (2018). Contribution of Industry 4.0 to the emergence of a joint cognitive and physical production system. In *HFES European Chapter, Berlin, Germany*
3. McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors*, *60*(2), pp. 262-273
4. Lemoine, M. P., Debernard, S., Crevits, I., & Millot, P. (1996). Cooperation between humans and machines: first results of an experiment with a multi-level cooperative organisation in air traffic control. *CSCW*, *5*, 299-321.
5. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.
6. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.
7. Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors*, *59*(3), 333-345.
8. Lyons, J. B. (2013, March). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

9. Wright, J. L., Chen, J. Y., Barnes, M. J., & Hancock, P. A. (2017). The effect of agent reasoning transparency on complacent behavior: An analysis of eye movements and response performance. In *3rd HFES Conference, Philadephia, USA*
10. Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *11th ACM/IEEE HRI Conference, Christchurch, New Zealand*
11. Simon, L., Guérin, C., Rauffet, P., Chauvin, C., & Martin, É. (2023). How humans comply with a (potentially) faulty robot: Effects of multidimensional transparency. *IEEE Transactions on Human-Machine Systems*, *53*(4), 751-760.
12. Rajabiyazdi, F., & Jamieson, G. A. (2020). A review of transparency (seeing-into) models. In *the 33rd IEEE SMC Conference, Toronto, Canada*
13. Chen, J. Y., Procci, K., Boyce, M., Wright, J. L., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency*. ARMY RESEARCH LAB ABERDEEN.
14. Lyons, J. B., Koltai, K. S., Ho, N. T., Johnson, W. B., & Smith, D. E. (2016). Engineering trust in complex automated systems. *Ergonomics in Design*, *24*(1), 13-17.
15. Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*, 85-104.
16. Schmidt, K., Rasmussen, J., Brehmer, B., & Leplat, J. (1991). Cooperative work: A conceptual framework. *Distributed decision making*, 75-110.
17. Hoc, Jean-Michel. "Towards a cognitive approach to human–machine cooperation in dynamic situations." *IJHCS* 54(4): 509-540.
18. Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, *20*(4), 440-458.
19. Wilson, R. S., Zwickle, A., & Walpole, H. (2019). Developing a broadly applicable measure of risk perception. *Risk Analysis*, *39*(4), 777-791
20. Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, *39*(5), 740-748.
21. Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, *35*(6), 456-467.
22. Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In 11th International VAMR Conference, Orlando, USA.
23. Rauffet P., Simon L., Guérin C. (2023). Transparence et Confiance au sein des équipes Humains Systèmes de l'Industrie 4.0. *CIGI QUALITA MOSIM, Trois Rivières, Canada*