Trust in Automation: Analysis and Model of Operator Trust in Decision Aid AI Over Time

Vincent Fer, Daniel Lafond, Gilles Coppin, Mathias Bollaert, Olivier Grisvard, Pierre de Loor

Trust in Automation: Analysis and Model of Operator Trust in Decision Aid AI Over Time

Vincent Fer
Thales DMS & IMT Atlantique
Brest, France
vincent.fer@imt-atlantique.fr

Daniel Lafond
Thales Research and Technology
Quebec, Canada
daniel.lafond@thalesgroup.com

Gilles Coppin
IMT Atlantique
Brest, France
gilles.coppin@imt-atlantique.fr

Mathias Bollaert
Thales Defence Mission Systems
Brest, France
mathias.bollaert@fr.thalesgroup.com

Olivier Grisvard
Thales Defence Mission Systems
Brest, France
olivier.grisvard@thalesgroup.com

Pierre De Loor
ENIB
Brest, France
pierre.deloor@enib.fr

***Abstract*—Understanding how human trust in AI evolves over time is essential to identify the limits of each party and provide solutions for optimal collaboration. With this goal in mind, we examine the factors that directly or indirectly influence trust, whether they come from humans, AI, or the environment. We then propose a summary of methods for measuring trust, whether subjective or objective, to show which ones are best suited for longitudinal studies. We then focus on the main driving force behind the evolution of trust: feedback. We justify how learning feedback can be transposed to trust and what types of feedback can be applied to impact the evolution of trust over time. After understanding the factors that influence and how to measure trust, we propose an application example on a maritime surveillance tool with an AI-based decision aid.**

***Index Terms*—Trust in Automation, Maritime Patrol, Longitudinal Experiment, Feedback, Human Factors, Cognitive Engineering**

I. Introduction

With the emergence of AI several years ago in the world of research and then a sudden acceleration in its accessibility to a large part of the population in the civilian or military sectors, it is becoming essential to determine the level of trust that a human should be able to place in these intelligent systems. Today, used in the medical world to detect disease, in the banking sector to detect fraud, in IT to develop ultra-fast applications, or in the military to assist operators in high-risk operations, AI is becoming a central element in boosting performance. This integration is not without consequences, since humans tend to become over- or under-confident in automation [1] over time, this can lead to catastrophic results such as fatal accidents [2]. Confidence can generally be distinguished from trust in the following sense: confidence is defined as the expectation of a certain level of performance, while trust is a human attitude based on the perception of an agent's ability to help him perform a task in a situation characterized by uncertainty and vulnerability [3]. Trust is obviously not unique to the Human-AI pairing, but is present in all areas of our society, whether it is trust in our institutions, trust in science, trust in business, trust in justice, or trust in others. The same is true for research with a wide variety of disciplines that have been working on this subject for many years, including sociology, psychology, philosophy, neurology, informatics, etc. What is interesting about this diversity is that it is becoming clear that trust cannot be dealt with by a single discipline. This is what Lee & See have attempted to do with the transposition of concepts related to Human-Human trust to Human-Automation trust. A model in which trust is ultimately just one state in a cycle of cognitive evolution in humans has been proposed. The current question is not "Is the human trustworthy?" but "Is AI trustworthy?" For this first question, which may be essential in mission-critical systems, studies are underway, such as Hou's [4] article on the integration of decision-support AI, in which the AI can make a decision when the Human is no longer in a position to do so, in order to complete a mission. We propose to address the second question in the following sections, with a focus in Section 2 on trust in automation, its influencing factors, and how to measure it. In Section 3, we transpose learning feedback to feedback as a driver of trust evolution. In Section 4, we define a synthetic model based on models in the literature and enhanced by the notion of feedback. In the final section, Section 5, we propose an example of an application for maritime surveillance. Finally, we conclude this paper by summarizing our proposals and talking about our future experiments.

II. State of the art

A. Trust

Trust, a term at the core of Human-AI cooperation, is a major issue to be understood in the future to optimize Human and AI performance in rich and varied contexts. Humans are strong in contexts requiring nuance, while AI can handle large amounts of generic data. We explore different processes impacting trust that have a unique influence and occur at different temporalities. We will also show that the decision to trust is not solely impacted by trust but can also be influenced by factors specific to each context. To understand what trust is, let us go back to Lee & See's [3] definition, which is "trust is

a human attitude based on the perception of an agent's ability to help them perform a task in a situation characterized by uncertainty and vulnerability" [3]. In the context of human-autonomy teams, this definition highlights the three groups of factors that influence trust: trustor factors (human), trustee factors (AI), and environmental factors. [5]–[7].
These three sets of factors can be described as follows:

- Trustor factors: These factors are related to human operators, including their personality traits, previous experience with AI systems, cognitive abilities, and emotional states [8]. For the same experience with an AI, an operator with specific features of personality [9] - such as extroversion, for instance - can reveal to be more trustful than someone with another personality profile.
- Trustee factors: These factors relate to the AI system itself, such as its performance [10], process [11], transparency [12], and explainability [13]. An AI system that continuously provides accurate recommendations and is easy to understand will gain more trust from the operator. Factors such as system errors, poor communication, or lack of transparency can negatively impact trust [14].
- Environmental factors: These factors include the context and conditions under which the Human-AI interaction occurs. Factors such as task complexity, time pressure [15], or situational risk [16] can influence the operator's trust or trust-related behavior toward the AI system. For example, in high-risk environments, operators will be more cautious when relying on the AI system, especially if the potential consequences of an error are severe [16], [17].

Although influenced by different actors, trust is part of a cognitive cycle. Lee & See [3] propose a general model for the evolution of trust, including beliefs, attitudes, intentions, and behaviors. We synthesize this model in Fig.1. Beliefs are initially formed with "external" elements, such as rumors, training, or knowledge of a similar system, and will evolve over the course of the user's experience. These latest influence the user's attitude which also depends on characteristics specific to each individual (culture, predisposition to trust). Trust will have an impact on an individual's intention to use the system or not. Factors related to the context (time pressure) or the individual (perceived risk [16], workload [18]) can also influence this intention. Intention is not directly observable and may only be expressed through a behavior that is concrete evidence of whether or not an individual trusts the system. The most classical examples [19] of such behaviors are reliance, compliance, and verification time. Reliance consists of asking the system for a recommendation. Compliance consists of changing one's decision in response to a system proposal. The characteristics of behavior, such as verification time, are also meaningful: The shorter the time required to verify a recommendation, the higher the level of trust. However, intention can be influenced by many other factors, so observed behaviors cannot be directly mapped onto trust itself. One such example is the time pressure that can be encountered in different operational contexts. Time pressure arises when individuals face limited time to perform tasks. In such instances, the integration of AI becomes an important aid in improving performance while reducing the risk of critical errors. Several benefits can be observed, such as reduced workload, reduced stress, and increased performance. However, the Rieger article [20] confirms the findings of Rice and Keller [15]: time pressure can result in excessive reliance and an escalation of errors. Nonetheless, Rice and Keller [15] demonstrates that under high time pressure, mitigate reliance on AI can enhance overall performance compared to an operator working alone.

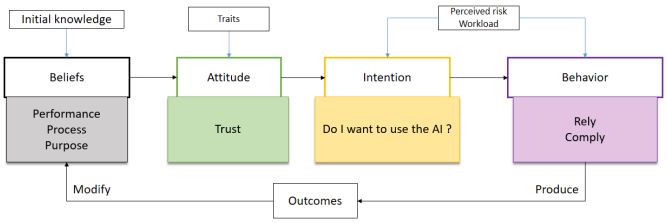

Fig. 1. The simplified Lee & See [3] model describing the evolution of trust in a cognitive cycle. Blue arrows represent influences outside the cycle, while black arrows represent evolution within the cycle.

To measure human trust in AI over its evolution at different moments, there are various subjective and objective methods [19], each with its own advantages and disadvantages. Among the subjective ones, self-report measures of trust, such as the Checklist for Trust [21] or the Measures of Trust & Trustworthiness [22] are easy to administer, but may be intrusive when used during experimentation. It can be used before or after an experiment to capture the attitude of trust and the beliefs. Behavioral and physiological measures, such as electroencephalography (EEG) [23], [24], provide more objective data but can be influenced by factors not related to trust and can require specialized equipment. It may be used during the experiment and will give online data about the behaviors but also on the attitude and intention. Researchers often use a combination of these methods to obtain a complete understanding of the dynamics of trust [1], [19], [25]. Currently, there is no general computational model of the evolution of trust, although some attempt to predict how trust evolves based on controlled variables limited to a specific context and with limited precision [**?**], [26], [27]. The gap lies in the fact that each context has unique variables with various impacts on the level of individuals trust. They are trying to estimate the level of trust at each point in time. Each event perceived by the system will change the value of trust or another variable. Estimating trust is important for regulating it. If a computational model is able to detect that an operator is becoming over- or under-confident towards the AI, then methods can be put in place to adjust it to the expected level before it is reached. General methods are proposed, for example, by de Visser [14] with apologies, explanation of repair, expression of system limits, or regular alerts for trust reduction. Certain methods are not applicable to all systems, and the appropriate ones must be selected for each context.

While there is currently no exact answer to the question

of how trust evolves, studies have highlighted the main factors that have an impact on trust [3], [5], [8], such as AI performance, its process, or its initial purpose. The impact of AI performance on trust has been demonstrated in numerous studies [5], [6], [10]. Few studies have been conducted to date on the process [28] or purpose [7].

- Performance is easily identifiable by comparing an expected result with the result obtained, or the speed with which a task is completed, with or without AI. Three metrics of performance are observable: Human performance (without AI), AI performance (without human) and team performance (Human + AI). Avril [10] studies the impact of AI performance on trust, and, indeed, trust increased when AI performance was high (87. 5%) compared to low (67.5%). Other information can be observed, such as the more information the AI gives about its choices, the better the Human-AI pairing performs. A change in behavior can also be observed when the information about a case is very precise and when no information is provided. No differences were observed when there was little or no information about potential errors. However, humans tend to see fewer AI errors with a high-performance AI (87.5%) than with a low-performance AI (67.5%). No correlation was found between self-confidence and AI performance.
- Concerning the process, which corresponds to the way the AI works or the steps it takes, this is not necessarily visible to the operator, and this is what studies on explainability and transparency are trying to resolve, to shed light on what are known as "black boxes".
- Purpose refers to how the AI is used, and more specifically how it is used in relation to its initial task. The operator perceives whether the AI's purpose is to help him in the task at hand, or whether it has not been developed to perform this task. A difference between the developers' initial goal and the goal expected by users can create disillusionment, which can reduce trust in automation. It is therefore important to communicate the latter's limits, with the goal of optimal collaboration.

Particular difficulties occur when trying to apply these kinds of indicators for the long-term analysis of operator trust, which is still a largely unexplored sub-field of research [29]. Among the rare references related to this topic, one can mention Beggiato's [11] study, which lasted two months. The research examined the evolution of trust and acceptance as drivers learned about the system. The results showed that trust and acceptance increased with familiarity, emphasizing the importance of learning in shaping users relationships with automation.

Our assumption, given Beggiato's findings on the similarities between the learning curve and the trust evolution curve, and Lee and See's model describing beliefs (perceived performance, process, and purpose) as the antecedents of trust, is that feedback, which is essential to learning, can be transposed to trust, since beliefs evolve with experience. Based on this premise, we turned to Bosc Miné's [30] article on feedback in a learning context. It describes all forms of feedback and allows us to target the ideal feedback for each context and need. To learn, feedback is necessary and essential for each learner (student, professional, expert, etc.). Using Lee & See's simplified model, behavior is linked to an outcome. This outcome is caused by the decision to trust or not. It can either agree with what was expected or create dissonance. In the former case, it will reinforce the initial beliefs; in the latter, it will modify the beliefs about the system. Therefore, we propose that the result is the feedback from our decision. Feedback is the element that allows beliefs to evolve. The particularity of feedback is that it can take different forms to reflect the state of the element impacted by the system user's decision. The final result of the decision to trust or not can therefore be interpreted by external actors and formatted to reflect, from a certain point of view, the whole process put in place by the human or AI leading to the final decision.

The previous section highlighted most of the components of trust in AI. The aim was to understand the processes by which trust evolves, the indirect influences on its behavior, and how to measure trust with different tools that have their own characteristics. In the following section, we will focus on feedback and its central role in the evolution of trust, including ways of integrating a type of feedback to a specific temporality (during the task and afterward). We then propose to integrate these different types of feedback into a general model of trust evolution, based in part on models found in the literature. We end by proposing an applicative framework in the context of maritime surveillance, where we will see, in particular, their constraints, how they work, and how to integrate the previous theoretical parts into this specific case.

B. Feedback

In each of the models in the literature on trust in automation, the concept of outcome is present. This result is the element that enables the modification of a human's belief in his autonomous system. In particular, it is the result of human-induced behavior in the use or non-use of automation. In the Mayer & Davis model, for example, it is described as an outcome. In its sense, this term refers to performance feedback on whether the task has succeeded or failed. It is simplifying and does not allow us to describe the elements that can constitute feedback being more than a success or failure. Then we have Lee's & See model, where the outcome is described as a "display", i.e. it is an integral part of the system, since it returns the result of using the system via an interface. What we are proposing brings nuance to the possible feedback given to the user to modify his initial knowledge of the system. In this approach, feedback is strongly associated with user learning. It provides an assessment or creates dissonance in the learner concerning his use of a system and its integration into his environment. The definition of beliefs shows that they evolve according to the use of the system and its results. This includes the notion of learning, since beliefs are what humans perceive of the system, including its performance, process, and purpose.

Over time and interactions, this perception evolves, and so does his or her knowledge. As feedback is a central element of learning and beliefs are being directly linked to learning, it seems obvious for us to transpose the descriptions of feedback for learning in our case on trust. In Bosc Miné's [30] article, she describes the different types of feedback applied to various contexts and shows the advantages and disadvantages of each. For example, some feedback is more suitable for experts, while others are more suitable for novices. We find groups of factors such as Human, System, and Environment at the source of the different feedback. The operator can provide himself with auto-feedback, i.e. it necessitates a kind of expertise to provide himself with feedback to learn from something observed. The system can provide direct or delayed performance feedback on its use, e.g., you succeeded in doing this or failed in doing that. As far as the environment is concerned, several types of feedback can be provided like physical feedback, such as a car accident, or feedback provided by a teacher (not linked to the operator-system pair) during a learning phase. These types of feedback are not necessarily independent of each other and can all be provided for the same event. Let us take the example of an autonomous car accident, in which different temporalities may exist:

- Before the accident, the car's automated system can trigger indicators (alarms, for example) to warn of an incoming accident;
- Just after the accident, the operator will question himself or the system, via auto feedback, which has not reacted as he expected;
- Several days later, an expert having analyzed the car's black box will be able to tell the driver everything what was done wrong, and thus provide the driver with elaborated feedback.

To provide some guidance, we need to be able to describe all types of feedback and understand the role of each. These three main types are elaborated feedback, verification feedback, and elaborated verification feedback. An overview of the feedback tree is given in Fig.2.

- Elaborated feedback comes from an external source and is intentional, the aim is to train the receiver of the feedback by providing clues to guide the individual towards a correct response;
- Verification feedback is an information concerning the correctness or incorrectness of the response, it can be separated into intentional feedback, external unintentional feedback and auto feedback;
  - Intentional feedback comes from something or someone that is external to the individual, such as a teacher or a device enabling the transmission of information; it allows the learner to have a teacher directly providing the necessary information;
  - Unintentional external feedback is the direct consequence of natural interactions with the physical or social environment. They can provoke auto feedback in the individual, giving him or her food for self interpretation;
  - Auto feedback is generated by the individual directly during or at the end of the action. These are the individual's own feelings and thoughts about the processes involved in performing the task. Its impact is difficult to calibrate, given the subjectivity of thought.
- Elaborated verification feedback is a mixture of the two previous sub-categories. They provide the correct answer, information explaining why the given answer is incorrect and why the expected answer is correct. This makes it possible to compare the answer given with an expert answer. An example would be the case-based reasoning feedback. This takes the current given answer and searches the database for past (or expert) answers, proposing the nearest "neighbor" for comparison. It has the ability to recall what the expert had proposed in this case, with parameters as close as possible.
  Each context and task allow one to define the appropriate answer for each question. Customized solutions for each category of learners are also necessary to optimize learning.

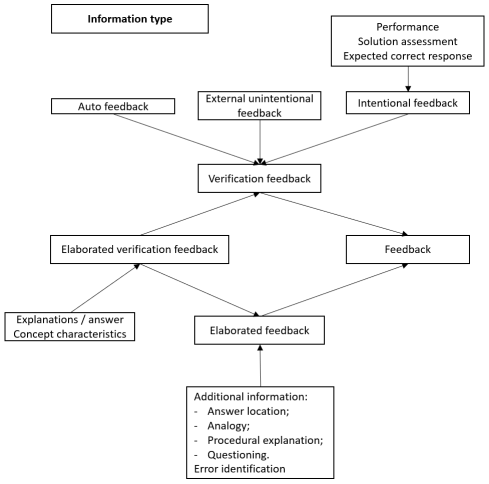

Fig. 2. Translation of the Bosc-Miné [30] model of feedback in learning. Feedback is divided into different sub-feedbacks. Each sub-feedback is associated with certain characteristics and contexts. The black arrow represents an heirloom. The source of the arrow is a subpart of the destination.

All these forms of feedback have common characteristics: from whom should it be given ? (itself,something,someone), how is it connoted ? (positively, negatively), what is focused on ? (task, process, self-regulation, person), presented with ? (alone, several), when should it be given ? (immediately, delayed), how should it be formalized ? (the solution, assessment criteria, individual characteristics).
Therefore, we propose to sort and integrate these different characteristics into a synthetic model of trust evolution. Our interest is to show that there are two temporalities in the evolution of trust and that different types of feedback need to be put in place to have a relevant effect on trust. This will be discussed in the next section.

III. Model

Our model incorporates elements from the literature such as the Parasuramann, Mayer, Lee & See and Hoff models, to which we add elements from the feedback description. Our interest lies in showing two types of feedback that affect trust over the course of interactions. These types of feedback act on two different temporalities, during the task and afterward (see Fig.3). The aim of immediate feedback is to maintain the

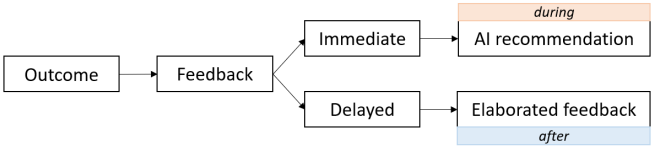

Fig. 3. Changes to the basic outcome. Feedback applies to two timeframes, each with its own specificities.

operator's awareness [18] on the mission. We want to avoid deterioration in mood or fatigue, which would degrade the operator's performance and make him less attentive to more critical tasks. Then, we have post-interaction feedback. This provides details on various elements encountered during the mission and highlights any disagreements. If the final state of an event is unknown, then this feedback would be constructed by an expert providing an inferred ground truth. This feedback incorporates the notion of learning.

There are a number of points to bear in mind. Immediate feedback must not be over-soliciting, i.e. it must not add workload on operator's in contexts that may already be overloaded. In terms of design integration, it must be simplistic and clear. It is also important to ensure that the first recommendations are not errors, as this could reduce the operator's trust to a level that would block further action [1]. For elaborate feedback, we need to be able to take into account the limitations of the field during experiments. Some knowledge are unknown, and we must not make the mistake of integrating it into experiments to establish whether a type of end-of-mission feedback is more effective than another. Being precise in the feedback certainly enables the learner to better integrate knowledge, but being as close as possible to the field gives a better understanding of the processes encountered.

We propose that immediate feedback should be in the form of a AI recommendation. Having direct feedback from the trustee seems to us to be important for the evolution of trust. This recommendation comes after the operator's initial choice and is there either to validate or refute the operator's initial decision. It fulfills the role of a decision-support AI, while the human still has the final decision. Therefore, this recommendation can influence human beliefs in AI. If the recommendation appears to be a mistake for the operator, then the beliefs through performance, process, and perceived goal will be degraded, while if the AI gives a recommendation in line with the initial decision, then its beliefs will be reinforced. No ground-truth information can be established at this stage of the interaction.

Regarding elaborated feedback, this takes as input all interactions captured by the system, including the event to be categorized (including all traces available up to that point), the operator's initial decision, the AI's recommendation, and the operator's final decision. Behavioral values such as reliance, compliance, and verification time are also captured. The aim is to compare, on the basis of all these factors, what decision an expert would have made in this specific case. This feedback contains a section on explicability that is essential for optimizing the way we explain what happened at a precise moment. The approach to explicability needs to be calibrated according to the target audience: some will prefer a mathematical approach, others a schematic or literary one.

Let's now turn to the integration of these terms into a model of trust evolution. The first phase occurs during interaction. A task is to be carried out, embodied either in a physical manifestation or via a computerized interface. In both cases, this task contains information that the AI can process so that it can make a recommendation. This recommendation, whether in dissonance or agreement with the initial decision, will modify the operator's beliefs (degrade or reinforce). The task will be added to a pre-existing or null task list. This addition increases the situational risk [16], as it increases the number of iterations to be processed. If the operator perceives this new iteration, then the perceived risk will also increase. There may be a dissonance between ground-truth and perceived risk. Perceived risk [2] is the only one that has an impact on a decision to trust or not. Next, we have the behavior related to trust. This behavior is affected by factors linked to the operator (such as perceived risk [16], self-confidence [31], workload [18], trust, beliefs), the environment (such as the event to be treated) and the system (such as the AI recommendation [10]). The elaborated feedback is fed by the trust-related behavior, the AI's recommendation, the initial decision, and the task information to be performed. As long as the mission is in progress, we have this interaction cycle.

The second phase takes place after the interaction. The elaborated feedback, previously fed by all the events that took place during the interaction, provides information to analyze any inconsistencies or dissonances. This feedback must be provided by an element external to the Trustor-Trustee pair, like another human or a device deemed effective. This feedback will have an impact not only on the trustor's beliefs towards the trustee but also on his self-confidence. It is a way of taking a step back and reflecting on the events that took place during the interaction.

In the following section, we propose to apply this model to the application context we will be experimenting with in the future: maritime surveillance.

IV. Maritime Surveillance Use Case

Maritime surveillance is a cognitively challenging work domain that involves high operational tempo and information overload. The crews of the maritime patrol aircraft are tasked with carrying out missions to control human trafficking, drugs, dangerous substances, illegal fishing, and sea rescue.

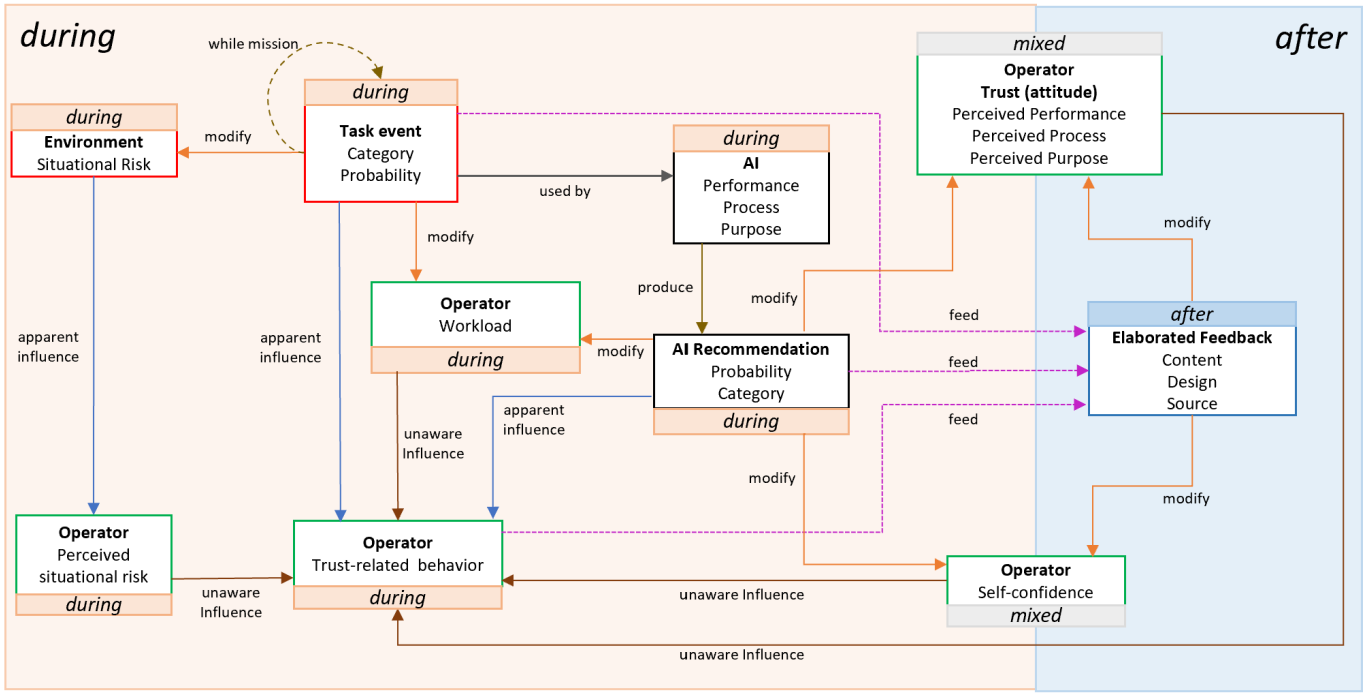

Fig. 4. Synthetic model of the evolution of trust. Two temporalities separate this model during (light orange) and after the task (light blue). The dynamics of this model arise from a task-related event. Red rectangles represent an environment-related state, green a human-related state, and blue a mixture of the three. Two types of arrow are visible. Solid arrows represent an influence on a process linked to the human, the system, or the environment. Orange is a modification of a state, blue is an apparent influence, i.e. perceived by the operator, and brown is a non-perceptible (or at least non-intellectualized) influence. The dotted arrows represent a transfer of information with no direct influence on any of the three actors.

Currently equipped with electronic surveillance equipment used manually by maritime surveillance operators, Thales is looking to integrating artificial intelligence tools to increase mission performance over the long term, to reduce the risk of missing illegal activities, and to increase operator resilience during high-pressure phases. To validate the value of such cooperation, we will re-use the application developed by Thales Defence Mission Systems (Thales DMS) and Thales Research Technology Canada (TRT Canada), which integrates Cognitive Shadow (TRT Canada) with AMASCOS (Thales DMS). AMASCOS is a tactical display that represents a situational context for maritime surveillance (see Fig.5). Cognitive Shadow (CS) [32] is a decision support AI composed of seven different supervised learning algorithms including Naive Bayes, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Neural Network, and Random Forest. A configurable model evaluation and aggregation method is integrated to select the best recommendation amongst these algorithms. All these algorithms can be trained on classifications made by experts and developed in consultation with them to determine which information is relevant to a specific task (target classification in our case). In figure 5, the operator is represented by the aircraft in the middle of the map. A radar detection field is used to detect vessels characterized by fourteen features that are a mix of categorical and numerical attributes, which include: platform type, speed, speed change, stationary, length, friend list, AIS(on-off), nearest track distance, cluster size, coastal proximity, interception, sea lane deviation, heading change, and nationality. With this information, the operator must be able to make decisions about the category of vessels such as 'Allied', 'Neutral', or 'Suspect'. At the end of an operation, the operator team carries out debriefs to see if they have made any mistakes or if they have made any clarifications that could help them improve for the next time. The problem is that they have to wait until the end of the mission to realize whether or not they may have made mistakes. Here comes CS. It is integrated in such a way that it checks an operator's classification after an initial decision has been made. If the check does not agree with the operator's decision, and a certain probability threshold is exceeded, then CS provides a recommendation for the operator to re-check this target. This enables the operator to keep an eye on certain situations where he may be less attentive due to factors linked to himself (fatigue, stress) or to the environment (shake, noise, etc.). The recommendation creates dissonance in the operator, enabling him to refocus on the mission, to question himself, and to evolve his trust in the AI. In this very specific case, the elaborated feedback produced will be equivalent to the discussions operators have at the end of a mission to discuss disagreements concerning targets encountered. The point is to have a report to back up discussions and to show when the operator was right to change his mind or not. The ground-truth must be fed by another expert. The linguistic content will be generated procedurally

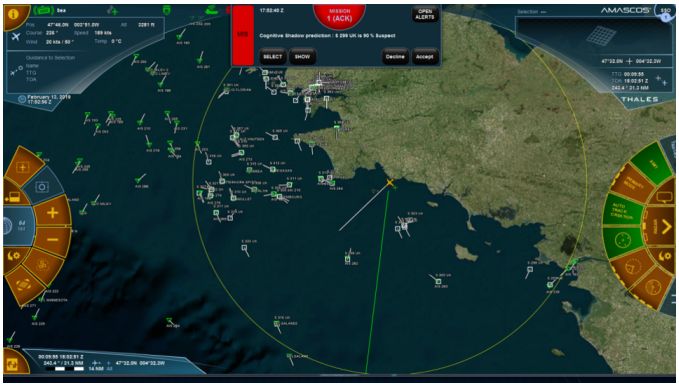

Fig. 5. AMASCOS interface representing an operational situation with a Cognitive Shadow recommendation showing that it is a suspect target.

to explain the data recovered during the mission.

V. Conclusion

The purpose of this study was to understand what trust is and what implications this has for the Human-AI pair when performing tasks in risky environments. The second aim was to introduce feedback terms into a model of trust evolution and to explain how different types of feedback can be introduced to optimize trust over time. For trust, we have drawn on general reference models covering all influencing factors, with Lee & See's [3] article as the main reference. Regarding feedback, the central article is by Bosc-Miné [30], in which feedback for learning is deconstructed to establish what can best be transposed to trust. We then presented a synthetic model of the evolution of trust, describing the main antecedents that have an impact on trust, or its behavior, over time. In particular, the perceived risk [16] described in Lee & See's [3] definition as a prerequisite for the study of trust. Concerning behavior, it is therefore associated with trust, but also with other factors such as self-confidence [31], workload [18], perceived risk [16], or the type of event to be dealt with. Two temporalities, during and after, allow us to target the types of feedback to be provided that will modify trust and self-confidence. The feedback from AI during the interaction will help the operator stay alert over time and reduce mental workload. Elaborated feedback from either the system (non-AI) or a person (external to the pair) helps to modify the operator's perceived beliefs (the basis of trust) of the AI and his own self-confidence. Other concepts related to trust have not been addressed, or only very briefly, such as explainability, transparency, acceptability, and so on. Explainability [7] is a major topic on its own with the goal of helping humans understand the basis of AI decisions. Different forms are being explored to determine which one will enable the most informed decisions to be made in the one-way relationship that is the Human-AI couple. Transparency, part of the explicability, aims to show what lies behind AI models. A definition of transparency provided by Chen [33] is "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process.". It has been shown that transparency can increase operator trust but, in return, can increase mental workload, which implies a design choice to best balance the impact of integrating transparency. Acceptability refers to the degree of acceptance of a tool's integration in a particular environment. It can be considered as an attitude or as behaviors [11]. The result is a level of trust in a tool. Beggiato et al. [11] show that acceptance evolves in the same way as trust during the learning phases. However, they are two different terms.

To the end, a case study applied to maritime surveillance was used to transpose the theory onto something more concrete. In this case, we will conduct experiments to study the evolution of trust over several weeks and validate this theoretical model proposed in this article. We will compare the impact of different types of feedback (immediate and delayed) and AI performance on operator trust over time.

References

[1] E. J. De Visser, P. J. Beatty, J. R. Estepp, S. Kohn, A. Abubshait, J. R. Fedota, and C. G. McDonald, "Learning from the slips of others: Neural correlates of trust in automated agents," *Frontiers in human neuroscience*, vol. 12, p. 309, 2018.

[2] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human factors*, vol. 39, no. 2, pp. 230–253, 1997.

[3] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[4] M. Hou, G. Ho, and D. Dunwoody, "Impacts: A trust model for human-autonomy teaming," *Human-Intelligent Systems Integration*, pp. 1–19, 2021.

[5] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Human factors*, vol. 58, no. 3, pp. 377–400, 2016.

[6] A. D. Kaplan, T. T. Kessler, J. C. Brill, and P. Hancock, "Trust in artificial intelligence: Meta-analytic findings," *Human factors*, vol. 65, no. 2, pp. 337–359, 2023.

[7] T. Saßmannshausen, P. Burggräf, J. Wagner, M. Hassenzahl, T. Heupel, and F. Steinberg, "Trust in artificial intelligence within production management–an exploration of antecedents," *Ergonomics*, vol. 64, no. 10, pp. 1333–1350, 2021.

[8] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human factors*, vol. 57, no. 3, pp. 407–434, 2015.

[9] S. Sousa and G. Beltrão, "Factors influencing trust assessment in technology," in *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part V 18*, pp. 416–420, Springer, 2021.

[10] E. Avril, "Providing different levels of accuracy about the reliability of automation to a human operator: impact on human performance," *Ergonomics*, vol. 66, no. 2, pp. 217–226, 2023.

[11] M. Beggiato, M. Pereira, T. Petzoldt, and J. Krems, "Learning and development of trust, acceptance and the mental model of acc. a longitudinal on-road study," *Transportation research part F: traffic psychology and behaviour*, vol. 35, pp. 75–84, 2015.

[12] J. B. Lyons, G. G. Sadler, K. Koltai, H. Battiste, N. T. Ho, L. C. Hoffmann, D. Smith, W. Johnson, and R. Shively, "Shaping trust through transparent design: theoretical and experimental guidelines," in *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA*, pp. 127–136, Springer, 2017.

[13] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 295–305, 2020.

[14] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[15] S. Rice and D. Keller, "Automation reliance under time pressure.," *Cognitive Technology*, 2009.

[16] S. Hoesterey and L. Onnasch, "The effect of risk on trust attitude and trust behavior in interaction with information and decision automation," *Cognition, Technology & Work*, vol. 25, no. 1, pp. 15–29, 2023.

[17] T. Sato, Y. Yamani, M. Liechty, and E. T. Chancey, "Automation trust increases under high-workload multitasking scenarios involving risk," *Cognition, Technology & Work*, vol. 22, pp. 399–407, 2020.

[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs," *Journal of cognitive engineering and decision making*, vol. 2, no. 2, pp. 140–160, 2008.

[19] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw, "Measurement of trust in automation: A narrative review and reference guide," *Frontiers in psychology*, vol. 12, p. 604977, 2021.

[20] T. Rieger and D. Manzey, "Human performance consequences of automated decision aids: The impact of time pressure," *Human factors*, vol. 64, no. 4, pp. 617–634, 2022.

[21] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[22] R. C. Mayer and J. H. Davis, "The effect of the performance appraisal system on trust for management: A field quasi-experiment.," *Journal of applied psychology*, vol. 84, no. 1, p. 123, 1999.

[23] K. Akash, W.-L. Hu, N. Jain, and T. Reid, "A classification model for sensing human trust in machines using eeg and gsr," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1–20, 2018.

[24] T. Xu, A. Dragomir, X. Liu, H. Yin, F. Wan, H. Wang, *et al.*, "An eeg study of human trust in autonomous vehicle basing on graphic theoretical analysis," *Frontiers in Neuroinformatics*, p. 70, 2022.

[25] Y. Lu and N. Sarter, "Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 560–568, 2019.

[26] K. Akash, W.-L. Hu, T. Reid, and N. Jain, "Dynamic modeling of trust in human-machine interactions," in *2017 American Control Conference (ACC)*, pp. 1542–1548, IEEE, IEEE, 2017.

[27] S. W. Jaffry and J. Treur, "Comparing a cognitive and a neural model for relative trust dynamics," in *Neural Information Processing: 16th International Conference, ICONIP 2009, Bangkok, Thailand, December 1-5, 2009, Proceedings, Part I 16*, pp. 72–83, Springer, 2009.

[28] E. T. Chancey and M. Politowicz, "Designing and training for appropriate trust in increasingly autonomous advanced air mobility operations: A mental model approach: Version 1," 2020.

[29] O. Vereschak, G. Bailly, and B. Caramiaux, "How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–39, 2021.

[30] C. Bosc-Miné, "Caractéristiques et fonctions des feed-back dans les apprentissages," *L'Année psychologique*, vol. 114, pp. 315–353, June 2014.

[31] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *International journal of human-computer studies*, vol. 40, no. 1, pp. 153–184, 1994.

[32] D. Lafond, K. Labonté, A. Hunter, H. F. Neyedli, and S. Tremblay, "Judgment analysis for real-time decision support using the cognitive shadow policy-capturing system," in *Human Interaction and Emerging Technologies: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHIET 2019), August 22-24, 2019, Nice, France*, pp. 78–83, Springer, 2020.

[33] J. Y. Chen, K. Procci, M. Boyce, J. L. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," tech. rep., ARMY RESEARCH LAB ABERDEEN PROVING GROUND MDUNIVERSITY OF CENTRAL FLORIDA ..., 2014.