CIGI QUALITA MOSIM 2023
Transparence et Confiance au sein des équipes Humains-Systèmes de l'Industrie 4.0

PHILIPPE RAUFFET[1], LOÏCK SIMON[1], CLEMENT GUERIN[1]

[1] Equipe FHOOX, Lab-STICC UMR CNRS 6285, UNIVERSITE BRETAGNE SUD
17, Boulevard Flandres Dunkerque, 56100 LORIENT
{ philippe.rauffet, loick.simon, clement.guerin}@univ-ubs.fr

***Résumé* – L'industrie 4.0 voit émerger de nouvelles équipes hybrides, constituées d'opérateurs humains et de systèmes cyber-physiques intelligents et autonomes. Au sein de ces équipes se posent de nouvelles questions sur la coopération, et notamment celle de la confiance de l'humain dans l'agent artificiel, lequel peut faire des recommandations en utilisant des données ou des capacités de traitement dont ne dispose pas l'humain. Cet article investigue cette question de la confiance au travers de deux cas d'étude, en modélisant les différentes dimensions de la confiance, en analysant leur influence sur l'acceptation de la recommandation, et en étudiant si la transparence de l'agent modifie cette confiance.**

***Abstract* – Industry 4.0 is seeing the rise of new hybrid teams, made up of human operators and intelligent and autonomous cyber-physical systems. Within these teams, new questions arise about cooperation, particularly about human trust in the artificial agent, where AI can make recommendations using data or processing capabilities that human does not have. This paper investigates the question of trust with 2 case studies, by modeling the different dimensions of trust, examining their influence on recommendation acceptance, and studying if agent transparency changes trust.**

***Mots clés* – Confiance dans le système autonome, Transparence de l'agence, Coopération Humain-Machine**

***Keywords* – Trust in autonomy, Agent Transparency, Human-Machine Cooperation**

1 INTRODUCTION

Le paradigme Human-Autonomy Teaming, appliqué à l'industrie 4.0, fait référence à l'intégration de travailleurs humains et de systèmes autonomes, tels que les robots et l'Intelligence Articifielle (IA), pour accroître l'efficacité, la productivité et la compétitivité globale dans les milieux industriels. De telles équipes tirent parti des compétences et des forces complémentaires des humains et des machines pour effectuer des tâches et prendre des décisions, ce qui permet d'améliorer les performances et la satisfaction des clients ([Pacaux-Lemoine et al., 2017], [Rauffet, 2021], [Guérin et al., 2019], [McNeese et al., 2018]). Dans cette nouvelle ère de la production manufacturière, les équipes humains-systèmes autonomes jouent un rôle crucial pour façonner l'avenir du travail et de la compétitivité dans l'industrie.

Dans de nombreuses situations industrielles, ces équipes hybrides sont structurées selon une coopération verticale, dans laquelle un agent hiérarchique de niveau supérieur possède l'autorité, et supervise un agent de niveau inférieur pouvant fournir des conseils (cf. [Lemoine et al., 1996]).

Plus particulièrement, l'IA peut analyser la situation, puis émettre un signal (comme une proposition d'action) à un opérateur humain. L'homme peut alors, ou non, faire confiance et décider de suivre l'IA, après avoir intégré ou débattu les conseils de l'agent autonome avec son propre modèle mental.

1.1 Modèle de la confiance humain-système et de l'acceptation des conseils de l'IA

Selon Lee et See (2004), l'humain est donc celui qui donne sa confiance (le « fiduciant », ou le "trustor" en Anglais), l'IA celui qui reçoit la confiance (le « fiduciaire », ou le "trustee"), et la relation de confiance entre eux est très dépendante des caractéristiques de la situation (objectifs de la tâche, contraintes de l'environnement).

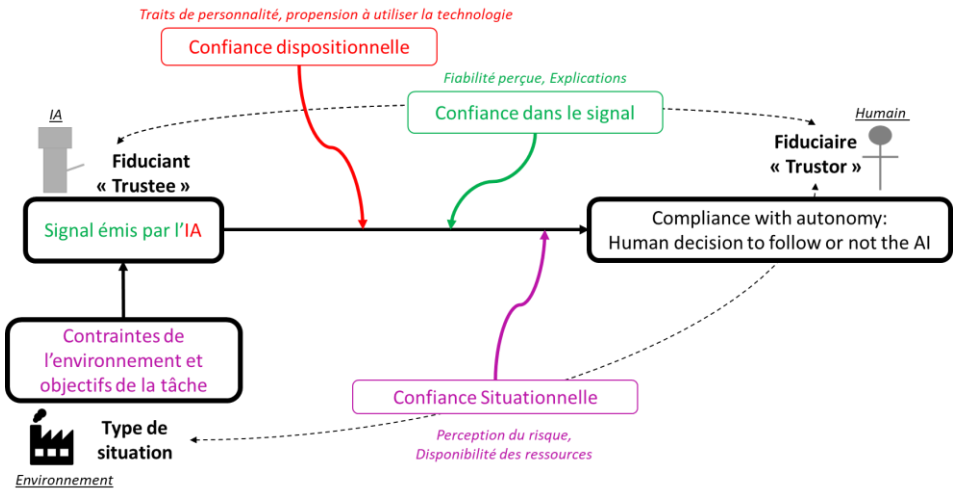

Figure 1: 3 types de la confiance humain-système, comme inducteur de l'acceptation des conseils de l'agent autonome

Par ailleurs, le fait que l'humain suive ou non les conseils de l'IA peut être influencé (cf. [Hoff & Bachir, 2015], [Chancey et al., 2017]) par trois types de confiance qui sont tous liés entre eux, comme le montre la figure 1. On distingue ainsi :

- La **confiance dispositionnelle** : il s'agit de la confiance a priori dans la technologie, motivée par des traits personnels, tels que l'âge, le sexe, mais aussi la propension à faire confiance à la technologie, etc.
- La **confiance dans le signal** : elle est aussi appelée confiance apprise (cf. [Hoff & Bachir, 2015]). Ce type concerne la façon dont l'humain perçoit la qualité de son expérience dans l'interaction avec l'IA. Elle est notamment influencée par la fiabilité perçue de l'IA (juste avant et pendant les interactions), mais aussi par les raisonnements que l'IA peut fournir.
- - Enfin, la **confiance situationnelle** : il s'agit de l'ensemble des éléments liés à l'environnement qui modifient la confiance, comme le risque perçu d'une situation donnée, ou le temps et les ressources disponibles pour l'opérateur humain.

De plus, selon Chancey et al. (2017), il pourrait y avoir un effet modérateur entre la confiance dans le signal et la confiance situationnelle. En effet, selon ces auteurs, les humains décident d'accepter ou non les conseils l'agent principalement sur la base de la confiance dans le signal. Cependant, ce comportement d'acceptation pourrait varier, notamment à cause de la perception du risque (confiance situationnelle) qui interfère avec la confiance dans le signal.

1.2 La transparence de l'agent pour corriger et calibrer la confiance humain-système

D'après les travaux récents sur le Human-Autonomy Teaming ([Lyons et al., 2016], [Wright et al., 2017]), **la transparence de l'agent** est un facteur clé qui va influer sur la volonté de l'humain d'interagir avec un agent autonome, comme un robot. De plus, la transparence permet de développer la capacité des robots à inspirer confiance et à être considérés comme un coéquipier légitime. Manipuler la transparence des agents pourrait donc être un moyen de corriger une confiance mal ajustée (une méfiance injustifiée, ou au contraire un excès de confiance), et même un moyen de calibrer la confiance ([Wang et al., 2016]), c'est-à-dire de trouver un moyen de rapprocher le taux d'acceptation des conseils de l'agent par l'humain avec la fiabilité réelle du système.

Comme le suggèrent deux revues de littérature récentes sur la transparence ([Bhaskara et al., 2020], [Rajabiyazdi & Jamieson, 2020]), deux approches principales ont été adoptées dans la littérature pour décrire et opérationnaliser le concept de transparence : La transparence des agents basée sur la conscience de la situation de Chen et al. (2014), connue sous le nom de modèle SAT, et le cadre de transparence de Lyons (2013) pour l'interaction homme-robot.

Le modèle SAT intègre trois niveaux de transparence des agents, basés sur la théorie de la conscience de la situation ([Endsley, 1995]). Au **niveau L1**, l'agent fournit des informations de base sur son état actuel, ses objectifs, ses intentions, ses plans, ses progrès, ses actions actuelles et proposées. Au **niveau L2**, l'agent fournit des arguments qui justifient son action ou sa décision. À ce deuxième niveau, l'opérateur humain reçoit des informations sur le raisonnement de l'agent, ses capacités comportementales et les contraintes qu'il prend en compte. Enfin, **au niveau L3**, l'agent fournit une projection des résultats futurs. À ce niveau, l'humain reçoit des informations concernant l'anticipation de l'état futur par l'agent, les conséquences prévues et les incertitudes.

Selon Lyons (2013), la transparence peut également être considérée selon différents modèles liés aux situations de coopération, regroupés en deux dimensions principales. D'une part, la transparence du robot vers l'humain (robot-TO-human) concerne les informations sur l'agent qui sont communiquées à l'humain. Dans ce cas, l'agent peut être transparent sur son intention ou son objectif (**modèle intentionnel**), sur la tâche en cours ou les tâches précédentes effectuées (**modèle de tâche**), sur les processus effectués qui ont conduit à une décision ou à une action (**modèle analytique**), ou sur les aspects de l'environnement (**modèle de l'environnement**). D'autre part, la transparence du robot sur l'humain (robot-OF-humain) concerne la conscience du robot sur "les autres" qui est communiquée à l'opérateur humain. Ici, l'agent peut être transparent sur sa perception de l'état de l'opérateur (**le modèle de l'opérateur**), ou sur la répartition des tâches (**le modèle du travail d'équipe**), en précisant qui est responsable d'une tâche ou d'un ensemble de tâches.

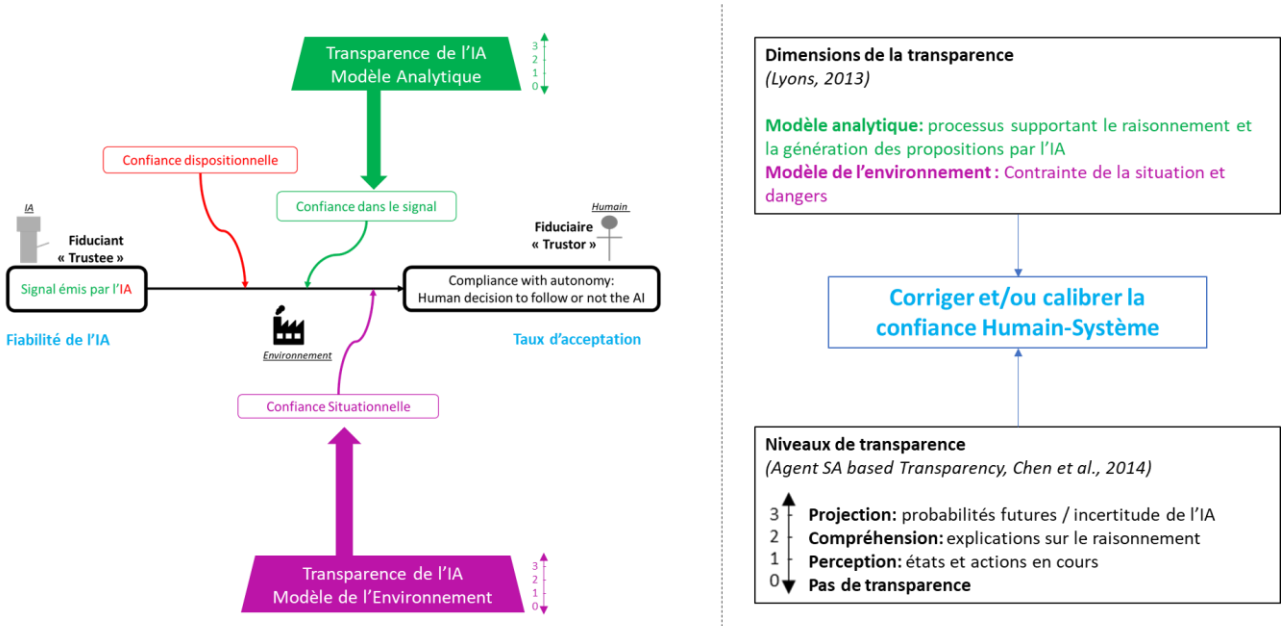

Figure 2: La transparence de l'agent, un modérateur de la confiance dans le signal et de la confiance situationnelle

1.3 Questions de recherche

A partir des différents concepts et modèles issus de la littérature que nous venons de présenter ci-dessus, cet article vise à répondre à trois principales questions ou hypothèses de recherche.

R1. Nous cherchons à étudier comment chacun des trois niveaux de confiance a un impact réel sur la conformité à l'IA.

R2. En outre, conformément à la littérature et aux hypothèses formulées par Chancey et al. (2017), nous souhaitons explorer comment la confiance situationnelle peut modérer l'effet de la confiance dans le signal sur la conformité à l'autonomie.

R3. Enfin, comme illustré dans la Figure 2, nous supposons que l'acceptation des conseils de l'IA, ainsi que les différentes confiances de notre modèle (dispositionnelle, dans le signal et situationnelle), peuvent être corrigées et améliorées en faisant varier la transparence de l'agent, à différents niveaux et sur différentes dimensions. En particulier, la transparence liée au risque de la situation et à la charge de travail de l'opérateur (fournie par les modèles de l'opérateur, du travail d'équipe ou de l'environnement) pourrait modifier la confiance situationnelle, et la transparence liée aux processus de l'agent générant les recommandations (fournie par les modèles analytique, intentionnel ou de la tâche) pourrait modifier la confiance dans le signal.

2 MÉTHODE

2.1 La question de la confiance et de la coopération à différents niveaux d'activité de l'industrie 4.0

L'industrie 4.0 a fait émerger de nouveaux systèmes humain-machine, composés d'opérateurs et de composants cyber-physiques, dont les interactions multiples et multi-niveaux (aux niveaux opérationnel, tactique et stratégique des activités de l'entreprise) contribuent à surveiller ou à contrôler les systèmes de fabrication. Ces différentes interactions peuvent être comprises à travers des cadres théoriques.

Schmidt (1991) adopte ainsi une approche fonctionnelle-structurelle en distinguant trois types de coopération :

- **La coopération augmentative** vise à accroître les capacités physiques ou intellectuelles puisque des agents supplémentaires aux compétences identiques effectuent la tâche lorsque la charge de travail augmente et qu'elle ne peut être gérée par un seul agent.
- **La coopération débative ou confrontative** permet la comparaison des points de vue entre les agents afin d'améliorer les solutions et de réduire les erreurs. Ce type de coopération nécessite que certains agents effectuent des vérifications et contrôlent d'autres agents. Comme dirait Ronald Reagan, la « confiance n'exclut pas le contrôle ».
- Enfin, **la coopération intégrative** vise la complémentarité d'agents aux compétences différentes.

Par ailleurs, selon l'approche fonctionnelle de Hoc (2001), la coopération entraîne des interférences entre l'agent humain et l'agent artificiel. Ces interférences doivent être gérées à trois niveaux de coopération :

- **La coopération dans l'action.** Ce niveau distingue différentes activités de coopération opérationnelle pour la gestion des objectifs et des procédures en temps réel et à court terme : création d'interférences locales (par exemple, désaccord), détection d'interférences locales (par exemple, redondance), anticipation et résolution des interférences.
- **La coopération dans la planification.** Ce niveau est caractérisé par des activités de coopération tactique pour la gestion d'une référence commune : maintien et développement d'un objectif commun, d'un plan commun, d'une division des fonctions.
- La Méta-coopération. Ce niveau facilite les deux niveaux précédents en développant un code de communication commun et des modèles de soi et du partenaire.

2.2 Deux cas d'étude pour explorer la transparence de l'agent et la confiance dans l'agent

Les travaux de recherche présentés ci-après dans le document sont issus de deux projets nationaux français.

D'une part, le projet **SEANATIC,** financé par l'ADEME, a réuni les expertises de partenaires académiques et industriels : le Lab-STICC UMR CNRS 6285, Azimut, IoT.bzh, Thalos et Piriou. Le projet portait sur le développement d'un ensemble complet de solutions, pour collecter, analyser et présenter des données techniques, afin d'aider les mécaniciens et les gestionnaires de flotte à améliorer la maintenance maritime préventive, en adoptant de nouveaux outils intelligents et prédictifs basés sur l'apprentissage automatique des données collectées.

D'autre part, le projet HUMANISM, financé par l'ANR, a été réalisé par trois partenaires universitaires : le Lab-STICC UMR CNRS 6285, le CReSTIC EA 3804, et le LAMIH UMR CNRS 8201. Les objectifs du projet étaient de modéliser et de développer de nouveaux mécanismes d'allocation de fonctions, ainsi que de dialogue et d'IHM, pour faciliter la coopération human-machine dans l'industrie 4.0 avec des robots intelligents et des IA, à un niveau opérationnel, tactique ou stratégique.

Le tableau 1 énumère les caractéristiques de ces deux études de cas qui sont présentées dans les sections 3 et 4, et souligne leurs différences, afin de considérer la question de la confiance homme-autonomie le long de différents paramètres (différents niveaux de gestion de l'activité, différents niveaux de précision de l'agent, etc.)

Tableau 1: caractéristiques des deux cas d'étude

PROJET	ANR Humanism	ADEME Seanatic
ACTIVITE INDUSTRIE 4.0	Préparation de commande sur différents postes sur une ligne de production	Planification des opérations de maintenance sur un navire de la marine marchande
TYPES DE COOPERATION HUMAIN-MACHINE	*Coopération confrontative* : l'agent propose des actions, l'humain décide de les accepter ou non *Coopération integrative :* l'agent collecte, traite et partage des informations auxquelles l'humain peut difficilement accéder	
NIVEAU D' INTERFERENCES A GERER DANS LA COOPERATION	Coopération dans l'action, à un niveau opérationnel	Cooperation dans la planification, à un niveau tactique
CAPACITES ET FIABILITE DE L'AGENT AUTONOME	Robot toujours myope (vue partielle du terrain)	IA avec un modèle prédictif très fiable
	Inexactitude systémique (à cause de la myopie) dans les recommandations	Grande exactitude (90%) dans les recommandations
CONNAISSANCE DES PARTICIPANTS SUR L'AGENT	Niveau de fiabilité non connu par les participants	Niveau de fiabilité connu par les participants

2.3 Evaluation de la confiance et de l'acceptation

Différents questionnaires et mesures sont proposés dans la littérature pour évaluer les trois types de confiance présentées ci-dessus, ainsi que le taux d'acceptation des recommandations de l'IA par les participants humains. Le tableau 2 liste les différentes mesures que nous avons utilisées à chaque fois dans les deux études présentées dans les sections 3 et 4.

Tableau 2: Mesures de confiance et d'acceptation

Type de confiance	*Métriques, questionnaires et références*
ACCEPTATION	Mesurée dans chaque situation après la décision du participant (acceptation ou rejet de la recommandation de l'agent).
CONFIANCE DANS LE SIGNAL	Mesurée au début puis après chaque situation d'interaction à l'aide de l'échelle IMOTRIS (traduction française de [Mayer et al, 1995], [Lyons & Guznov, 2019]). Mesurée à l'aide d'une échelle de likert à un item après chaque situation.
CONFIANCE SITUATIONNELLE	Perception du risque mesurée avec des échelles de perception du risque (telles que la traduction française de [Wilson, Zwickle & Walpole, 2018]). Charge mentale mesurée avec ISA (traduction française de Tatersall & Foord, 1996]), ou NASA-TLX (traduction française de [Hart & Staveland, 1988]) après chaque situation d'interaction.
CONFIANCE DISPOSITIONNELLE	Affinité du participant envers la technologie mesurée à l'aide de l'échelle ATI (traduction française de [Franke et al, 2019]). Propension du participant à faire confiance à la technologie mesurée à l'aide de l'échelle PTT (traduction française de [Jessup et al, 2018]). Propension générale au risque du participant mesurée à l'aide de l'échelle GriPS (traduction française de [Zhang et al, 2019]).

2.4 Traitement et analyse des données

Enfin, nous avons étudié les différentes relations entre les types de la confiance et les paramètres de transparence en calculant des régressions logistiques et des modèles linéaires mixtes. Nous avons particulièrement utilisé R studio, et notamment le package lme4 ([Bates & al., 2012]), pour les régressions logistiques, et les analyses des effets mixtes linéaires (fonction lmer pour l'étude des effets sur les variables continues). Pour les analyses des effets linéaires, l'inspection visuelle des tracés des résidus n'a pas révélé de déviations évidentes de l'homoscédasticité ou de la normalité. Pour tous les modèles statistiques, nous avons introduit les dimensions de la transparence et les caractéristiques de la situation comme effets fixes (avec des termes d'interaction) dans le modèle complet. Comme effet aléatoire, nous avions un intercept pour les participants. En ce qui concerne les effets fixes, une sélection de modèle par étapes par AIC (stepAIC) a été effectuée. À chaque étape, un nouveau modèle a été ajusté, dans lequel un des termes du modèle a été éliminé et testé par rapport au modèle précédent.

3 1ER CAS D'ETUDE : HUMAN-AUTONOMY TEAMING DANS UNE ACTIVITE OPERATIONNELLE DE PREPARATION DE COMMANDE

3.1 Expérimentations

Dans le projet Humanism, l'expérimentation a été conçue pour instancier une coopération entre des opérateurs humains et un cobot, préparant chacun différentes commandes clients sur leurs postes de travail respectifs, en partageant les mêmes ressources (des produits stockés dans un entrepôt, et amenées sur les différents postes à l'aide d'un anneau de convoyage). Régulièrement, le cobot alertait sur un possible problème d'interférence lié à cette situation de co-activité, prenant la forme d'une rupture de stock sur une ressource partagée (cf. Figure 3). Le cobot demandait aux participants, jouant le rôle de superviseur de l'équipe, un transfert de ressources du stock du coéquipier humain vers son propre stock. Cependant, cette demande pouvait être non pertinente en raison d'une myopie informationnelle, le cobot ne regardant que dans son propre stock sans vérifier la disponibilité des ressources critiques dans l'entrepôt général.

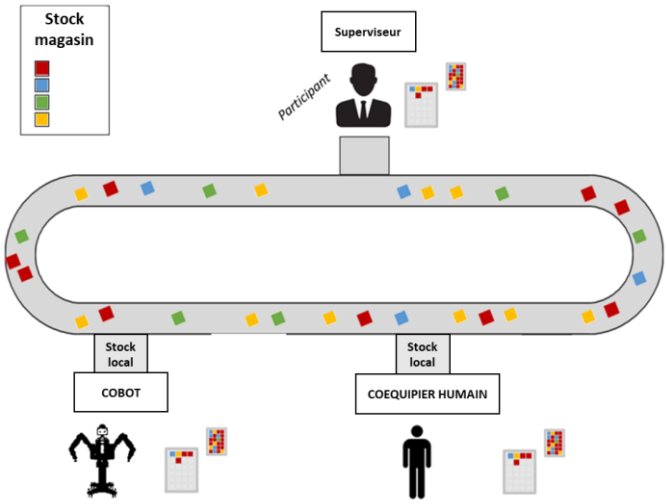

Figure 3 : Expérimentation sur la préparation de commandes 4.0

53 élèves ingénieurs, spécialisés en génie industriel (17 femmes, 36 hommes, âge moyen = 21,34 ans, écart-type = 1,67) ont participé à cette expérimentation. Ils ont été confrontés à des situations répétées, dans lesquelles la transparence variait en niveaux (cf. le cadre SAT de Chen et al.), le long de deux dimensions différentes (modèle Analytique et modèle Opérateur, cf. Lyons), comme expliqué dans le tableau 3.

Cette expérience visait à répondre à deux de nos questions de recherche :

- **R1.** Quel type de confiance a un effet significatif sur le taux d'acceptation des conseils de l'IA ?
- **R2.** Comment la confiance situationnelle modère-t-elle l'effet de la confiance dans le signal sur cette acceptation ?

Tableau 3 : Modalités de transparence dans Humanism

CONDITIONS	NIVEAU DE TRANSPARENCE SUR LE MODELE ANALYTIQUE (INFLUANT SUR LA CONFIANCE DANS LE SIGNAL)	NIVEAU DE TRANSPARENCE SUR LE MODELE DE L'OPERATEUR (INFLUANT SUR LA CONFIANCE SITUATIONNELLE)
S1	**A_L2 :** Le cobot alerte sur un problème potentiel de rupture de stock, mais sans transparence sur sa myopie. Le participant pourrait penser que Cobot a pris en compte le stock de l'entrepôt.	**O_L0 :** il n'y a pas de transparence sur la charge de travail des coéquipiers
S2		**O_L1+ :** le coéquipier humain est moins occupé que le cobot
S3		**O_L1- :** le coéquipier humain est plus occupé que le cobot, et la demande du cobot peut perturber l'activité humaine.
S4	**A_L3 :** Le cobot précise en plus qu'il ne tenait pas compte du stock de l'entrepôt, étant transparent sur sa myopie. Les participants sont donc certains des limites de cobot.	**O_L0 :** décrit ci-dessus
S5		**O_L1+ :** décrit ci-dessus
S6		**O_L1-:** décrit ci-dessus

3.2 Résultats

Cette première étude de cas a mis en évidence les principaux résultats suivants.

R1. Relations entre acceptation et types de confiance

Confiance dispositionnelle. Nous n'avons pas trouvé d'effet significatif de l'affinité du participant envers la technologie (en utilisant l'échelle ATI) et de la propension du participant à faire confiance à la technologie (en utilisant l'échelle PTT) sur la conformité.

Confiance dans le signal. Les analyses de Mann-Whitney ont révélé que la conformité et la confiance dans le signal sont significativement associées. Les participants qui acceptaient les demandes du cobot ont obtenu des scores plus élevés en ce qui concerne la compréhensibilité du robot (W=50846, p = .008), la fiabilité (W=51383, p = .004) et la confiance dans le signal (W=55450, p < .001), par rapport aux participants qui ont rejeté les requêtes du cobot.

Confiance situationnelle. De même, il existe une association significative entre l'acceptation des propositions du robot et la perception du risque. Les participants acceptant la requête du cobot ont rapporté une perception du risque plus faible (W=23378, p=.038) que ceux qui l'avaient refusée.

R2. Modération de la confiance situationnelle sur l'effet de la confiance dans le signal sur l'acceptation

Le taux d'acceptation des requêtes du cobot était significativement plus élevé lorsque cobot n'était pas transparent sur la situation (condition O_L0, sans information sur l'activité du coéquipier humain), par rapport à la condition O_L1-, où le cobot était transparent sur le fait que le coéquipier est plus occupé que lui (OR = 8,47, p < .001).

De plus, un effet d'interaction a été observé entre la confiance situationnelle et la confiance dans le signal sur l'acceptation. En effet, un cobot, faiblement transparent sur son modèle analytique puisque n'indiquant pas sa myopie, et transparent sur une perturbation potentielle du coéquipier humain (A_L2 et O_L1-) a entraîné un taux d'acceptation significativement plus faible qu'un cobot très transparent sur son manque de fiabilité dû à sa myopie et sur une situation positive du coéquipier humain (A_L3 et O_L1+) (OR = 7.29, p < .01). Cette baisse de l'acceptation a également été observée en comparant (A_L2 et O_L1-) à (A_L3 et O_L0, sans transparence sur l'activité du coéquipier humain), avec un effet de tendance (OR = 2.5, p = .08).

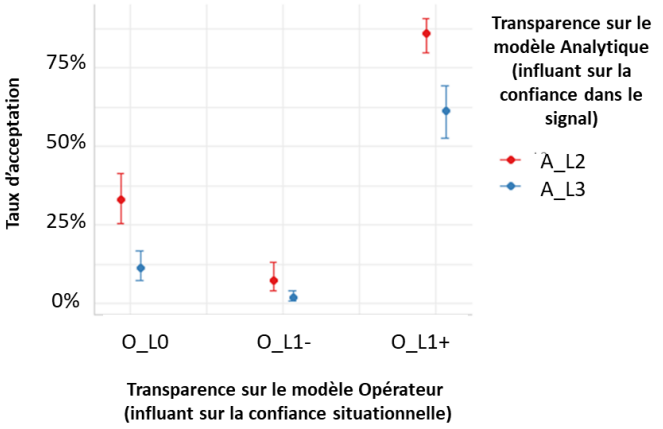

Figure 4 : Interactions entre les types de confiance

4 2EME CAS D'ETUDE : HUMAN-AUTONOMY TEAMING DANS UNE ACTIVITE TACTIQUE DE PLANIFICATION DE MAINTENANCE

4.1 Expérimentations

Dans le projet Seanatic, nous avons conçu un scénario, dans lequel un humain et une IA coopèrent pour la planification de la maintenance ([Simon et al., 2021]). L'IA peut proposer d'avancer ou de reporter certaines opérations, et l'humain décide d'accepter cette modification, ou de la refuser en conservant la date initiale de l'outil de GMAO (cf. Figure 5).

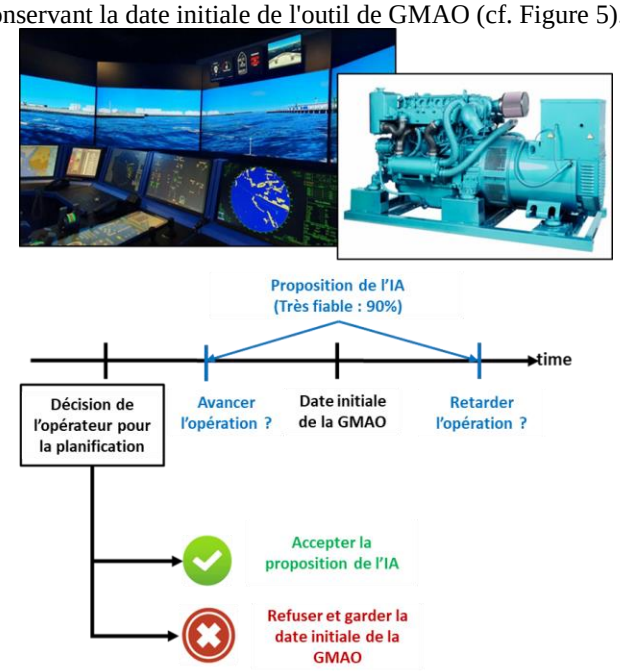

Figure 5 : Expérimentation sur la planification de la maintenance

39 participants (25 hommes, 14 femmes, âge moyen = 22,15 ans, ET = 2,77), élèves ingénieurs en génie industriel, ont été confrontés à des situations répétées, où l'IA suggérait différentes décisions (avancer ou reporter une opération de maintenance) dans différents contextes (composant critique ou non à changer), avec différents niveaux de transparence de l'agent. En plus de la transparence par défaut de l'IA sur les modèles analytiques et environnementaux dans chaque situation, comme illustré dans la Figure 6, des informations ont été ajoutées, soit sur la fiabilité du système (dans la condition « Fiab », affichant des informations soutenant la confiance dans le signal), soit sur le risque environnemental sur l'équipement et la logistique (dans la condition « Risque », affichant des informations liées à la confiance situationnelle). Une dernière condition expérimentale (« Fiab&Risque ») combinait les deux paramètres précédents (pour plus de détails, voir [Simon et al., 2022]).

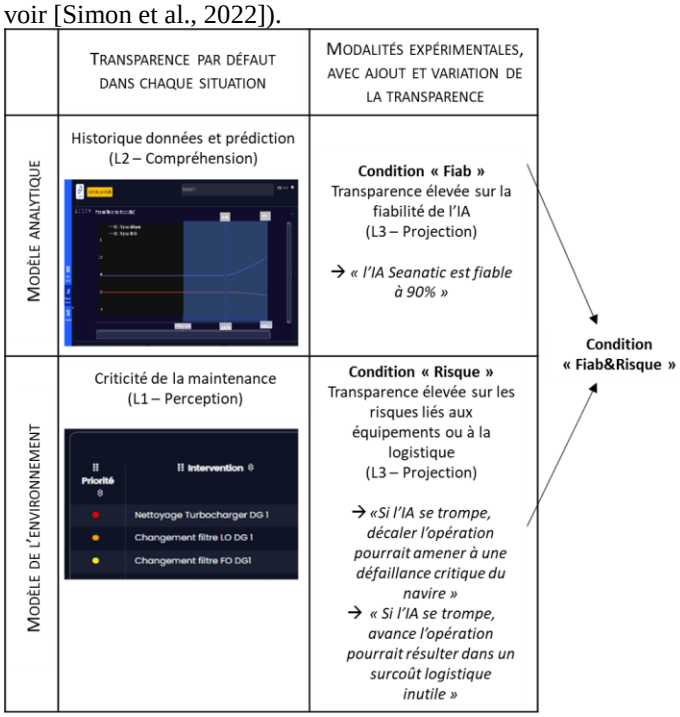

Figure 6 : Modalités de transparence dans Seanatic

Dans cette expérimentation, nous cherchons à répondre à deux de nos principales questions de recherche :

- **R1.** Quel(s) type(s) de confiance induit un changement de l'acceptation de l'IA par l'opérateur humain ?
- **R3.** Comment les paramètres de transparence peuvent-ils modifier et corriger les différents types de confiance ?

4.2 Résultats

Cette deuxième étude de cas sur la planification de la maintenance a mis en évidence les résultats suivants.

R1. Relation entre les types de confiance et l'acceptation

Confiance dispositionnelle. Nous n'avons pas trouvé d'effet significatif de l'affinité du participant envers la technologie (en utilisant l'échelle ATI) et de la propension du participant à faire confiance à la technologie (en utilisant l'échelle PTT) sur le taux d'acceptation de l'IA par les participants.

Confiance dans le signal. Les analyses de Mann-Whitney ont montré que la confiance dans le signal et l'acceptation sont significativement associées, avec une confiance plus élevée pour les participants suivant les suggestions de l'outil de maintenance prédictive (W = 3857.5, p < .001).

Confiance situationnelle. De même, il existe une association significative entre l'acceptation et la perception du risque. Les participants acceptant les suggestions de maintenance prédictive ont rapporté une perception du risque plus faible (W=13911, p < .001).

R3. Effet de la transparence sur la confiance dans le signal et la confiance situationnelle

Comme le montre la figure 7, lorsque l'IA Seanatic n'est transparente que sur la "fiabilité" (condition "Fiab"), on observe une augmentation de la confiance dans le signal par rapport aux situations où l'IA communique les risques (respectivement pour "Fiab&Risque" : OR = 0,39, p < 0,05 ; et pour "Risque" : OR = 0,26, p = 0,001).

De plus, pour la perception du risque, lorsque l'IA était transparente uniquement sur la fiabilité (condition "Fiab"), on observe une diminution de la perception du risque par rapport aux situations où l'IA était transparente sur les risques (respectivement pour "Fiab&Risque" : OR = 4,73, p < 0,001 ; et pour "Risque" : OR = 6,02, p < 0,001).

Enfin, la criticité des opérations de maintenance n'a pas influencé la confiance dans le signal. Au contraire, nous avons observé que lorsque la criticité de la proposition est "Modérée", la perception du risque est plus élevée par rapport à une criticité "Elevée" (OR = 3.21, p < .005).

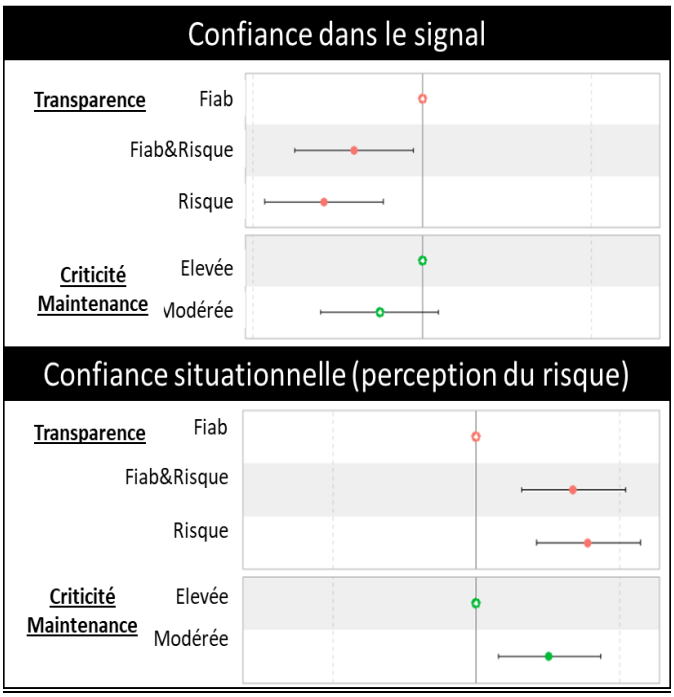

Figure 7 : Effets de la transparence sur la confiance

5 Discussion

Ces deux études de cas ont permis d'explorer les relations entre la transparence des agents, la confiance et l'acceptation, au sein d'équipes humain-système impliquées dans des activités représentatives de l'industrie 4.0. Nous avons ainsi pu étudier ces relations à la fois dans des activités opérationnelles (coopération dans l'action) et des activités tactiques (coopération dans la planification), ainsi qu'avec des systèmes intelligents plus ou moins fiables (très myope et inexact pour le projet Humanism, très fiable pour Seanatic).

Ces différentes expérimentations ont montré une certaine convergence et une certaine reproductibilité des résultats, notamment en ce qui concerne la question R1. Dans les deux cas, la confiance dispositionnelle n'affecte pas significativement le taux d'acceptation des propositions de l'IA par l'opérateur humain. Au contraire, les deux autres niveaux de confiance sont fortement corrélés avec l'acceptation : la conformité augmente avec une confiance plus élevée dans le

signal et une perception du risque plus faible (cette perception du risque étant associée à la confiance situationnelle).

De plus, les résultats liés à la question R2 et présentés dans l'étude de cas Humanism sont en accord avec les hypothèses de Chancey et al. (cf. Figure 4), suggérant que la perception du risque peut être considérée comme un modérateur de la confiance humaine dans la fiabilité de l'agent autonome, et que cet effet modérateur a un impact sur l'acceptation humaine des messages émis par un robot.

Enfin, en considérant la question R3, nous avons constaté que le fait de jouer sur les niveaux et les dimensions de la transparence de l'agent peut modifier de manière significative la confiance dans le signal et la confiance situationnelle, et donc la manière de suivre ou non les propositions des systèmes intelligents (comme souligné dans la Figure 7). Ce résultat corrobore les différents travaux de recherche mentionnés au §1.2, et ouvre de nouvelles perspectives, en phase de conception ou d'exploitation, pour corriger ou mieux calibrer la confiance en l'autonomie, afin d'éviter la méfiance, la défiance et l'effet de complaisance.

6 CONCLUSION

Ce travail de recherche a permis d'étudier la question de la confiance dans les équipes humaines-systèmes autonomes dans des situations typiques de l'industrie 4.0. Cet article viser à articuler les différents travaux conceptuels sur la confiance dans l'autonomie, en reliant la performance comportementale (acceptation des conseils de l'IA) avec les différents types de confiance (confiance dispositionnelle, confiance situationnelle et confiance dans le signal). En outre, sur la base de deux études de cas, il donne un aperçu de la façon dont ces types de confiance interagissent entre elles, et comment elles peuvent jouer sur l'acceptation. Enfin, il démontre comment la confiance dans l'autonomie peut être modifiée et manipulée en variant la transparence de l'agent, ouvrant ainsi des perspectives pour la conception et le contrôle opérationnel des équipes humain-système.

7 REMERCIEMENTS

Les auteurs tiennent à remercier l'ANR et l'ADEME pour leur soutien et le financement respectif des projets Humanism et Seanatic, qui ont permis la réalisation des études présentées dans ce document.

8 BIBLIOGRAPHIE

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes [Software]. R package version 1.1-7.

Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent transparency: A review of current theory and evidence. *IEEE Transactions on Human-Machine Systems*, *50*(3), 215-224.

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.

Chen, J. Y., Procci, K., Boyce, M., Wright, J. L., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency*. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MDUNIVERSITY OF CENTRAL FLORIDA ORLANDO.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*, 85-104.

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.

Guerin, C., Rauffet, P., Chauvin, C., & Martin, E. (2019). Toward production operator 4.0: modelling human-machine cooperation in industry 4.0 with cognitive work analysis. *IFAC-PapersOnLine*, *52*(19), 73-78.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies: 11th International Conference, VAMR 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21* (pp. 476-489). Springer International Publishing.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

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. *Computer Supported Cooperative Work (CSCW)*, *5*, 299-321.

Lyons, J. B. (2013, March). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

Lyons, J. B., Koltai, K. S., Ho, N. T., Johnson, W. B., Smith, D. E., & Shively, R. J. (2016). Engineering trust in complex automated systems. *Ergonomics in Design*, *24*(1), 13-17.

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.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*, 709–734. doi:10.2307/258792

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), 262-273

Pacaux-Lemoine, M. P., Trentesaux, D., Rey, G. Z., & Millot, P. (2017). Designing intelligent manufacturing systems through Human-Machine Cooperation principles: A human-centered approach. *Computers & Industrial Engineering*, *111*, 581-595.

Rajabiyazdi, F., & Jamieson, G. A. (2020, October). A review of transparency (seeing-into) models. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 302-308). IEEE.

Rauffet, P., Guerin, C., Chauvin, C., & Martin, E. (2018, October). Contribution of Industry 4.0 to the emergence of a joint cognitive and physical production system. In *HFES European Chapter*.

Rauffet, P. (2022). Tools and methods for Human-Autonomy Teaming: Contributions to cognitive state monitoring and system adaptation. *Habilitation à diriger des recherches. Université Bretagne Sud*

Simon, L., Guérin, C., Rauffet, P., & Lassalle, J. (2021,

September). Using cognitive work analysis to develop predictive maintenance tool for vessels. In *31st European Safety and Reliability Conference*.

Simon, L., Rauffet, P., Guérin, C., & Seguin, C. (2022). Trust in an autonomous agent for predictive maintenance: how agent transparency could impact compliance. *Industrial Cognitive Ergonomics and Engineering Psychology*, *35*

Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, *39*(5), 740-748.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016, March). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 109-116). IEEE.

Wilson, R. S., Zwickle, A., & Walpole, H. (2019). Developing a broadly applicable measure of risk perception. *Risk Analysis*, *39*(4), 777-791.

Wright, J. L., Chen, J. Y., Barnes, M. J., & Hancock, P. A. (2017, September). The effect of agent reasoning transparency on complacent behavior: An analysis of eye movements and response performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1594-1598). Sage CA: Los Angeles, CA: SAGE Publications.

Zhang, D. C., Highhouse, S., & Nye, C. D. (2019). Development and validation of the general risk propensity scale (GRiPS). *Journal of Behavioral Decision Making*, *32*(2), 152-167.