Utilisation dynamique de la transparence pour la coopération humain-machine

Loïck Simon, Clément Guerin, Philippe Rauffet

▶ **To cite this version:**

Loïck Simon, Clément Guerin, Philippe Rauffet. Utilisation dynamique de la transparence pour la coopération humain-machine. 12e Colloque EPIQUE Paris, ARPEGE, Jul 2023, Paris, France. hal-04111732v1

HAL Id: hal-04111732

https://hal.science/hal-04111732v1

Submitted on 31 May 2023 (v1), last revised 7 Jun 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉPIQUE 2023
Utilisation dynamique de la transparence pour la coopération humain-machine

Loïck SIMON, Clément GUERIN, Philippe RAUFFET

Université Bretagne Sud, Lab-Sticc, UMR 6285, loick.simon@univ-ubs.fr

Université Bretagne Sud, Lab-Sticc, UMR 6285, clement.guerin@univ-ubs.fr

Université Bretagne Sud, Lab-Sticc, UMR 6285, philippe.rauffet@univ-ubs.fr

Catégorie de soumission : communication longue

RÉSUMÉ

Les intelligences artificielles, fortes de l'utilisation de big data, sont désormais des outils d'aide à la décision. Ces machines peuvent proposer des solutions à un opérateur qui aura la charge de les accepter ou de les refuser. Cette coopération nécessite que l'opérateur ait une confiance calibrée dans la machine. Une confiance calibrée permet d'éviter que l'opérateur accepte ou refuse de façon inconsidérée les propositions d'une machine. Afin de calibrer la confiance, le concept de transparence semble un levier prometteur. Dans le cadre de la maintenance prévisionnelle en milieu maritime, nous testons ici l'utilisation de la transparence d'une intelligence artificielle via une interface adaptative. Les résultats indiquent que la transparence peut être un levier permettant d'adapter la communication de l'intelligence artificielle et permettant à l'opérateur de calibrer sa confiance. Nos résultats montrent un lien entre les changements de transparence et les changements de comportement de l'opérateur. L'utilisation de cette dynamique est une voie prometteuse pour l'amélioration de la coopération humain-machine.

MOTS-CLÉS

Transparence, Interface adaptative, Confiance, Coopération Humain-Machine, Prise de décision

1 INTRODUCTION

Les intelligences artificielles (IA) sont désormais assez développées pour aider les opérateurs dans leurs prises de décision et ce dans de multiples domaines tels que l'armée, la santé ou la finance (Chong et al., 2022). Lorsqu'une IA fait des propositions, la coopération qui s'opère avec l'humain est de type débative (Schmidt, 1991 ; Hoc, 1996). L'opérateur n'ayant pas les compétences pour manipuler de grand ensemble de données (i.e. big data), l'IA pourra lui proposer de reconsidérer certaines décisions en prenant en compte les données dont elle dispose. Si nous prenons l'exemple de la maintenance prévisionnelle en milieu maritime, l'IA va pouvoir utiliser les données d'un navire issues de capteurs, afin de simuler les états futurs d'un équipement. De cette projection, l'IA peut reconsidérer une date d'intervention de maintenance prévue et proposer une alternative à l'opérateur (Seguin et al., 2022).

Lorsque qu'une IA émet une proposition, l'opérateur doit prendre une décision : accepter ou refuser. Si l'opérateur accepte, cette décision reflète un comportement de confiance qui est issu de l'attitude de confiance de l'opérateur envers cet agent autonome (AA) (Hoff & Bashir, 2015). Selon Lee et See (2004), la confiance est définie comme l'attitude d'un individu envers un agent, qu'il considère comme pouvant l'aider à atteindre un but. La confiance nécessite d'être calibrée (de Visser et al., 2020) car un niveau de confiance trop élevé ou trop faible conduit à un risque de mauvaise utilisation (voir les célèbres travaux de Parasuraman & Riley (1997) : Use, Misuse, Disuse, Abuse). Zhang, Liao et

Bellamy (2020) ont montré que le concept de transparence est un levier prometteur pour calibrer la confiance, en permettant la construction d'un modèle mental de la situation et de l'AA.

Pour Chen (2021), la transparence peut être définie comme le taux d'information que transmet un AA à un opérateur humain. Il existe deux modèles principaux dans la littérature sur la transparence. Le modèle de Chen et al. (2018) distingue trois niveaux de transparence, en référence aux travaux d'Endsley (1995), pour aider à la construction et au maintien des trois niveaux de la conscience de la situation. Dans ce modèle, les informations que communique l'AA vont améliorer la conscience de la situation de l'opérateur. Cette amélioration permet de calibrer la confiance de l'opérateur dans l'AA et d'augmenter la performance du couple agent autonome/opérateur. Le second modèle est celui de Lyons (2013) qui distingue différentes dimensions pouvant être communiquées par l'AA, ce qui facilite la coopération avec un opérateur. Par exemple, un AA qui communique sur la dimension « analytique » fournira des informations sur les éléments qu'il a pris en compte, ses algorithmes et sa la fiabilité de ses algorithmes. Si l'agent communique sur la dimension « environnement », il fournira des informations relatives à son environnement (par exemple les risques associés à une situation).

Il est possible de considérer les interfaces adaptatives comme une autre solution pour calibrer la confiance (Naiseh et al., 2021). Ici, l'interface se modifie à partir d'un déclencheur (Sarter, 2007). Ce déclencheur peut être le niveau d'attention ou l'expérience de l'opérateur, ou bien encore une condition particulière de l'environnement. Akash et al. (2020) ou Okamura et Yamada (2020) ont quant à eux utilisé la confiance de l'opérateur comme déclencheur de l'adaptation des interfaces. Dans ces deux expérimentations, l'adaptation des interfaces s'opérait en termes de transparence des informations sur l'interface. Dans l'étude de Okamura et Yamada (ibid.), si la confiance du participant était trop basse, alors l'IA augmentait sa transparence en émettant une information sur sa fiabilité. Leurs premiers résultats sont concluants et mettent en évidence que l'utilisation adaptative de la transparence peut servir à calibrer la confiance grâce à une amélioration du modèle mental de l'IA et de la conscience de la situation.

Enfin, l'étude de Hoesterey et Onnasch (2022) a montré que dans un contexte de coopération entre un opérateur et une aide à la décision, le risque associé à une situation a un effet sur les comportements de confiance. Dans la perspective des travaux de Akash et al. (ibid.) et d'Okamura et Yamada (ibid.), l'expérimentation présentée dans cette communication cherche à faire varier la transparence d'un AA, sur sa dimension analytique, en fonction des comportements de confiance d'un opérateur. C'est en ce sens que notre étude propose d'utiliser la transparence de façon dynamique. Notre question de recherche est la suivante : Un agent autonome à transparence adaptative aura-t-il une influence sur le comportement de confiance d'un opérateur ?

2 PROTOCOLE EXPERIMENTAL

L'expérimentation s'est déroulée à l'ENSM (école nationale supérieure maritime) du Havre. Les 16 participants (*M* = 25,31 ans, *E-T* = 9,85) avaient tous une expérience de navigation (*M* = 37,31 mois ; *E-T* = 118,09). 15 participants étaient des élèves de l'ENSM et 1 participant était un ancien chef mécanicien. Le recrutement a été fait en interne et aucune rétribution n'était prévue. Les participants étaient conviés à une expérimentation de coopération avec une IA de maintenance prévisionnelle.

La tâche des participants était de coopérer avec une IA de maintenance prévisionnelle sur 60 situations, dépendantes les unes des autres, en acceptant ou en refusant ses propositions. À la suite de cette prise de décision les participants recevaient un feedback (Tab. 1) et avaient accès à une nouvelle proposition avec la transparence actualisée en fonction d'un taux d'acceptation (Fig. 1).

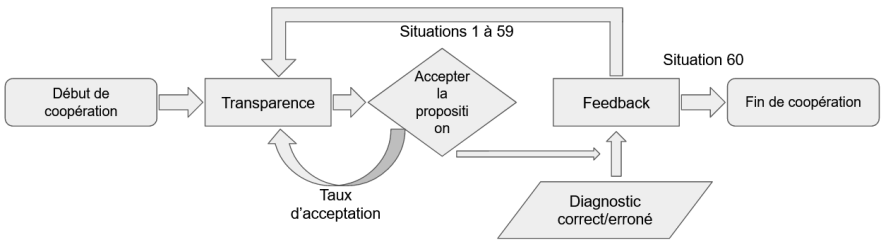

Figure 1 : Protocole expérimental

La transparence de l'intelligence artificielle (IA) variait en fonction du taux d'acceptation du participant calculé sur les 7 dernières propositions. L'IA était transparente soit exclusivement sur la fiabilité de sa proposition (i.e. 90%, « F », Fig. 2 en jaune), soit sur les risques liés à cette proposition (en cas d'acceptation et de refus, « R », Fig. 2 en vert), soit sur les deux informations (« F+R »). Sur les dix premières situations l'IA communiquait la fiabilité et le risque (« F+R »). Ce feedback textuel pouvait être **positif** ou **négatif** en fonction de la décision du participant (i.e. accepter ou refuser la proposition) et de la proposition (i.e. proposition sur un diagnostic correct ou erroné) (Fig. 1et Tab. 1). Il portait sur les conséquences de son choix sur l'équipement. Le rôle de ce feedback était de rendre la situation de coopération écologique et de créer une boucle de rétroaction entre la décision et la construction de la confiance de l'opérateur envers l'agent autonome.

Tableau 1 : Feedback de l'IA en fonction du diagnostic lié à la proposition de l'IA et du choix du participant

Proposition de l'IA	**Choix du participant**	**Feedback de l'IA (conséquence)**
Basée sur un diagnostic correct (55 propositions)	Accepter	Feedback positif « *Le filtre a tenu sans conséquences négatives jusqu'à la date proposée par Seanatic* »
	Refuser	Feedback négatif « *Après inspection suite à son changement, le filtre aurait pu tenir sans conséquences négatives jusqu'à la date proposée par Seanatic* »
Basée sur un diagnostic erroné (5 propositions)	Accepter	Feedback négatif : « *Le filtre était encrassé lors de votre maintenance* »
	Refuser	Feedback positif « *Le filtre présentait les traces d'un début d'encrassement lors de votre maintenance* »

Toutes les propositions de l'IA portaient sur le fait de décaler la maintenance d'un équipement important, dans le sens où un défaut de maintenance pouvait conduire à des conséquences graves. Les propositions de l'IA, fiables à 90%, étaient toutes associées à une criticité forte : si le participant acceptait, il y avait une forte probabilité de survenue d'un évènement grave. Par exemple, reculer le changement de filtre peut conduire à son encrassement prématuré, ce qui peut générer une panne moteur. L'IA n'étant fiable qu'à 90%, 5 de ses propositions étaient basées sur un diagnostic erroné et amenaient des conséquences négatives (i.e. le filtre s'encrassait plus tôt que prévu). Ces 5 propositions ne pouvaient pas être différenciées des autres propositions et survenaient de façon prédéfinie et identique à tous les participants. Entre chaque situation (i.e. une proposition de l'IA suivie d'une décision du participant) une ellipse temporelle était simulée, afin d'intégrer un aspect écologique au protocole. Cette ellipse était de 14 jours si le participant refusait (la maintenance était alors réalisée à la date prévue), et de 18 jours s'il acceptait (la maintenance se faisait à la date proposée par l'IA).

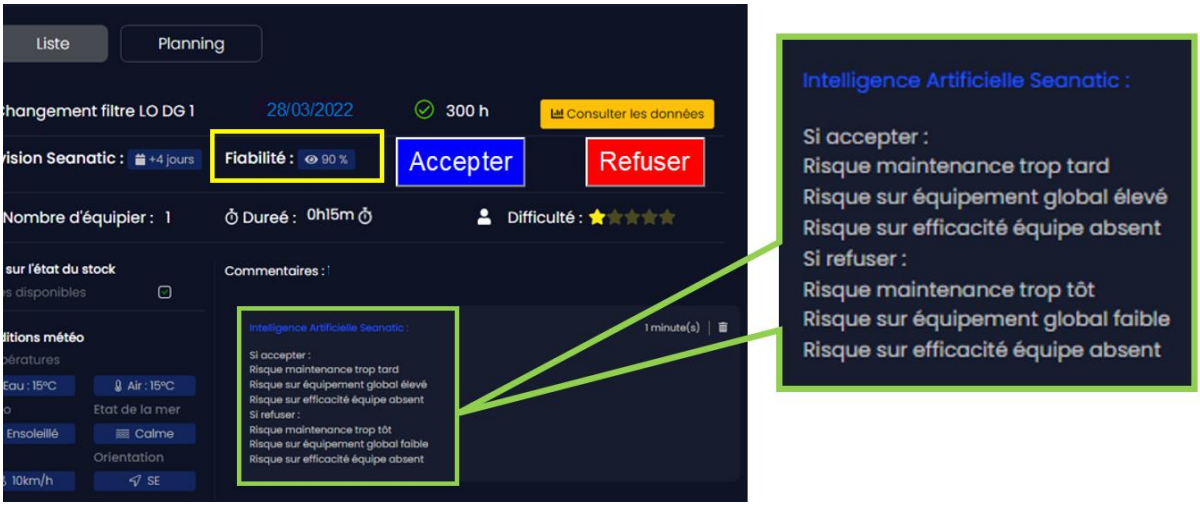

Figure 2 : Interface de dialogue avec l'IA Seanatic

Pour accepter ou refuser la proposition de l'IA, le participant utilisait le logiciel Seanatic sur lequel étaient aussi communiquées les informations liées à chaque proposition (Fig. 2).

Dans le cadre de cette expérimentation nous avons investigué la variable **comportement de confiance**, mesurée après chaque décision du participant (accepter ou refuser) suite aux propositions de l'IA Seanatic. Notre hypothèse est la suivante : **les changements de transparence font varier les comportements de confiance** (Akash et al., 2020 ; Okamura & Yamada, 2020). Les changements de transparence (par ex., de "F+R" à "R"), de feedback (par ex., d'un feedback positif à un feedback négatif) et d'acceptation (par ex., d'accepter à refuser) sont nommés "Delta". L'absence de delta est nommée "none". Nous avons également contrôlé l'affinité envers la technologie, la propension à faire confiance à l'agent autonome, la propension au risque et la confiance de l'opérateur dans l'IA. Nous avons utilisé des tests de Chi² suivi d'une régression logistique multinomiale pour analyser les effets de la transparence et des feedbacks sur l'acceptation, ainsi que les effets des deltas.

3 RÉSULTATS

Premièrement nous avons identifié deux groupes de participants selon leur comportement. Le premier groupe (nommé « alternant »), correspondant à 7 participants, avait un comportement d'alternance de réponse après les propositions. Le second groupe (nommé « conservateur »), correspondant à 9 participants, conservait la même décision durant les 60 situations. Ce groupe avait tendance à toujours accepter les propositions. L'utilisation de T-tests de Student n'a montré aucune différence significative entre les groupes « alternant » et « conservateur » pour :

a. L'affinité envers la technologie (« alternant » (M = 4.41, $E\text{-}T$ = 0.98) et « conservateur » (M = 4.39, $E\text{-}T$ = 0.74), ($t(14) = 0.041$, p = .484));
b. La propension à la confiance (« alternant » (M = 3.62, $E\text{-}T$ = 0.82) et « conservateur » (M = 3.74, $E\text{-}T$ = 0.30), $t(14) = 0.414$, p = .452));
c. La propension au risque (« alternant » (M = 3.32, $E\text{-}T$ = 1.10) et « conservateur (M = 3.83, $E\text{-}T$ = 1.53), $t(14) = 0.748$, p = .467));
d. La confiance dans l'IA (« alternant » (M = 3.71, $E\text{-}T$ = 1.08) et « conservateur (M = 3.82, $E\text{-}T$ = 0.64), $t(14) = 0.258$, p = .8)).

Nous concentrerons donc nos analyses statistiques sur le groupe « alternant ». Les deux groupes ayant une forte tendance à accepter, la modalité de transparence « F » n'a été activée qu'une fois sur toutes les situations (420 situations). Nous l'excluons également des analyses.

Nous constatons un effet principal de la transparence. Le taux d'acceptation est plus bas lorsque l'IA communique uniquement sur les risques « R » (72%) que lorsque l'IA communique sur la fiabilité et les risques « F+R » (93%) (X^2 (1) = 32.279, p <.001) (Fig. 3). L'utilisation de la régression logistique binomiale montre que lorsque l'IA utilise la transparence « R » la probabilité d'acceptation est significativement plus faible qu'avec une transparence « F+R » (OR = 0,26, p <.005) (Tab. 2).

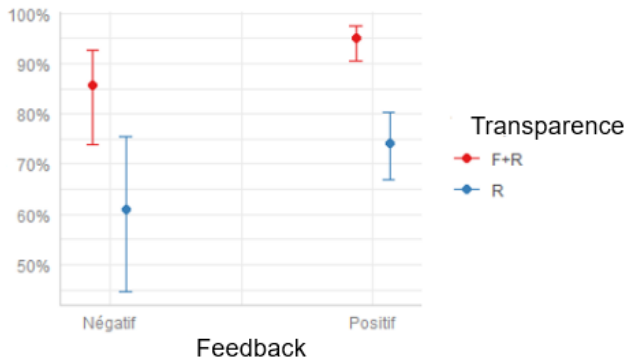

Figure 3 : Taux d'acceptation selon la transparence et le feedback N-1

Nous ne constatons pas de différence significative entre le taux d'acceptation lorsque le feedback est « Négatif » (76%) et lorsque le feedback est « Positif » (84%) (X^2 (1) = 2.95, p > .05) (Fig. 3). Le modèle issu de la régression logistique prévoit une probabilité d'acceptation plus élevée lorsque le

feedback est positif par rapport à un feedback négatif (*OR* = 3,21, *p* <.05) (Tab. 2). Ces premiers résultats supportent notre hypothèse.

Tableau 2 : Effet de la transparence et du feedback sur l'acceptation (Régression logistique binomiale, référence « F+R » et « Négatif »)

Modalités	OR[1]	95%IC[2]	*p-value*
Transparence (réf. F+R)			
R	0.14	-2.75, -1.24	< .01
Feedback (réf. Négatif)			
Positif	3.21	1.12, 9.17	< .05
*Transparence * Feedback (réf. F+R * Négatif)*			
R * Positif	0.57	0.16, 2.07	0.4

[1]OR = Odd Ratio, [2]IC = intervalle de confiance

Pour regarder l'effet du delta de la transparence et du delta de feedback sur le delta de comportement, nous utilisons une régression logistique multinomiale par étape afin de conserver le meilleur modèle. Celui-ci inclut deux effets principaux liés aux deltas de transparence et aux deltas de feedback, mais pas d'interaction entre ces deltas.

- Nous constatons un effet principal du delta de transparence. L'utilisation de la régression logistique multinomiale nous permet de constater que lorsque l'IA ne communique plus sa fiabilité « -F » (i.e. elle passe de « F+R » à « R »), il y a plus de probabilité de passer d'une acceptation à un refus « A=>R » par rapport aux valeurs de référence (pas de changement de transparence et d'acceptation : « None ») (*OR* = 9,1, *p <.001*). A l'inverse, lorsque l'IA communique à nouveau sur sa fiabilité « +F » (i.e. elle passe de « R » à « F+R »), il y a plus de probabilité de passer d'un refus à une acceptation « R=>A » par rapport aux valeurs de référence (*OR* = 754, *p < .001)* (Tab 3.).
- Nous constatons également un effet principal du delta de feedback. Lorsque le feedback passe de positif à négatif « P=>N », les probabilités d'un delta d'acceptation « A=>R » et d'un delta « R=>A » augmentent par rapport aux valeurs de référence (delta transparence « None » et delta acceptation « None ») (respectivement : *OR* = 2.61, *p* < .005 et *OR* = 25.9, *p* < .001) (Tab 3).

Tableau 3 : Effet du delta transparence et du delta feedback sur le delta d'acceptation (Régression logistique multinomiale)

Modalités	OR[1]	95%IC[2]	*p-value*
A=>R			
Delta Transparence (réf. None)			
-F	9.09	4.66, 17.7	<.001
Delta Feedback (réf. None)			
N=>P	0.41	0.14, 1.07	0.067
P=>N	2.61	1.35, 5.04	0.004
R=>A			
Delta Transparence (réf. None)			
-F	2.17	0.43, 10.5	0.4
+F	754	92.7, 6.138	<.001
Delta Feedback (réf. None)			
P=>N	25.9	12.7, 53.0	<.001

[1]OR = Odd Ratio, [2]IC = intervalle de confiance

Ces résultats, qui montrent que des changements d'acceptation sont corrélés à des changements de transparence et de feedback, supportent notre hypothèse.

4 DISCUSSION

Cette expérimentation avait pour but de tester l'utilisation adaptative de la transparence lors d'une coopération humain-machine. Premièrement, nous constatons un effet principal de la

transparence sur le taux d'acceptation de l'opérateur. Un agent autonome (AA) qui est plus transparent conduira l'opérateur à plus accepter ses propositions. Secondement, nous constatons que les changements de comportements de l'opérateur (par ex., passer d'accepter à refuser) sont corrélés aux changements de transparence de l'AA. Cependant, le changement de comportement induit par ces changements de transparence de l'AA semble diverger selon le profil des utilisateurs.

Comme dans l'expérimentation de Simon et al. (sous presse), nous constatons un effet de la transparence sur le risque qui est corrélé à un taux d'acceptation plus faible. Les participants du groupe « alternant » ont utilisé les informations relatives à la dimension « environnement » afin de prendre leur décision. Cette transparence sur la dimension de l'environnement permettait au participant de comprendre que l'IA faisait une proposition à criticité élevée. Les participants ont donc eu tendance à refuser la proposition. L'IA n'étant plus transparente sur la dimension analytique, la fiabilité ne pouvait pas compenser les risques liés à la proposition. Conformément à l'étude de Wright, Chen et Lakhmani (2020), lorsque l'AA est transparente et fiable, la confiance du participant augmente. Cette augmentation de la confiance se traduit ici par un comportement de confiance lorsque l'AA est transparent sur la dimension « analytique » (i.e. accepter la proposition). Nous voyons également un effet du feedback positif. Ici les participants continuaient de faire confiance à l'AA en acceptant ses propositions, si le précédent feedback qu'il communiquait avait un résultat positif (c'est-à-dire, que le choix de l'opérateur n'avait pas eu de conséquence sur l'équipement). En d'autres termes, si l'on analyse la dynamique des interactions, nos résultats montrent que lorsque la dernière interaction avec un AA s'est bien passée, les opérateurs ont tendance à maintenir leur confiance en acceptant les propositions de l'IA. Ces résultats sont congruents avec le modèle de Hoff et Bashir (2015) selon lequel les expériences passées avec l'AA contribuent à la construction et au maintien de la confiance de l'opérateur. Cette attitude, si elle est positive, prend alors la forme d'un comportement d'acceptation.

Les résultats liés aux changements de comportement de l'opérateur nous permettent de constater une corrélation avec les situations où l'AA modifie sa transparence. Lorsque l'AA ne communique plus sur sa fiabilité, les opérateurs « alternant » modifient leur comportement et se mettent à refuser. A l'inverse, lorsque l'AA communique à nouveau sur sa fiabilité, les participants modifient à nouveau leur comportement et se remettent à accepter. Conformément aux études précédentes (Akash et al., 2020 ; Okamura & Yamada, 2020) un opérateur se trouvant dans une situation de coopération avec un AA en capacité d'adapter sa transparence modifie son comportement en fonction de ces changements. Pour les participants du groupe « alternant », les changements de transparence étaient utilisés pour calibrer leur comportement et l'accorder en fonction des informations à leur disposition. En fonction de ce que l'on va juger comme une confiance appropriée (ici le taux d'acceptation), l'AA doit avoir la capacité de rendre visible ou non certaines dimensions de la transparence, telles que les dimensions proposées par Lyons (2013). Lorsque l'AA n'est plus transparent sur la dimension analytique (et en particulier sur sa fiabilité) cela accentue l'importance de la dimension de l'environnement dans la décision de l'opérateur. A l'inverse, lorsque l'AA communique sur ces deux dimensions, l'opérateur les prend en compte dans sa décision. Interagir avec un AA pouvant modifier sa transparence dynamiquement est une voie intéressante à explorer pour améliorer la coopération homme-machine. Pour cela, l'AA doit pouvoir mettre l'accent sur une dimension ou sur une autre s'il perçoit que la confiance de l'opérateur n'est pas appropriée. Concernant l'effet des feedbacks, le modèle issu de la régression logistique multinomiale prévoit que le changement de feedback, positif vers négatif, est prédicteur de deux comportements différents. Ce changement de feedback augmente la probabilité que l'opérateur modifie sa décision, c'est-à-dire soit le passage d'un refus vers une acceptation ou inversement, d'une acceptation vers un refus. Un changement de feedback semble donc générer un questionnement chez l'opérateur, et l'amène à remettre en cause son choix précédent. Notons toutefois que la probabilité qu'un changement de feedback conduise l'opérateur à passer d'un refus vers une acceptation est plus élevée que celle qui conduit à passer d'une acceptation vers un refus. Nous interprétons le cas où l'opérateur refuse une proposition, alors que les acceptations précédentes conduisaient à des feedbacks positifs, comme la volonté, pour l'opérateur, de tester la fiabilité de l'AA. Son refus amenant à un feedback négatif, l'opérateur est conforté dans l'idée qu'accepter est corrélé à un feedback positif et se remet à

accepter. A l'inverse, lorsque les participants acceptent et que le feedback est négatif, ils perdent confiance dans l'AA. De ce fait, ils considèrent comme probable que l'AA refasse une erreur. L'événement précédent ayant amené des conséquences négatives, il est probable que l'AA se trompe à nouveau et je ne devrais plus lui faire confiance. Contrairement à Salem et al. (2015) nous constatons un effet de l'erreur de l'AA sur le comportement de l'opérateur. Comme souligné par les auteurs, le degré de révocabilité de l'erreur a un effet sur l'attitude. Or le contexte maritime a pour objectif d'éviter toute erreur de maintenance (Simon et al., 2021). Ici l'importance de cette révocabilité est telle qu'elle dépasse l'attitude et vient modifier le comportement, malgré la fiabilité élevée de l'AA.

Comme dans l'expérimentation de Kortschot et al. (2022), nous constatons que les participants ne sont pas tous sensibles à la transparence d'un AA puisque nous avons identifiés 2 profils. Nos résultats ne permettent pas de montrer avec certitude quels déterminants individuels pourraient expliquer cette différence, puisqu'aucun résultat significatif n'a été observé sur les 4 déterminants que nous avons contrôlés (affinité envers la technologie, propension à la confiance et au risque, confiance dans l'IA). En revanche, nous constatons que ces déterminants convergent. En effet, comparé au groupe « conservateur », le groupe « alternant » connaissait mieux la technologie et ses limites (i.e. une affinité envers la technologie plus élevée), avait une propension plus faible à faire confiance dans les AA, ainsi qu'une propension plus faible au risque, et également une confiance plus faible dans l'IA.

Nous relevons quelques limites à cette étude. Premièrement, l'algorithme d'adaptation n'inclut que les 7 dernières interactions. Se pose donc la question de ce qu'il se passerait sur une plus longue durée (et avec un AA qui adapterait sa transparence moins régulièrement), ou au contraire sur des échéances plus courtes (et avec un AA adaptant sa transparence plus régulièrement). Également, la fiabilité de l'AA était fixée à 90%. Il serait intéressant d'observer les effets que produit un changement de transparence avec différents niveaux de fiabilité. Enfin, il est nécessaire d'améliorer la compréhension des facteurs et déterminants individuels qui sous-tendent la sensibilité à ce type d'AA.

5 CONCLUSION

Dans cette étude, nous voulions tester la pertinence de l'utilisation d'une transparence adaptative en s'appuyant sur les dimensions analytiques et de l'environnement du modèle de Lyons (2013). Pour ce faire nous avons mis en place un protocole simulant une interaction entre un AA et un opérateur avec des situations dépendantes entre elles. Nous avons pu confirmer que les changements de transparence de l'AA étaient corrélés avec les changements de comportement des opérateurs. Ainsi, les résultats de cette étude montrent l'intérêt d'utiliser les dimensions de la transparence pour calibrer la confiance d'un opérateur envers un agent autonome. Il est également possible d'utiliser le taux d'acceptation comme déclencheur des mécanismes d'adaptation des interfaces. Ces deux résultats pourraient permettre, dans un contexte industriel, d'éviter que les opérateurs se trouvent dans une situation de complaisance en les faisant se questionner sur leur choix d'accepter ou de refuser.

Les AA adaptatifs utilisant la transparence comme levier d'adaptation semblent être une voie prometteuse pour la calibration de la confiance. Dans le cadre où l'AA est force de proposition, comme dans de nombreux secteurs tels que la santé, l'industrie ou l'armée (Chong et al., 2021), cette propriété permettrait de calibrer et de maintenir un niveau de confiance adapté. L'objectif étant de ne pas utiliser la proposition de façon déconcertée, mais de toujours rester vigilant afin d'éviter de surutiliser, ou de sous-utiliser l'AA (Parasuraman & Riley, 1997) et éviter biais et phénomène de contentement (Parasuraman & Manzey, 2010). Quatre voies de recherches restent à investiguer sur l'utilisation d'une interface adaptative. Premièrement, il serait intéressant de changer les algorithmes basés sur le taux d'acceptation pour regarder s'il existe une différence entre un algorithme court-terme (par exemple la dernière décision) ou long-terme (les cinquante dernières). Deuxièmement, il est nécessaire de mieux comprendre les déterminants individuels influencent l'utilisation des interfaces adaptatives. Troisièmement, la question de la fiabilité reste à approfondir. Il serait opportun de moduler cette fiabilité pour constater quels sont les effets sur l'utilisation d'une interface adaptative. Quatrièmement, il semble nécessaire d'utiliser d'autres dimensions de la transparence pour observer leur potentiel en tant que levier d'adaptation, et ce par rapport à d'autres déclencheurs.

6 REMERCIEMENTS

Les recherches présentées dans ce document ont été menées dans le cadre du projet SEANATIC (N°2082C0023). Ce projet est soutenu par le Programme d'Investissements d'Avenir de l'ADEME.

7 BIBLIOGRAPHIE

Akash, K., McMahon, G., Reid, T., & Jain, N. (2020). Human Trust-based Feedback Control: Dynamically varying automation transparency to optimize human-machine interactions.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science, 19(3), 259-282.

Chen, J. Y. C. (2021). Agent Transparency. In Smart and Intelligent Systems. CRC Press.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018.

de Visser, E., Peeters, M. M. M., Jung, M., Kohn, S., Shaw, T., Pak, R., & Neerincx, M. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. International Journal of Social Robotics, 12.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 37(1), 32-64. https://doi.org/10.1518/001872095779049543

Hoc, J.-M. (1996). Supervision et controle de processus : La cognition en situation dynamique. PUG.

Hoesterey, S., & Onnasch, L. (2022). The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. *Cognition, Technology & Work*.

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

Kortschot, S. W., Jamieson, G. A., & Prasad, A. (2022). Detecting and Responding to Information Overload With an Adaptive User Interface. Human Factors, 64(4), 675-693.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), 50-80.

Lyons, J. B. (2013, mars 15). Being Transparent about Transparency: A Model for Human-Robot Interaction. 2013 AAAI Spring Symposium Series. 2013 AAAI Spring Symposium Series.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendation: When design meets trust calibration. World Wide Web, 24. https://doi.org/10.1007/s11280-021-00916-0

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. PLOS ONE, 15(2).

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230-253.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52(3), 381–410.

Pynadath, D. V., Gurney, N., & Wang, N. (2022). Explainable Reinforcement Learning in Human-Robot Teams: The Impact of Decision-Tree Explanations on Transparency. 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 749-756.

Sarter, N. (2007). Coping with Complexity Through Adaptive Interface Design. International Conference on Human-Computer Interaction, 493-498.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 141-148.

Schmidt, K., Rasmussen, J., Brehmer, B., & Leplat, J. (1991). Cooperative work: A conceptual framework. Distributed decision making: Cognitive models for cooperative work, 75-110.

Seguin, C., Rioual, Y., Diguet, J.-P., & Gogniat, G. (2022). Data Extraction and Deep Learning Method for Predictive Maintenance in Vessel's Engine Room. 32nd European Safety and Reliability Conference (ESREL 2022)., 1983-1990.

Simon, L., Guérin, C., Rauffet, P., & Lassalle, J. (2021). Using cognitive work analysis to develop predictive maintenance tool for vessels. 31st European Safety and Reliability Conference.

Simon, L., Guérin, C., Rauffet, P., Chauvin, C., & Martin, E. (sous-presse). How humans comply with a (potential) faulty robot: effects of multidimensional transparency. *IEEE Transactions on Human-Machine Systems*

Wright, J. L., Chen, J. Y. C., & Lakhmani, S. G. (2020). Agent Transparency and Reliability in Human–Robot Interaction: The Influence on User Confidence and Perceived Reliability. *IEEE Transactions on Human-Machine Systems*, *50*(3), 254-263.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295-305.