Expérimentation de la confiance d'un utilisateur de système d'IA

Nicolas Maille, Kahina Amokrane-Ferka, Benoit Leblanc, Nicolas Heulot

Expérimentation de la confiance d'un utilisateur de système d'IA

Nicolas. Maille[1], Kahina. Amokrane-ferka[2], Benoit. Leblanc[3], Nicolas. Heulot[2]

[1] ONERA The French Aerospace Lab, Salon de Provence
[2] IRT SystemX, 91120 Palaiseau
[3] IMS UMR CNRS 5218, ENSC Bordeaux INP, Bordeaux,

kahina.amokrane-ferka@irt-systemx.fr

Résumé

Les systèmes à base d'IA deviennent de plus en plus présents dans la vie de tous les jours et surtout dans les activités professionnelles. Se pose alors la question de savoir ce que les humains éprouvent vis-à-vis de ces systèmes. Leurs attitudes s'apparentent déjà aux relations que peuvent avoir entre eux plusieurs humains, dans le cadre professionnel. Avec l'accroissement des services associés à ces systèmes (aide à la décision, recommandation, etc.), la question de la confiance devient particulièrement critique pour la sécurité et la performance des outils industriels.
L'objectif de ce travail est d'éclairer comment la confiance que l'opérateur se construit dans le système à base d'IA vient modifier son ressenti, son comportement et ses performances globales dans la réalisation de sa tâche. Pour ce faire, un micro-monde a été développé et une étude expérimentale faisant intervenir 32 sujets a été conduite afin de mieux cerner la question. Les résultats montrent une adaptation du comportement en fonction du niveau de fiabilité du système, avec en particulier une confiance rapportée plus forte et une supervision moins importante quand la fiabilité de l'IA est forte.

Mots-clés

Confiance, Système à base d'IA, aide à la décision, collaboration Human-IA.

Abstract

AI-based systems are becoming increasingly present in everyday life, and especially in professional activities. This raises the question of how humans feel about these systems. Their attitudes are already similar to the relationships that several humans may have with each other in the workplace. With the increase in services associated with these systems (decision support, recommendations, etc.), the question of trust becomes particularly critical for the safety and performance of industrial tools.
The aim of this work is to clarify how the trust that the operator builds up in the AI-based system modifies his feelings, behavior and overall performance in carrying out his task. To this end, a micro-world was developed and an experimental study involving 32 subjects was conducted to further investigate the issue. The results show an adaptation of behavior according to the system's level of reliability, with in particular higher reported confidence and less supervision when the AI's reliability is high.

Keywords

Trust, AI based System, decision support system, Human Machine Teaming

1 Introduction

1.1 Contexte applicatif

Notre cadre de travail porte sur un système d'analyse visuelle de conformités, déployé dans des usines automobiles. Ce cas d'usage s'inspire des systèmes conçus pour surveiller différentes parties d'une ligne de production de véhicules, venant prendre des photos de plusieurs points de contrôle spécifiques sous différents angles. Ces points de contrôle peuvent concerner les gaines électriques, les soudures, les peintures, et sont validés visuellement par des opérateurs. Sur la base de ces photos, une IA peut venir analyser la conformité des points et pousser le système à déclencher des alarmes visuelles et sonores à chaque fois qu'une anomalie est détectée (cf Figure 1). L'objectif de ce système est de permettre aux opérateurs, en charge du contrôle qualité, de gagner du temps sur la détection de non conformités afin de disposer de plus de temps pour corriger physiquement ces anomalies lorsque cela est possible.
Ce type de système doit idéalement s'intégrer de manière transparente dans le processus de travail des opérateurs.Or, les premiers retours terrains suite au déploiement de cette solution semblent montrer que les opérateurs sont très sensibles à la robustesse du système. Il semblerait que lorsqu'un système réalise trop souvent des fausses détections ou laisse passer des non conformités, les opérateurs s'en détournent progressivement ce qui interroge les mécanismes de la confiance de l'opérateur dans ce système à base d'IA.

1.2 État de l'art

Les progrès accomplis dans le domaine de l'IA au cours de ces dernières années et leur utilisation croissante dans de nombreux outils, tant de la vie courante que dans l'environnement professionnel, montrent que l'IA va de plus en plus impacter notre quotidien ainsi que nos façons de travailler. Néanmoins, il reste encore des efforts pour ces systèmes passent d'une position de support de la collaboration, à un

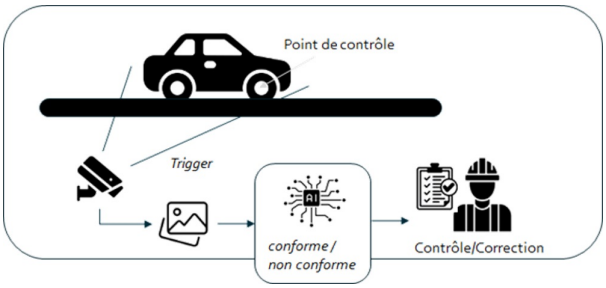

FIGURE 1 – Cas d'usage de conformité visuelle par IA.

système capable d'engagement dans les étapes de la résolution de problème et de prise de décision [12]. La littérature sur le travail d'équipe associant humains et machines identifie la confiance comme un élément essentiel pour que les interactions puissent avoir lieu de manière constructive [2] et ainsi s'inscrire dans une démarche de collaboration.

Cependant, la notion même de confiance, que ce soit dans les individus ou dans les systèmes ne fait pas consensus [13], d'autant que la confiance n'est pas une caractéristique propre au système. Elle s'entend plutôt comme un processus associant un confiant (ou 'trustor') et un mandant (ou 'trustee') [7], couple dans lequel on peut observer l'établissement ou la rupture de ce lien de confiance [11]. Dans la lignée de [3] nous considérons ici que la confiance d'un agent dans un autre représente sa croyance que cet autre agent ne va pas entreprendre des actions qui lui sont préjudiciables. La question de la confiance se pose de manière complexe car cette relation dépend de plusieurs éléments liés au contexte d'interaction [5]. Ces facteurs peuvent être regroupés en trois catégories [3] : ceux liés à l'opérateur, ceux liés à l'automate ou robot et enfin ceux liés à l'environnement. Pour une revue récente de la notion de confiance dans le contexte de la coopération avec des systèmes à base d'IA et des éléments qui l'impactent, le lecteur est invité à se reporter à la méta-analyse de Kaplan et al. [6]. De part sa nature, la confiance n'est pas directement mesurable mais va pouvoir être appréhendée soit à travers le ressenti de la personne faisant confiance (questionnaires), soit par des mesures comportementales car la confiance dans l'autre agent vient aussi modifier leurs interactions [14].

De nombreux facteurs influençant la mise en place de cette confiance ont fait l'objet d'études expérimentales. En particulier, dans le cas où le système d'aide repose sur une IA basée sur de l'apprentissage, l'opacité du système est identifié par la littérature comme nuisant à la confiance qu'il lui accorde l'utilisateur [1]. Ainsi, la manière dont le système explique sa décision peut alors être un facteur déterminant pour l'établissement de ce lien de confiance. Cependant, d'autres facteurs peuvent être tout aussi essentiels. Par exemple l'étude réalisée dans [8], montre l'impact de la performance du système à base d'IA sur la confiance. Les auteurs ont constaté que la confiance dans un système diminue lorsque sa performance est faible. De plus, ils ont souligné que dans les équipes moins performantes, les individus ayant peu confiance dans le système manifestent également peu de confiance envers leurs partenaires humains. Une autre étude [10] s'est focalisée sur l'impact de la nature de l'agent avec lequel un opérateur interagit sur sa prise de décision. L'observation indique que dans une situation où le risque est élevé et où les deux sources d'information (humain et machine) sont équivalentes en termes de fiabilité, le participant préfère choisir une information provenant d'une aide humaine plutôt que celle provenant d'une aide automatisée. La confiance accordée dans l'agent dépendrait donc aussi de sa nature et les résultats acquis pour la confiance entre individus ne peuvent donc pas forcément être étendus à la confiance entre un opérateur et un système à base d'IA. Dans cette même perspective, l'objectif de notre travail est d'éclairer comment la relation de confiance, que l'opérateur se construit avec le système à base d'IA, vient modifier son ressenti, son comportement et finalement ses performances globales dans la réalisation de différentes tâches. Pour cela, le choix est fait de s'appuyer uniquement sur différents niveaux de fiabilité du système d'IA pour modifier cette confiance induite. Dans cet article, nous présentons une expérience de mise en situation d'opérateurs humains dans un micro-monde inspiré du cas d'application de conformité visuelle introduit précédemment.

1.3 Visée de l'étude

Cette étude vise à contribuer à la compréhension de la manière dont un opérateur ressent et s'adapte aux performances d'un système d'assistance à base d'IA. Elle repose sur des sessions de travail répétitives avec ce même système afin que l'opérateur se forge par lui-même une représentation des capacités du système d'aide qui lui est proposé.
Les deux hypothèses formulées sont les suivantes :

- (H1) : l'utilisateur va effectivement percevoir et adapter sa manière de faire en fonction de la fiabilité de l'IA qui lui vient en support (moins l'IA sera fiable, moins l'opérateur lui fera confiance et plus il passera du temps à superviser les résultats proposés par cette IA).
- (H2) : pour une session donnée, le niveau initial de confiance que l'opérateur a dans le système d'aide va impacter sa perception de la fiabilité réelle du système.

Ces hypothèses sont formulées pour répondre aux questions de recherche suivantes :

- Comment la confiance dans le système d'aide modifie-t-elle le comportement de l'utilisateur et ses performances ?
- Comment les performances de l'IA affectent-elles le ressenti de l'utilisateur vis à vis du système ?
- Est-ce que l'on peut conditionner sur la durée, le comportement de l'utilisateur ?

2 Méthodologie

2.1 Participants

Cette étude mobilise un ensemble de 32 participants (15 hommes et 17 femmes) âgés de 18 à 31 ans (M 23.25, ET 3.35), étudiants ou doctorants à l'École Nationale Su-

périeure de Cognitique [1]. Tous ont déclaré avoir une acuité visuelle normale ou corrigée, n'avoir aucun antécédent de troubles neurologiques, et étaient naïfs quant au sujet de l'étude. Ils étaient volontaires pour réaliser l'expérimentation et ont signé un document de consentement éclairé.

2.2 Matériel et stimuli

L'expérimentation s'est déroulée dans un laboratoire de recherche, sur un ordinateur portable (écran 15"). Les sujets étaient installés dans une position proche de la position de travail des opérateurs en atelier (tabouret haut, plan de travail type établis). Chaque passation dure environ 40 minutes. La variable indépendante est la fiabilité de l'IA, et les différentes variables dépendantes mesurées concernent le ressenti de l'opérateur, son comportement et ses performances.
Deux niveaux de fiabilité de l'IA ont été étudiés dans l'expérimentation : (1) une IA très fiable qui ne fait aucune erreur de classification et (2) une IA peu fiable qui fait entre 10% et 30% d'erreur de classification. Chaque sujet travaille avec l'IA pendant 2 blocs consécutifs de 3 essais chacun. Pour un opérateur donné et pour un bloc donné, le niveau de fiabilité de l'IA reste constant ('très fiable' ou 'peu fiable'). Les 32 participants sont répartis sur les quatre combinaisons possibles : 2 (bloc) x 2 (fiabilité de l'IA). Cela constitue donc quatre groupes indépendants de 8 personnes (cf Figure 2).

Groupe	Bloc 1	Bloc 2
G1	Fiable	Fiable
G2	Peu fiable	Peu fiable
G3	Fiable	Peu fiable
G4	Peu fiable	Fiable

FIGURE 2 – Conditions expérimentales associées à chaque groupe.

Les participants ne savent pas à quel groupe ils appartiennent, ils n'ont pas d'information sur la fiabilité de l'IA avec laquelle ils vont travailler et n'ont pas été avertis que cette fiabilité pouvait éventuellement changer d'un bloc à l'autre (donc au bout de 3 essais).

2.3 Procédure

Pour la passation, les participants sont placés dans le contexte d'un travail de classification d'images par distinction des lettres (lettre O) et des chiffres (chiffre 0) à l'aide d'un assistant basé sur un algorithme émulé d'IA (cf Figure 3). Leur tâche consiste à vérifier la validité de la classification proposée par l'IA pour certaines images et de réaliser par eux-mêmes la classification des images que l'IA n'a pas su traiter.
Lors de chaque session, une centaine d'images sont à traiter, comprenant environ 70 images classifiées par l'IA comme

1. https ://ensc.bordeaux-inp.fr

FIGURE 3 – Deux exemples de lettres (à gauche) et de chiffres (à droite).

étant des lettres, environ 10 images classifiées par l'IA comme étant des chiffres et environ 20 images non classifiées. L'interface (cf Figure 4) comprend 3 pages distinctes permettant d'accéder aux images non classées, aux images classées comme des lettres par l'IA et à celles classées comme des chiffres. L'utilisateur peut naviguer librement entre ces trois pages et modifier à son gré la classification de toutes les images, soit pour classer les images non traitées par l'IA, soit pour corriger ce qu'il estime être des erreurs de l'IA. Quand il considère avoir fini sa tâche, le bouton « terminer la session » le conduit à une page lui permettant de donner son ressenti sur la session réalisée, avant de passer à la session suivante. Le temps attribué à chaque session est au maximum de 3 minutes.

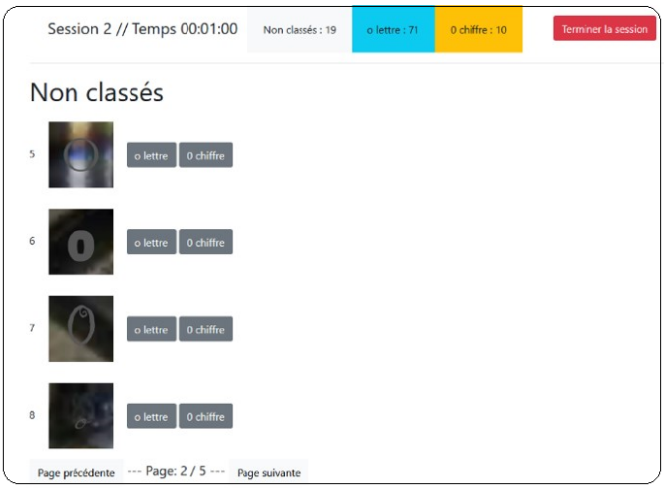

FIGURE 4 – Interface montrant la page des images non classées par l'IA. Les boutons en haut de la page (gris clair, bleu et orange) permettent d'accéder aux trois types de pages. Pour chaque image les deux boutons gris associés permettent de classer l'image dans une catégorie (le bouton devient alors coloré). Les boutons gris (page précédente / page suivante) en bas permettent d'accéder aux autres images de la page choisie.

Le déroulement de la passation comprend (cf Figure 5) : (1) un temps d'accueil (présentation de l'expérimentation au sujet), (2) un temps d'entraînement à la tâche et (3) l'expérimentation en elle-même (la réalisation des 6 sessions).

2.4 Données collectées et statistiques

Les variables dépendantes mesurées sont de 3 types. D'abord des mesures subjectives relatives au ressenti de

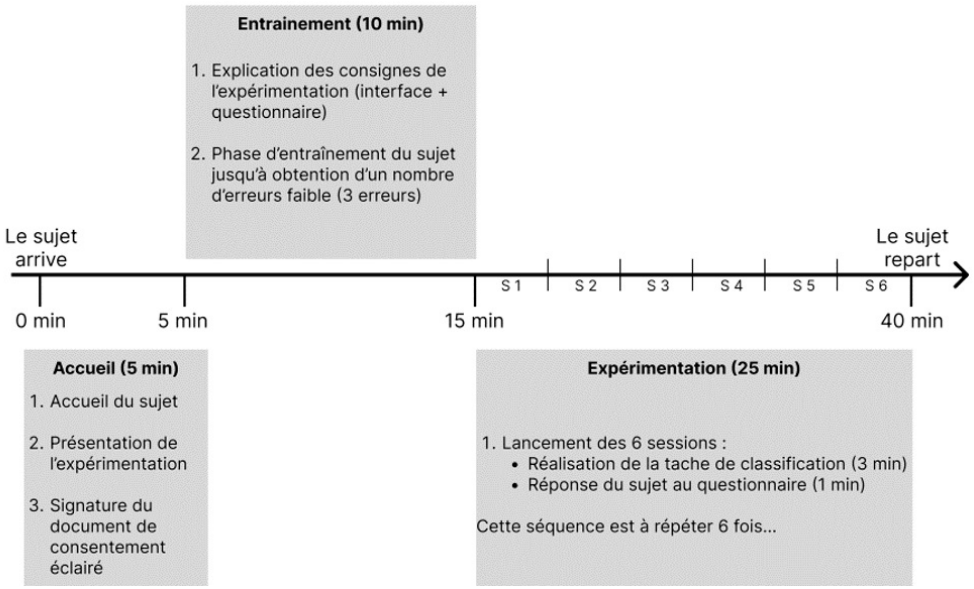

FIGURE 5 – Déroulé d'une passation.

l'opérateur. Elles comprennent : (1) l'évaluation de son niveau de confiance dans la classification réalisée par la machine (VD1 : Confiance_Rapportée), (2) son sentiment de paternité sur le travail fait (i.e. est-ce que la classification vient plutôt de moi ou bien de la machine ; VD2 : Paternité [9]) et (3) son appréciation du taux d'erreur de la machine qui l'amène à demander un ré-entraînement de l'algorithme d'IA (VD3 : ré-entraînement). Pour ces 3 mesures, la réponse est faite sur une échelle non segmentée comprise entre « Pas du tout » et « Totalement ». De plus, une mesure subjective de la charge de travail (VD4 : Charge_Travail) est réalisée à travers une version française du questionnaire du NASA TLX [4] qui comprend 6 dimensions (Exigence mentale, Exigence physique, Exigence temporelle, Succès, Effort, Frustration).

Des mesures comportementales sont recueillies à travers le temps passé sur les différentes pages de l'interface durant l'essai. Le comportement de supervision de l'IA par l'utilisateur est évalué par le temps total passé sur les pages contenant les images classifiées par l'IA (VD5 : Temps_Supervision). Le temps global passé pour réaliser la session est aussi enregistré (VD6 : Temps_Global).

Enfin, deux mesures de performance sont également collectées : le pourcentage d'erreurs de classification faites par le sujet (nombre d'images mal classées par l'opérateur / nombre d'images non classées par l'IA ; VD7 : Erreurs_Classification) ainsi que le nombre d'introduction d'erreurs par le sujet dans la classification de l'IA (image classée correctement par l'IA mais que l'opérateur choisit de classer autrement ; VD8 : Erreurs_Ajoutées).

Des statistiques inférentielles entre les groupes sont réalisées avec un seuil de signification de 0.05. Le terme de 'tendance' est utilisé dans cet article pour un seuil de 0.1. De manière générale, les résultats présentés concernent les comparaisons de 2 groupes indépendants et ces dernières ont été réalisées grâce à des t-tests (test unilatéral).

2.5 Hypothèses détaillées

La première hypothèse de cette étude porte sur la compréhension et l'adaptation de l'opérateur au niveau de fiabilité de l'IA. De manière plus spécifique les hypothèses suivantes sont formulées :

- H1.1 : la confiance dans le système d'aide sera plus faible quand la fiabilité de cette aide est plus faible.
- H1.2 : l'opérateur considérera que sa charge de travail est d'autant plus forte que la fiabilité de l'IA est faible.
- H1.3 : L'opérateur prendra plus de temps pour réaliser la tâche et supervisera plus son système d'aide lorsque celui-ci est peu fiable.
- H1.4 : les performances globales du couple Opérateur/système d'IA seront plus faibles quand l'IA est moins fiable.

La deuxième hypothèse porte sur l'impact de la confiance acquise par l'opérateur envers le système d'IA et sur son évaluation de la fiabilité actuelle de ce système. De manière plus précise, il est attendu que pour une fiabilité du système donnée, une confiance initiale plus basse de l'opérateur l'amène à reporter une confiance moindre dans le système.

3 Résultats

3.1 Hypothèse 1 : Adaptation du comportement

Il est attendu que chaque opérateur évalue (consciemment ou non) pendant les trois sessions du bloc 1 la fiabilité du système d'IA avec lequel il coopère et adapte son comportement à ce niveau de fiabilité. Si la fiabilité de l'IA reste la même, le comportement de l'opérateur pendant les 3 essais du bloc 2 devrait alors être représentatif du comportement final de cet opérateur pour ce niveau de fiabilité de l'IA.

Pour tester si les opérateurs adoptent des stratégies de travail avec l'IA qui dépendent de la fiabilité de l'IA, nous comparons donc les données du bloc 2 pour les groupes G1 et G2. Cette modification du ressenti et du comportement envers l'IA est décomposée en quatre parties que nous évaluons séparément.

H1.1 : fiabilité de l'aide et confiance rapportée

Pour mettre à l'épreuve cette sous-hypothèse, nous utilisons les variables dépendantes VD1, VD2 et VD3. Pour la confiance rapportée (VD1, voir figure 6) il est attendu que celle-ci soit plus faible dans le groupe 2 (IA peu fiable) que dans le groupe 1 (IA fiable). Le test de Student unilatéral montre une différence significative (t(14)=2.99,p<0.005).

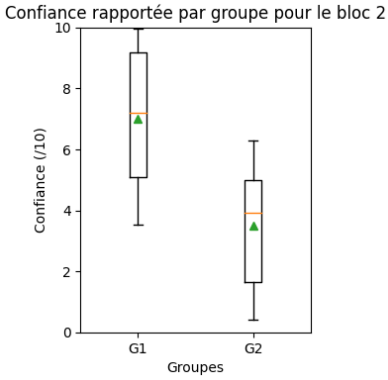

FIGURE 6 – VD1 : Niveau de confiance rapporté par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

Il est attendu que le sentiment de paternité (VD2, voir figure 7) soit plus fort dans le groupe 2 que dans le groupe 1, ce qui n'est pas le cas (t(14)=-2.59, p=0.01), même si les résultats statistiques indiquent tout juste une tendance.

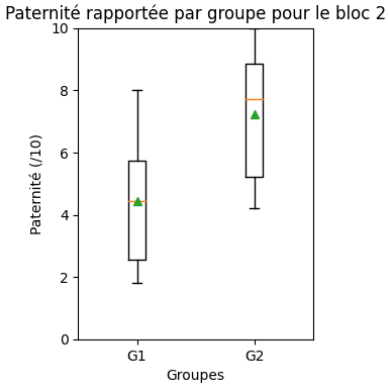

FIGURE 7 – VD2 : Paternité rapportée par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

Enfin, la demande de ré-entraînement de l'IA (VD3, voir figure 8) devrait être plus forte dans le groupe 2 que dans le groupe 1, ce qui est vérifié (t(14)=-2,75, p<0.001).

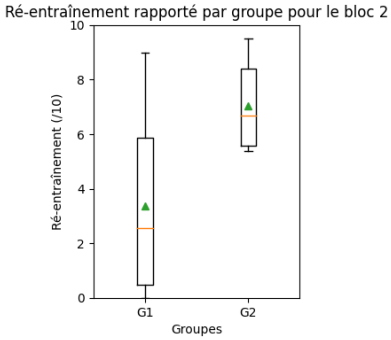

FIGURE 8 – VD3 : Ré-entraînement demandé par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

H1.2 : fiabilité et charge de travail ressentie

La deuxième composante de cette hypothèse concerne la charge de travail ressentie par l'opérateur qui devrait être plus forte quand l'IA est moins fiable. On attend donc une charge de travail rapportée (VD4, voir figure 9) plus forte dans le groupe 2 que dans le groupe 1, ce qui n'est pas le cas (t(14)=-1.08, p=0.15). En fait seule la composante liée à la pression temporelle augmente de manière significative pour le groupe 2 (t(14)=-1.92, p<0.05).

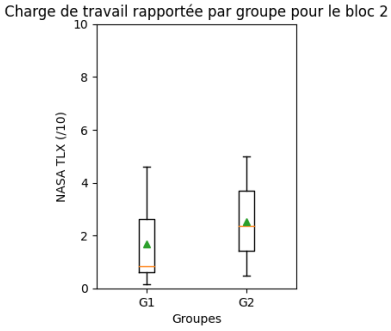

FIGURE 9 – VD4 : Charge de travail rapportée par les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

H1.3 : fiabilité et temps de travail

La troisième partie de l'adaptation de l'opérateur concerne le temps passé par l'opérateur pour réaliser la tâche. Il est attendu que quand la fiabilité de l'IA est plus faible, l'opérateur consacre plus de temps à superviser ce qui lui est proposé (VD5, voir figure 10) et que ceci se répercute sur le temps total passé pour réaliser la tâche (VD6, voir figure 11). Les analyses statistiques confirment une augmentation du temps de supervision de l'IA pour le groupe 2 (t(14)=-2.63, p<0.01) et une augmentation significative du temps total (t(14)=-1.83, p<0.05).

H1.4 : fiabilité et performances

Enfin la dernière partie de cette première hypothèse concerne l'impact de la fiabilité de l'IA sur les performances du couple opérateur-IA pour la réalisation de la

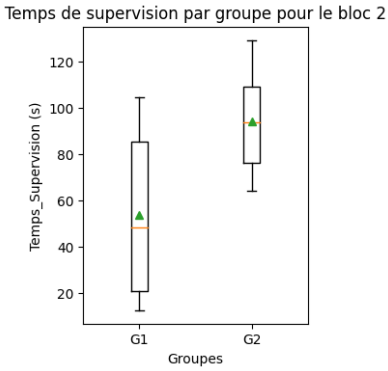

FIGURE 10 – VD5 : Temps de supervision pour les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

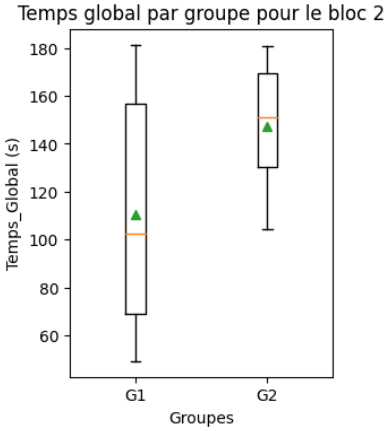

FIGURE 11 – VD6 : Temps total pour réaliser la tâche pour les 2 groupes d'utilisateurs. Le triangle vert indique la moyenne, la barre orange la médiane.

tâche. Il est attendu d'une part que la tâche de classification des images réalisée par l'opérateur seul (VD7) contienne plus d'erreurs quand la fiabilité de l'aide est faible (car l'opérateur est plus occupé par le travail de supervision de l'IA). Le nombre d'erreurs de classification faites par l'opérateur devrait donc être plus important pour le groupe 2, ce qui n'est pas le cas (t(14)=-0.50, p=0.31). De plus il est attendu que l'opérateur introduise plus d'erreurs dans les résultats de classification générés par l'IA (VD8) pour le groupe 2, ce qui n'est pas non plus le cas (t(14)=-0.47, p=0.32).

3.2 Hypothèse 2 : Influence de la confiance acquise

La deuxième hypothèse concerne l'impact de la confiance acquise dans le système d'aide sur la manière dont l'opérateur apprécie les performances de ce système à un moment donné. Pour cela, nous comparons d'une part les groupes 1 et 4 sur le bloc 2 pour regarder si une confiance apprise forte (G1) ou faible (G2) modifie la perception d'une même IA fiable pendant le bloc 2. Puis d'autre part nous regardons les groupes 2 et 3 pour évaluer si une confiance apprise faible (G2) ou forte (G3) impacte l'évaluation d'une même IA peu fiable pendant le bloc 2.

Seules les VD1 (confiance rapportée) et VD3 (nécessité de ré-entraîner l'IA) qui sont les plus caractéristiques du jugement conscient de la qualité de l'IA sont indiquées dans cet article.

Il est attendu ici que la confiance apprise module la confiance rapportée pour la session suivante et que de ce fait la confiance rapportée dans l'IA fiable à la première session du bloc 2 soit plus forte pour le Groupe 1 (confiance apprise forte) que pour le groupe 4. Ceci est confirmé par l'analyse statistique (t(14)=1.91, p<0.05). Il est de plus attendu que la nécessité de ré-entraînement de l'IA (VD3) soit plus élevée pour le groupe 4 que pour le groupe 1 (confiance apprise forte), ce qui n'est pas confirmé par les tests statistiques (t(14)=-1.13, p=0.14).

De même il est attendu que la confiance rapportée (VD1) dans l'IA peu fiable à la première session du bloc 2 soit plus forte pour le Groupe 3 (confiance apprise forte) que pour le groupe 2. Cette conjecture n'est pas confirmée (t(14)=-1.13, p=0.14). Il est de plus attendu que la nécessité de ré-entraînement de l'IA (VD3) soit plus élevée pour le groupe 2 que pour le groupe 3 (confiance apprise forte), ce qui n'est pas non plus confirmé par les tests statistiques (t(14)=-0.46, p=0.67).

4 Discussion

L'expérimentation menée dans cette étude permet de comparer deux à deux les comportements de groupes distincts d'opérateurs qui réalisent une tâche de classification d'images tout en bénéficiant d'un système d'aide à la décision. Le premier bloc de l'expérimentation repose sur trois essais successifs nécessitant de classer à chaque fois une centaine d'images. Ce bloc 1 permet aux sujets de se familiariser avec la tâche mais aussi avec le système d'assistance proposé. Ces essais leur permettent d'optimiser leur manière de coopérer avec cette aide, même si aucune information ne leur est donnée, ni sur la fiabilité de l'aide qui leur est proposée, ni sur la qualité du travail final qu'ils ont réalisé conjointement.

Les deux premiers groupes d'utilisateurs considérés réalisent la tâche soit avec une assistance fiable qui ne fait pas d'erreur de classification des images (Groupe1), soit avec une assistance peu fiable ayant un taux d'erreur de l'ordre de 20%. Les résultats obtenus montrent que les deux groupes d'opérateurs perçoivent cette différence dans la performance du système d'aide et modulent leur comportement en fonction de ce ressenti. Le report conscient de la qualité du système d'aide qui leur est proposé se retrouve tout d'abord dans la confiance attribuée au système. Ceci montre bien que les opérateurs ont supervisé le travail du système d'aide et qu'ils sont sensibles aux erreurs commises par cette IA. La présence d'erreurs dégrade la confiance attribuée à l'aide. En parallèle, et de manière cohérente, les opérateurs collaborant avec le système le moins fiable demandent plus que les autres un réajustement de

cette aide, confirmant cette prise en compte du taux d'erreur de l'aide. La fiabilité de l'aide proposée est donc bien dans cette étude un facteur qui module la perception qu'en a l'opérateur.

Par contre, l'opérateur ne se considère pas fortement plus impliqué dans la classification globale des images. Même si une tendance vers une implication plus forte se dessine, les opérateurs ne s'attribuent pas réellement une part plus importante du travail réalisé. Ceci est corroboré par le fait que la charge de travail ressentie n'est pas différente pour les deux groupes, ce qui tend à montrer que l'opérateur continue à tirer profit de cette aide.

Il y a donc à la fois une acceptation de l'aide et une adaptation de la confiance qui lui est attribuée. Ceci amène l'opérateur à modifier la manière dont il supervise les propositions de cette IA. Les résultats montrent qu'en effet le groupe utilisant une IA moins fiable consacre plus de temps à la superviser, c'est-à-dire à vérifier la validité de la classification des images qu'elle propose. Cette augmentation du temps nécessaire pour valider les propositions du système d'aide ne se fait pas au détriment du temps consacré à sa propre tâche de classification des erreurs, mais en augmentant le temps total utile pour finaliser l'essai.

Cette adaptation comportementale résulte en partie de ce que le temps laissé à l'opérateur pour réaliser la tâche (3 minutes) n'amenait pas une contrainte temporelle forte. Quand la fiabilité de l'aide est faible et que sa supervision devenait plus importante cette limite temporelle commençait à devenir une contrainte plus forte, ce qui est confirmé par le résultat sur la dimension « pression temporelle » de la charge de travail qui est significativement plus forte pour le groupe travaillant avec l'IA de plus faible fiabilité.

Enfin les résultats montrent que, contrairement à ce qui était attendu, cette diminution de la fiabilité de l'aide ne fait pas baisser les performances de l'opérateur dans sa tâche spécifique, ni la performance globale. Il semble donc ici que l'ajustement de la supervision du travail réalisé par le système d'aide ne vienne pas contraindre de manière sensible la tâche de l'opérateur qui garde le même niveau d'exigence pour sa partie de la classification et s'attribue suffisamment de temps pour juger de manière adéquate le travail de son système d'aide.

Ces résultats montrent que quand la pression temporelle n'est pas trop forte, l'opérateur reste investi dans son travail collaboratif avec une IA dont la fiabilité est plus faible. Dans un cadre industriel, par exemple pour un contrôle qualité où le temps disponible pour réaliser l'ensemble des contrôles est souvent assez contraint, la tolérance de l'opérateur à un taux d'erreur plus important du système d'aide pourrait être plus faible. Le protocole présenté dans cette étude reste en mesure d'investiguer ce type de problématique, soit en réduisant le temps disponible pour la tâche, soit en augmentant le taux d'erreur de l'IA jusqu'à observer un point de rupture dans l'utilisation de l'aide. En particulier il serait pertinent de regarder de manière plus précise le comportement de l'opérateur, dans un premier temps avec les stratégies de parcours des différentes pages de l'aide (existe-t-il des stratégies de parcours différentes selon la pression temporelle ou le taux d'erreur ?), et de manière plus fine à l'aide d'un oculomètre (le regarde parcourt-il toutes les images ?).

La deuxième partie des résultats a permis d'explorer si l'apprentissage de la confiance dans le système influençait, à un instant donné, l'évaluation que se fait l'opérateur de la fiabilité du système d'aide. Les résultats obtenus sont moins tranchés. L'évaluation d'une IA fiable semble dépendre de la confiance apprise, le groupe ayant appris à avoir peu confiance restant plus méfiant que l'autre groupe. Par contre, en sens inverse, quand les deux groupes sont confrontés à une IA peu fiable, il ne semble pas y avoir de différence. Ceci tendrait à montrer que, dans ce contexte, il est plus compliqué (plus long) de gagner la confiance que de la perdre. Ce résultat est certainement lié au fait que la situation expérimentale amenait les utilisateurs à rester impliqués dans leur tâche, de par la partie active de classification qu'ils avaient à réaliser et par le fait que la durée était relativement restreinte (6 sessions). Il semblerait que dans ce cadre il n'y ait pas de phénomène de surconfiance qui se soit installé chez les opérateurs, ce qui ne serait pas forcément le cas avec un usage journalier, beaucoup plus répétitif. Le protocole proposé pourrait permettre d'investiguer plus en profondeur cette problématique et chercher s'il est possible de mettre en évidence une habituation au système qui atténuerait les facultés de l'opérateur à identifier une baisse de performance du système.

5 Conclusion

Le travail présenté dans ce papier a consisté à explorer les attitudes d'utilisateurs mis face à un système à base d'IA. Aujourd'hui, expérimentales, ces situations ne vont pas tarder à se généraliser dans le monde professionnel. Le cas d'usage que nous avons choisi s'inspire de situations réelles du contrôle en continu de la qualité d'une production industrielle. Certains de ces contrôles, essentiellement visuels, vont très vite être opérés par des systèmes à base d'IA, charge à l'opérateur de contrôler ce contrôle. Dans ces situations hybrides où humains et machines vont devoir collaborer, la question de la confiance de l'opérateur dans le système à base d'IA devient une question centrale. Pour étudier cette question, nous proposons un dispositif où l'opérateur doit distinguer visuellement, sur une image bruitée, des chiffres zéros et des lettres O. Il s'agit d'une tâche de classification très facilement compréhensible, facile à opérer, mais pouvant demander un fort investissement cognitif en fonction du rythme imposé et du bruit dans les images à classer. L'apport d'un système d'IA y est donc salutaire tout en restant expérimentalement contrôlable dans ses performances.

Nous retrouvons dans ce dispositif expérimental les tendances attendues dans la construction de la confiance que l'opérateur se fait dans le système. Cela s'observe dans les modifications de son ressenti, de ses performances et de son comportement. Les résultats montrent une adaptation du comportement en fonction du niveau de fiabilité du système. C'est ainsi que nous mettons en évidence que

la confiance rapportée est plus forte et la supervision est moins importante lorsque la fiabilité de l'IA est forte. Ce genre d'investigation est nécessaire pour préparer la mise en service d'outils d'aide à la décision dans la production industrielle.

Bien que les premiers résultats soient assez attendus, comme par exemple que la fiabilité de l'IA ait un impact direct sur la confiance qui lui est accordée et sur la manière de la superviser, cette étude montre que ces résultats sont vérifiables empiriquement et apporte un protocole expérimental réutilisable. Ceci ouvre de nouvelles perspectives pour l'étude des interactions entre un opérateur et un système d'IA. Plusieurs pistes sont actuellement explorées.

La première piste porte sur la compréhension des distinctions engendrées par l'utilisation d'un système d'assistance traditionnel par rapport à un système d'aide à base d'IA. Une étude récente (Munoz et al., en cours de soumission) met en lumière l'impact de l'agent proposant une assistance (humain, système automatisé classique ou IA) sur la décision d'un opérateur lorsqu'il doit choisir entre différentes formes d'aide. Les résultats suggèrent que la nature de l'agent influence l'acceptation de l'aide, avec des implications différentes selon qu'il s'agisse d'un système automatisé classique ou basé sur l'IA. Par conséquent, la confiance accordée à ces systèmes et leur acceptation varient en fonction de leur spécificité, ce qui soulève la question de leur utilisation dans des contextes de travail répétitif et de la manière dont la confiance se développe au fil du temps.

La seconde perspective de recherche aborde la question de l'explicabilité des systèmes d'IA et de son influence sur la construction de la confiance envers le système d'assistance. De nombreuses études se penchent sur la nature de ces explications et leur impact sur la collaboration avec l'opérateur, notamment en ce qui concerne l'acceptabilité de l'aide, la confiance dans les décisions prises, et les performances dans l'exécution des tâches (voir par exemple Larasati et al., 2023 ; Le Guillou et al., 2023). Le protocole expérimental proposé dans cette étude offre une opportunité de manière plus pragmatique pour évaluer l'effet de différents types d'explications sur la construction de la confiance dans les systèmes d'IA. Dans cette optique, la fiabilité de la classification demeurerait constante, mais des éléments explicatifs de la classification des images seraient ajoutés dans une des conditions afin de mesurer leur effet sur l'exécution de la tâche (voir par exemple Van der Waa et al., 2021 pour une introduction à l'évaluation des explications dans les systèmes d'assistance basés sur l'IA).

Les recherches menées par Merritt et ses collègues (2015) examinent les représentations que les individus ont des systèmes d'aide et leur impact sur la confiance qu'ils accordent à ces systèmes. Ils démontrent notamment que certaines personnes ont des attentes dichotomiques (confiance totale ou défiance) et que la relation entre la fiabilité de l'assistance et la confiance reportée n'est pas forcement linéaire. De manière plus fondamentale, notre étude vise à éclairer les mécanismes et les facteurs qui contribuent à l'établissement d'un lien de confiance entre un opérateur et un système d'aide, quelle que soit sa nature. Une perspective à explorer ici consiste à examiner expérimentalement la nature de cette relation entre la fiabilité de l'assistance et la confiance établie, en la faisant varier de manière plus graduelle (en testant des niveaux de fiabilité intermédiaires). Cela permettrait d'obtenir des données pour mieux comprendre la transition entre les comportements de confiance et de méfiance envers un système d'aide.

Remerciements

Ce travail a obtenu le soutien du gouvernement français dans le cadre du programme "France 2030", au sein de l'Institut de Recherche Technologique SystemX et du projet Confiance.ai [2].

Références

[1] Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. Explainable artificial intelligence : an analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5), September 2021.

[2] Laurent Chaudron, Jean-Marie Burkhardt, Lisa Chouchane, Pauline Muñoz, Nicolas Maille, and Anne-Lise Marchand. Trust : The vital fluid of interactions. In *AHFE 2023*, 2023.

[3] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5) :517–527, 2011.

[4] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index) : Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[5] Kevin Anthony Hoff and Masooda Bashir. Trust in automation : Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3) :407–434, 2015.

[6] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. Trust in Artificial Intelligence : Meta-Analytic Findings. *Human Factors : The Journal of the Human Factors and Ergonomics Society*, 65(2) :337–359, March 2023.

[7] Laurent Karsenty. Comment appréhender la confiance au travail. *La confiance au travail*, pages 13–51, 2013.

[8] Nathan McNeese, Mustafa Demir, Erin Chiou, Nancy Cooke, and Giovanni Yanikian. Understanding the role of trust in human-autonomy teaming. 2019.

[9] Adrien Metge. *Opérateurs et systèmes intelligents : se comprendre pour décider. Application à la supervision de drones.* PhD thesis, Université de Bordeaux, 2022.

2. https ://www.confiance.ai/

[10] Pauline Munoz, Anne-Lise Marchand, Laurent Chaudron, and Nicolas Maille. Cl : Confiance en l'autre : approche expérimentale de l'arbitrage entre le partenaire humain et le partenaire automatisé. In *12ème Colloque de Psychologie Ergonomique EPIQUE 2023*, 2023.

[11] Yiteng Pan, Fazhi He, Haiping Yu, and Haoran Li. Learning adaptive trust strength with user roles of truster and trustee for trust-aware recommender systems. *Applied Intelligence*, 50 :314–327, 2020.

[12] Isabella Seeber, Eva Bittner, Robert O Briggs, Triparna De Vreede, Gert-Jan De Vreede, Aaron Elkins, Ronald Maier, Alexander B Merz, Sarah Oeste-Reiß, Nils Randrup, et al. Machines as teammates : A research agenda on ai in team collaboration. *Information & management*, 57(2) :103174, 2020.

[13] Thomas B Sheridan. Individual differences in attributes of trust in automation : measurement and application to system design. *Frontiers in Psychology*, 10 :1117, 2019.

[14] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to Evaluate Trust in AI-Assisted Decision Making ? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2) :1–39, October 2021.