Confiance et biais d'automatisation : différences entre novices et experts dans un contexte militaire

Trust and automation bias: differences between novices and experts in a military context

Anne-Lise Marchand[1], Nicolas Maille[2], Pauline Munoz[1], Laurent Chaudron[1]

[1] CREA, Centre de recherche de l'Ecole de l'Air et de l'Espace, Salon de Provence, France
[2] ONERA, Salon de Provence, France

RÉSUMÉ. L'étude vise à examiner si le biais d'automatisation dans des situations d'arbitrage entre une aide humaine et une aide basée sur l'IA varie en fonction des caractéristiques psychosociales des individus. La littérature met en évidence la robustesse du biais d'automatisation dans les situations de prise de décision avec une seule aide, mais quelques études récentes mobilisant le paradigme de la double aide à la décision identifient des résultats plus nuancés, notamment en fonction des caractéristiques des participants. 2 groupes de participants (37 élèves pilotes militaires vs 37 pilotes opérationnels) sont engagés dans une simulation de mission aérienne d'attaque au sol, où ils doivent choisir entre les informations fournies par une aide humaine et celles fournies par une aide automatisée basée sur l'IA. La confiance en ces aides est induite *a priori* par des niveaux de fiabilité prédéfinis (20%, 50%, 70% 90%). A fiabilité égale, lorsque les jeunes participants et les experts sont confrontés à une aide humaine et une aide basée sur l'IA, ils ont une préférence pour l'aide humaine. Toutefois cette préférence est plus importante pour les experts. L'étude remet en question l'invariabilité du biais d'automatisation, soulignant l'impact des caractéristiques psychosociales de l'opérateur sur la prise de décision. Il semble nécessaire de reconsidérer le biais d'automatisation dans des contextes modernes au travers des représentations individuelles des technologies pour optimiser la conception des systèmes d'aide à la décision.

ABSTRACT. This study examines whether the automation bias in situations of arbitration between human and AI-based assistance varies as a function of individuals' psychosocial characteristics. The literature highlights the robustness of the automation bias in decision-making situations with a single aid, but a few recent studies mobilizing the dual decision aid paradigm identify more nuanced results, particularly as a function of participants' characteristics. 2 groups of participants (37 military pilot students vs. 37 operational pilots) are engaged in an close air support mission simulation, where they must choose between information provided by a human aid and that provided by an AI-based automated aid. Trust in these aids is induced a priori by predefined levels of reliability (20%, 50%, 70% 90%). With equal reliability, when young participants and experts are confronted with a human aid and an AI-based aid, they have a preference for the human aid. However, this preference is greater for experts. The study questions the invariability of automation bias, highlighting the impact of the operator's psychosocial characteristics on decision-making. It seems necessary to reconsider the automation bias in modern contexts through individual representations of technologies to optimize the design of decision support systems.

MOTS-CLÉS. Biais d'automatisation, Confiance, Fiabilité, Man Machine Teaming, Intelligence artificielle, Prise de décision.
KEYWORDS. Artificial Intelligence, Automation Bias, Decision-making, Man Machine Teaming, Reliability, Trust.

1. Introduction

Dans un environnement où l'intelligence artificielle (IA) redéfinit les limites de l'automatisation, Il est pertinent de s'interroger sur l'évolution du comportement des individus en présence d'aides mettant en œuvre des technologies avancées, notamment celles qui utilisent de l'IA. Des études antérieures suggèrent que les représentations que les individus ont des systèmes d'aide peuvent varier en fonction de leur âge [EZE 08]. Par extension, on peut supposer que ces représentations influent sur le comportement des individus, notamment en modulant leur propension à faire confiance de manière trop privilégiée aux systèmes automatisés, appelé biais d'automatisation. Cette dynamique soulève un questionnement : lorsqu'un individu est confronté au choix entre deux types d'aides, sa prise de décision et sa confiance entre une aide automatisée basée sur l'IA et une aide humaine varie-t-elle en fonction de ses caractéristiques psychosociales ? Cet article aborde cette question au travers d'un cadre théorique assimilant confiance interpersonnelle et confiance en l'aide automatisée. Fondée sur les récents résultats de la littérature sur l'apparition affaiblie du biais d'automatisation en situation

d'arbitrage [DZI 02] [EZE 08] [PER 10] [RAJ 08], cette étude vise à tester si la tendance à privilégier l'aide automatisée est plus marquée chez les personnes plus âgées et plus expérimentées [EZE 08].

1.1. *Le biais d'automatisation*

1.1.1. *Un phénomène robuste ?*

Les erreurs humaines contribuent à 70-80% des accidents aériens [SAR 00]. Parmi ces erreurs, celles résultant de la sensibilité au biais d'automatisation ont fait l'objet d'études empiriques approfondies sur les interactions entre le pilote et le système [LAY 94] [MCG 06] [MOS 92] [MOS 98] [SAR 01].

Le biais d'automatisation, défini par Mosier et al. [MOS 98] résulte du fait que les humains utilisent des aides à la décision *« comme un remplacement heuristique de la recherche et du traitement vigilants de l'information »* (p. 205). En d'autres termes, le biais d'automatisation se caractérise par une tendance à accorder une confiance excessive aux systèmes automatisés, ce qui peut entraîner une dépendance inappropriée en ces systèmes [PAR 10]. Dans ce cadre ce biais est directement lié au niveau d'implication de l'opérateur humain dans le suivi de l'activité.

D'après la revue de littérature de Parasuraman et Manzey [PAR 10], le biais d'automatisation est un phénomène robuste qui s'observe dans divers contextes, tels que la médecine [WES 05] ou l'aviation [MOS 98] [SAR 01]. Sa persistance semble être étroitement liée à la fiabilité globale de l'aide automatisée [ROV 07] et ce biais se montre résistant aux tentatives d'élimination au travers de formations ou d'instructions explicites visant à vérifier les recommandations de l'aide automatisée [MOS 01]. Ses répercussions s'étendent à la prise de décision, qu'elle soit individuelle ou collective au sein d'équipes [SKI 00].

Cependant, cette robustesse est remise en question par de récentes études [DZI 02] [PER 10] [RAJ08] [EZE 08] qui mettent en évidence des nuances dans la manifestation de ce biais. Il semble notamment que le biais d'automatisation soit sensible au contexte ; en effet, en présence d'un niveau de risque élevé, la propension à opter pour l'automatisation diminue [PER 10] [RAJ08] [SAT 17]. Les variations individuelles comme l'âge jouent également un rôle dans l'apparition du biais [EZE 08]. Enfin ce phénomène semble être sensible aux caractéristiques concernant le type d'aide à la décision [DZI 02]. Autrement dit, les récentes recherches suggèrent que le biais d'automatisation est influencé par des variables telles que le risque [LIE 21] [PER 10] [RAJ08] [SAT 17], les différences individuelles [EZE 08] et les caractéristiques des aides automatisées [DZI 02].

1.1.2. *Un nouveau paradigme expérimental : le paradigme de la double aide à la décision*

L'affaiblissement du biais d'automatisation est également observé dans plusieurs études mobilisant un autre paradigme expérimental : celui de la double aide à la décision [LER 97] [LYO 12] [MER 15] [PEA 16] [PEA 19] [ZHA 21]. Ce paradigme est caractérisé par une situation d'arbitrage dans laquelle un individu prend une décision assistée par deux aides différentes. Une différence notable est que ce nouveau paradigme ne repose pas sur une implication plus ou moins grande de l'opérateur (un biais entre lui-même et une aide), mais sur le choix de déléguer la tâche en privilégiant une aide par rapport à une autre (un biais entre les deux aides, l'une automatisée et l'autre humaine). Ce paradigme ressemble davantage à certaines situations aéronautiques contemporaines où l'opérateur doit prendre des décisions en comparant plusieurs sources d'informations d'origines humaines ou automatisées.

L'ensemble des résultats des études mobilisant ce paradigme fragilisent la robustesse du biais d'automatisation même si leurs résultats sont nuancés. Par exemple, certaines études mobilisant ce paradigme se sont intéressées à la perception des aides en fonction de leur expertise perçue (pedigree). Lerch et al. [LER 97], Pearson et al. [PEA 19] ou Zhang et al. [ZHA 21] constatent que lorsque l'expertise perçue des aides est forte, l'individu a tendance à attribuer une confiance plus élevée à l'humain qu'à l'aide automatisée, suggérant que la préférence à choisir l'aide humaine lorsque les deux aides sont expertes est due à un biais de perception du pedigree, la perception de l'expertise pour l'humain étant

plus élevée que celle de l'automatisation [PEA 19]. Merritt et al. [MER 15] choisissent de manipuler l'expertise perçue des aides au travers de la fiabilité c'est-à-dire que la fiabilité de l'aide est induite au cours de plusieurs interactions avec le participant. Les résultats de cette étude montrent des choix initiaux en faveur de l'automatisation qui évoluent au fur à mesure des interactions en faveur de l'humain.

D'autres études obtiennent des résultats contradictoires en modulant la perception du risque ; ainsi, alors que Lyons et Stokes [LYO 12] montrent que plus le risque augmente, plus les participants choisissent l'aide automatisée, Pearson et al. [PEA 16] n'observent pas de différence significative dans les choix des participants entre l'aide humaine et l'aide automatisée. Toutefois, à l'instar des variations obtenues lors de protocoles classiques [EZE 08], ces résultats contradictoires peuvent être expliqués par des caractéristiques psychosociales différentes des participants, notamment l'âge et l'expérience professionnelle ; dans le cas de l'étude de Lyons et Stokes [LYO 12], les 40 participants sont tous issus d'une base aérienne (dont 40% de militaires) et ont 36 ans en moyenne, alors que dans le cas de l'étude de Pearson et al. [PEA 16], les 126 participants sont des étudiants de 19 ans en moyenne.

Ces résultats, contrastés dans le paradigme de la double aide à la décision, et plus généralement dans l'étude du biais d'automatisation, soulèvent des questions sur la manière dont la représentation de l'aide automatisée influe sur ce phénomène complexe. En effet, les aides automatisées mobilisées dans les années 70' n'ont parfois plus grand-chose à voir avec celles utilisées aujourd'hui [MOS 98] [ZHA 21]. D'ailleurs, dans les protocoles actuels, la caractérisation des aides automatisées proposées aux participants varie fortement, ainsi que l'explicitent Kaplan et al. [KAP 21] dans leur méta-analyse. Ils définissent notamment l'IA comme une "technologie logicielle permettant aux machines automatisées de percevoir leur environnement et de prendre des décisions intelligentes sur la base des données disponibles" (p.1). Cette adaptabilité de l'IA constitue l'une de ses caractéristiques les plus notables, la distinguant des systèmes automatisés basiques [FRA 03] [GLI 20] [RAH 19]. Autrement dit, cette caractéristique crée une distinction entre l'IA et d'autres formes d'automatisation, à partir de laquelle il est plausible de supposer que les résultats des expériences menées dans les années 70 ne peuvent pas être comparées directement aux résultats des expériences menées de nos jours.

Pris ensemble, l'hétérogénéité des résultats du champ de recherche étudiant le biais d'automatisation, à travers l'un au l'autre des paradigmes, pourrait s'expliquer non seulement par le fait que les aides automatisées mobilisées dans les protocoles antérieurs ne correspondent plus à celles mobilisées plus récemment, mais également parce que la représentation des individus sur les aides automatisées a évolué. Cette différence peut être attribuée à une familiarité avec les technologies numériques et à une adaptation plus rapide aux nouvelles innovations, ce qui renforce leur confiance dans ces systèmes [YIG 22]. Cette évolution de la représentation peut donc découler de divers facteurs tels que la familiarité croissante avec les technologies automatisées, une exposition plus régulière à son fonctionnement et donc à ses limites, les avancées constantes dans ce domaine, ou même les représentations médiatiques changeantes de l'automatisation. Autrement dit, la représentation des aides automatisées et en particulier de l'IA s'est hétérogénéisé parce que l'exposition à l'IA est différente en fonction du groupe social auquel l'individu appartient. Cette hétérogénéité devrait se retrouver en particulier entre les générations. Ainsi, le postulat considéré ici est que les individus plus jeunes ont construit une représentation plus exacte des possibilités de l'IA que leurs ainés, car ils y sont plus souvent confrontés. Cette évolution de la représentation pourrait contribuer à la variabilité des réponses observées dans les études sur le biais d'automatisation.

Les résultats des études mobilisant le paradigme de la double aide à la décision élargissent le champ d'étude du biais d'automatisation. Celui-ci ne se limite plus au choix entre réaliser la tâche soi-même ou la déléguer à une aide automatisée mais prend aussi en compte les cas où l'opérateur doit choisir entre déléguer cette tâche soit à un autre humain soit à un système automatisé. Dans cet article, il est donc considéré qu'un biais d'automatisation se manifeste quand une personne fait reposer son comportement ou sa décision sur une aide automatisée alors qu'un autre comportement ou une autre décision est à sa portée en utilisant d'autres type d'aides ou en s'impliquant davantage dans la tâche.

1.2. *Biais d'automatisation et confiance*

Au cours de ces différentes études, il apparaît que le biais d'automatisation est modulé par un élément clé : la confiance [DZI 01] [MER 15] [PAR 10] [PEA 19]. Les perspectives de ces différents travaux suggèrent que la manière dont un individu fait confiance à un système automatisé influence directement la prévalence et l'intensité du biais d'automatisation qu'il peut manifester. La confiance attribuée est également identifiée comme un facteur prédictif dans la prise de décision, et plus précisément dans le biais d'automatisation [MER 15] [PEA 19]. Ce constat est renforcé par les résultats de Parasuraman et al. [PAR 10] qui, en reliant la définition même du biais d'automatisation à la confiance, établissent un lien direct entre ces deux concepts. Cette relation est nécessaire pour comprendre comment le biais d'automatisation se forme et évolue dans divers contextes. De plus, il apparaît que le biais d'automatisation réagit aux mêmes facteurs que ceux qui influent la confiance envers l'automatisation dans le modèle de Hoff et Bashir [HOF 15] . Ces facteurs communs suggèrent qu'une approche globale peut donner accès à une meilleure compréhension de la dynamique du biais d'automatisation.

Le modèle proposé par Hoff et Bashir [HOF 15] indique que la confiance en l'automatisation se construit antérieurement à l'interaction sur la base de trois facteurs principaux (*Cf.* Figure 1) : la confiance dispositionnelle, la confiance acquise et la confiance situationnelle.

- La confiance dispositionnelle est liée à la personne qui est amenée à faire confiance, nommée le trustor : elle est influencée par des traits de personnalité, l'âge, l'expérience et la culture, et elle représente une tendance générale d'un individu à accorder ou à refuser sa confiance envers l'automatisation en fonction de ses caractéristiques ;
- La confiance situationnelle est liée à la situation : elle est influencée par l'environnement externe et les caractéristiques spécifiques au contexte de l'opérateur ;
- La confiance apprise est liée au système auquel la confiance va être accordée, le trustee : elle est influencée par les performances passées de l'automatisation et la connaissance préexistante du système automatisé par l'opérateur.

Par ailleurs, une fois que l'opérateur interagit avec le système automatisé, un processus de confiance dynamique se met en place et modifie cette confiance apprise initiale dans le système automatisé au cours du temps.

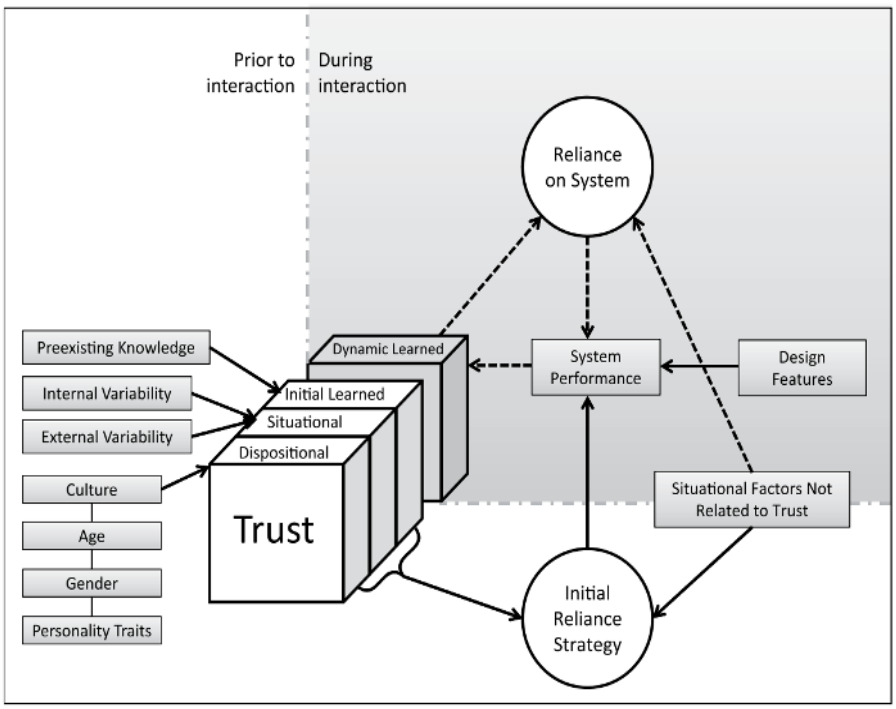

Figure 1. *Modèle de la confiance en l'automatisation [HOF 15]*

Ces trois sources principales de variabilité – confiance dispositionnelle, confiance situationnelle et confiance apprise – correspondent en partie aux influences du biais d'automatisation identifiées dans la littérature ; [EZE 08] montrent l'influence de l'âge des participants, qui est un des paramètres de la confiance dispositionnelle, sur le biais d'automatisation. De même la manipulation du risque, qui contribue à la confiance situationnelle, perturbe l'apparition du biais d'automatisation [PER 10] [RAJ08] [SAT 17]. Enfin, le niveau d'expertise perçue du partenaire, qui contribue à la confiance apprise, module aussi ce biais d'automatisation [PEA 19] [ZHA 21]. Cette cohérence conceptuelle entre les mécanismes du biais d'automatisation et le modèle d'Hoff et Bashir [HOF 15] met en évidence le rôle central de la confiance dans ce phénomène, comme souligné par Parasuraman & Manzey [PAR 10] . Toutefois ce modèle a été conçu pour les situations à simple conseiller, traditionnellement étudiées dans la littérature sur le biais d'automatisation. Ainsi il est nécessaire de l'adapter pour mobiliser son cadre d'analyse à des situations à plusieurs conseillers, notamment des conseillers humains.

Dans le cas d'une aide à la décision humaine, la confiance se construit autour de caractéristiques, telles que la bienveillance, l'intégrité et la compétence [MAY 95]. Dans le cas d'un système automatisé, les éléments principaux sur lesquels se construit la confiance incluent la performance, le processus, et l'alignement des objectifs [LEE 04] . Ces différences apparentes ne sont toutefois pas disqualifiantes pour un modèle commun : en effet, la fiabilité du trustee est au cœur des modèles de la confiance interpersonnelle et de la confiance dans un système automatisé [PAR 97] [REM 85]. Ainsi, Lee et See [LEE 04] mettent en avant la performance et la fiabilité comme éléments centraux de la confiance en l'automatisation, affirmant que la capacité d'un système à produire des résultats prévisibles et constants constitue un pilier fondamental de cette relation. De manière similaire, Rempel et al. [REM 85] soulignent que la fiabilité est la base principale de la confiance interpersonnelle, jouant un rôle déterminant dans la solidité des relations humaines. Dans le cadre du paradigme du double conseiller, la fiabilité des trustees s'avère également cruciale [MER 15]. Pris ensemble, ces différents cadres théoriques montrent que la fiabilité perçue du trustee est un facteur commun majeur de la confiance apprise. Toutefois la variabilité résiduelle des éléments pris en compte dans la confiance apprise indique la pertinence de différencier celle-ci en fonction de la nature du *trustee*, qu'il soit humain ou automatisé (*Cf.* Figure 2).

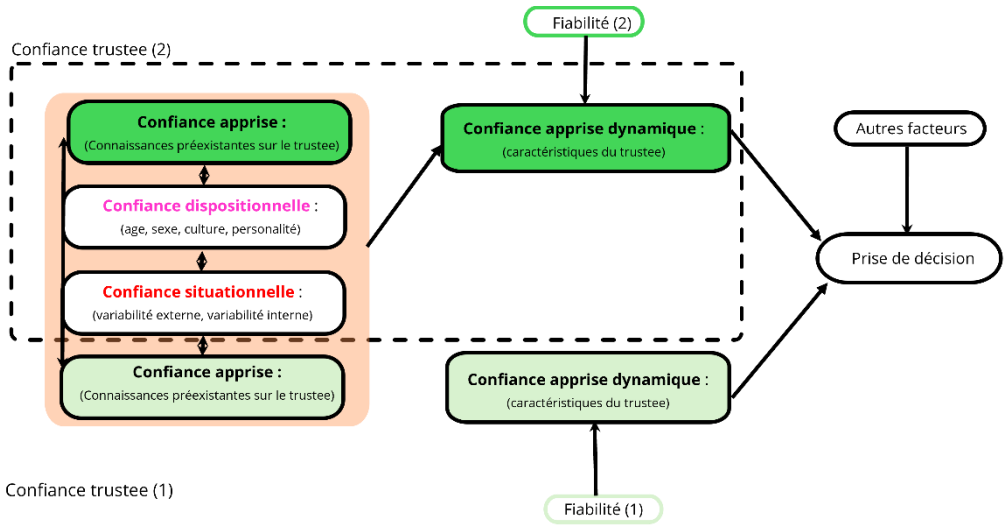

Figure 2. *Modèle de la confiance dans le paradigme du double conseiller*

Ce modèle propose une approche qui combine les éléments de confiance dispositionnelle et situationnelle tout en distinguant les caractéristiques spécifiques de la confiance apprise dynamique pour chaque type de *trustee*. Contrairement au modèle de Hoff et Bashir [HOF 15] , qui s'intéresse principalement à la confiance développée à travers des interactions répétées, ce modèle se distingue par sa focalisation sur la confiance générée uniquement à la première interaction. Cette approche est

particulièrement pertinente dans des contextes où les utilisateurs doivent prendre des décisions rapides sans disposer d'un historique d'interaction avec le conseiller, qu'il soit humain ou automatisé.

A partir de ce modèle, cette étude explore donc les liens entre confiance et biais d'automatisation dans le paradigme de la double aide à la décision : pour cela, elle examine si la confiance apprise, au travers de la fiabilité perçue et de la nature des deux aides, constitue un facteur explicatif de ce biais, et évalue l'influence de la confiance dispositionnelle au travers de l'effet de l'âge et de l'expérience de l'opérateur sur ce biais d'automatisation. Ainsi, dans cette expérience, la confiance apprise pour chacune des aides est manipulée au travers de la fiabilité des trustees, la confiance dispositionnelle est manipulée au travers de caractéristiques psychosociales des participants, spécifiquement leur âge et leur expérience, et la confiance situationnelle est contrôlée par des conditions expérimentales uniformes.

1.3. *Problématique et hypothèses*

Pour vérifier le lien entre confiance et biais d'automatisation, il s'agit d'observer des différences d'arbitrage en fonction de la confiance dans chacune des aides. En ayant choisi de fixer la confiance situationnelle (même tâche, dans les mêmes conditions) la confiance globale peut alors être modifiée soit par sa composante dispositionnelle (les caractéristiques du trustor) ou apprises (celles du trustee). Or, pour Parasuraman et Riley [PAR 97], Sheridan [SHE 02], Lee et See [LEE 04] , Wohelber [WOH 16], ou Chavaillaz et al. [CHA 16], la fiabilité est le principal facteur influençant la confiance envers un partenaire. De même pour Kaplan et al. [KAP 21] , la fiabilité est le plus grand prédicteur de la confiance. Il est donc supposé que la fiabilité du trustee aura plus de d'influence que sa nature ou que les caractéristiques du trustor sur la confiance accordée. La fiabilité est donc mobilisée pour faire varier la confiance apprise et ainsi faire varier les arbitrages en faveur de l'aide automatisée ou humaine.

Hypothèse 1 : Dans une situation d'arbitrage, les individus choisissent l'aide présentant le plus fort taux de fiabilité.

Par ailleurs, l'hétérogénéité des résultats récents sur le biais d'automatisation et l'évolution rapide des technologies amènent à questionner dans quelle mesure la représentation des aides automatisées par les individus pourrait avoir évoluée en entraînant un affaiblissement du biais d'automatisation. Autrement dit, le biais d'automatisation demeure-t-il invariable, ou peut-il être conditionné par l'âge et l'expérience de l'opérateur ? Notamment, l'acquisition d'une meilleure représentation de la capacité d'adaptation de l'IA à des situations dynamiques et non prédéfinies pourrait jouer un rôle déterminant dans la manière dont les utilisateurs perçoivent et accordent leur confiance à ces systèmes ? Cela ouvrirait la voie à des formations spécifiques des individus visant à modifier leur représentation de l'IA (mieux appréhender ses forces mais aussi ses limites) en vue de limiter le biais d'automatisation. Deux autres hypothèses sont donc formulées. Elles reposent sur le consensus scientifique antérieur aux années 2000 et les résultats de Lyons et Stokes [LYO 12] ou Ezer et al. [EZE 08] qui indiquent que les individus expérimentés et plus âgés sont vulnérables à un biais d'automatisation (qui correspond à une surestimation de ses possibilités), alors que les plus jeunes le sont moins [PEA 16].

Hypothèse 2 : Dans une situation d'arbitrage et à fiabilité égale entre une aide automatisée basée sur l'IA et une aide humaine, les individus plus âgés et plus experts, choisissent plus souvent l'aide basée sur l'IA que l'aide humaine.

Hypothèse 3 : Dans une situation d'arbitrage et à fiabilité égale entre une aide basée sur l'IA et une aide humaine, les individus plus jeunes et moins experts, choisissent aussi souvent l'aide humaine que l'aide basée sur l'IA.

2. Méthode

Cette expérimentation s'inspire des protocoles développés par Lyons et al. [LYO 12] et Pearson et al. [PEA 16] [PEA 19], qui mobilisent le paradigme de la double aide à la décision en contexte militaire.

L’approche se décline dans un contexte aéronautique et plus précisément dans une mission de CAS (Close Air Support). Une mission CAS est une mission aérienne effectuée par un ou plusieurs aéronefs contre des cibles hostiles au sol et à proximité de forces amies, impliquant ainsi un niveau de risque important. Pour mener à bien cette mission, les pilotes peuvent être assistés par diverses aides. Celles-ci peuvent être soit automatisées, soit humaines comme le JTAC (Joint Terminal Attack Controller) : celui-ci est un collaborateur qui guide l'avion depuis le sol pour aider le pilote à identifier la cible ennemie. Le protocole présenté ici s’inspire d’une mission CAS dans laquelle le participant doit arbitrer entre deux sources d’informations contradictoires : un JTAC d’une part, et une aide automatisée constituée par un drone autonome doté d’IA, capable de transmettre des images aériennes, et dénommé *IA*.

Le protocole de Lyons et Stokes [LYO 12] manipule l'expertise des aides au travers de labels. Dans cette étude, il a été choisi de manipuler la perception des aides au travers d’un niveau de fiabilité en raison de son fort lien avec la confiance [CHA 16] [HAN 11] [KAP 21] [LEE 04] . De plus, la fiabilité a déjà été manipulée empiriquement dans le paradigme de la double aide à la décision [MER 15]. Les auteurs modifiaient la fiabilité des aides à la décision en amont, au cours de plusieurs interactions entre le participant et le système. Cette étude se différencie en choisissant d'appliquer une fiabilité labélisée, c'est-à-dire que la fiabilité est induite uniquement sous forme textuelle, pour faire varier exclusivement la confiance antérieure à l’interaction. Quatre niveaux de fiabilité (20%, 50%, 70%, 90%) sont choisis en s'inspirant des niveaux de fiabilité induits dans l'expérience de Merritt et al. [MER 15], soit 50% et 70% de fiabilité. Les autres valeurs ont été sélectionnées en partie pour créer une différence de fiabilité suffisamment marquée entre les quatre conditions expérimentales.

2.1. *Participants*

Cette étude comprend 2 groupes de participants :

- Le premier groupe (*élèves*) est constitué de 37 participants, tous des hommes, âgés de 20 à 28 ans (M = 23.8, ET = 2.28) et tous élèves pilotes militaires avec des connaissances théoriques des missions aériennes.
- Le second groupe (*experts*) est constitué de 37 participants, tous des hommes, âgés de 29 à 46 ans (M = 37.9, ET = 4.39) et tous pilotes ou navigateurs opérationnels en escadron de chasse et ayant une expérience réelle de la mission de Close Air Support.

2.2. *Matériel et stimuli*

Avant la mise en œuvre des expérimentations, des caractérisations de l’aide humaine JTAC et de l’aide automatisée basée sur l’IA ont été soumises à des pré-tests (questionnaires et entretiens) afin de valider une formulation des aides sans ambiguïté pour les deux groupes sociaux visés. Une fois cette formulation validée, celle-ci a été intégrée à la consigne :

« Cette expérience se déroule dans le cadre d’une mission CAS (Close Air Support). Durant l’expérimentation, vous pilotez un aéronef au sein du SCAF (Système de Combat Aérien du Futur). Vous êtes engagé dans une mission de type CAS qui est une action aérienne contre des cibles hostiles au sol et à proximité de forces amies. Afin de vous aider dans cette mission, deux agents (JTAC et IA) vous donnent la position de l’ennemi : Le JTAC (Joint Terminal Attack Controller) est un militaire, qui vous guide depuis le sol. L'IA est un drone autonome capable de vous transmettre l'image de la cible. Pour chaque contexte et pour chaque agent un indice de fiabilité est indiqué. Chaque agent vous fournit une image aérienne. L’objectif principal de cette mission est de traiter la cible ennemie à partir d’une de ces images. Laquelle choisissez-vous ? Vous cliquez sur le nom de l'agent (JTAC ou IA) et vous avez un temps limité pour prendre votre décision.

Tâche : Choisir l'image qui correspond selon vous à la cible »

La consigne était associée à une demande de consentement éclairé.

Lors de chaque essai, les participants collaborent à la fois avec une aide humaine (JTAC) et une aide basée sur l'IA (IA). Chaque aide se voit attribuée une des quatre valeurs de fiabilité retenues (20%, 50%, 70% et 90%). Les seize combinaisons possibles : 4 (fiabilité humaine) x 4 (fiabilité de l'aide basée sur l'IA), sont proposées de manière pseudo-aléatoires afin que chaque participant passe exactement quatre essais par combinaison, soit 64 essais par participant.

Une base de données de paires d'images est utilisée pour l'expérience. Cette base de données a été conçue spécifiquement pour ce protocole pour garantir que les deux images sont "équilibrées" entre elles, c'est-à-dire que chaque image a autant de probabilité d'être choisie que l'autre image.

Le délai imparti de prise de décision pour chaque essai est fixé à cinq secondes. Ce temps de réponse est volontairement court pour induire une pression temporelle forte. Cette durée a été déterminée lors de la phase de construction du matériel expérimental, à partir de la moyenne des temps de réponse des testeurs du protocole.

2.3. *Procédure*

Au cours de l'expérience, les participants sont placés dans le contexte d'une mission de Close Air Support. Dans cette expérience, le participant reçoit deux images de cibles possibles provenant de deux aides différentes, une aide humaine appelée "JTAC" et une aide à base d'IA appelée "IA". Les instructions de l'expérimentation sont présentées aux participants avant l'expérience.

Lors de chaque essai, la description de la cible et la fiabilité des deux aides sont d'abord présentées au participant sur une diapositive. La fiabilité de chaque aide est présentée sous forme textuelle (20% ,50%, 70%, 90%) et schématique à l'aide d'une jauge (*Cf.* Figure 3). Ensuite, sur la même diapositive, deux images satellites, fournies l'une par l'aide JTAC et l'autre par l'aide IA, sont affichées. Les deux images satellites sont conflictuelles, car elles n'affichent pas le même emplacement au sol alors qu'elles correspondent toutes deux à la description de la cible (*Cf.* Figure 3). La tâche du participant est de cliquer sur l'image qui, selon lui, correspond à la cible ennemie décrite dans un temps contraint. Si le participant ne parvient pas à répondre dans le temps donné, un essai dans les mêmes modalités de fiabilité lui est proposé après les 64 essais. Une fois le choix effectué le participant peut passer à l'essai suivant.

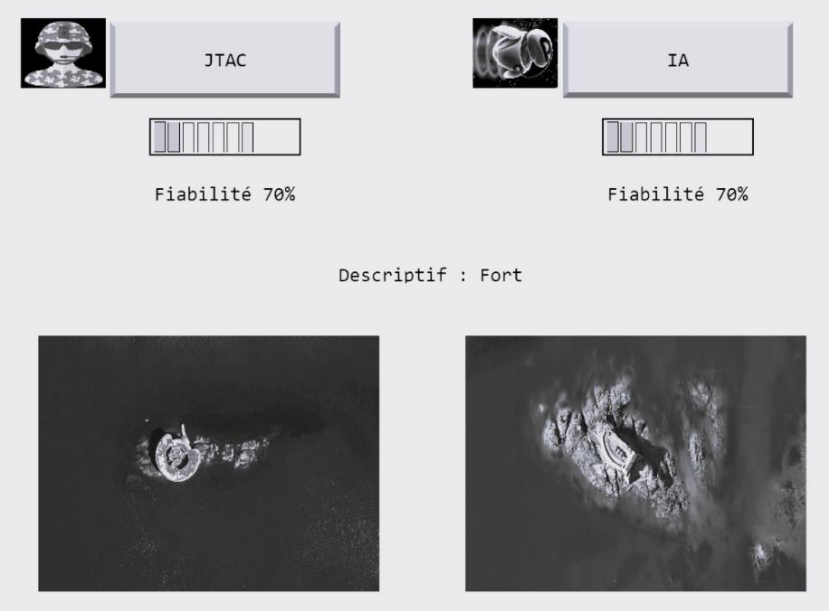

Figure 3. *Exemple d'un essai avec des fiabilités équivalentes de 70% entre l'aide humaine (JTAC) et l'aide à base d'IA (IA). Le participant doit cliquer, à l'aide de la souris, sur l'image correspondant pour lui à la cible*

2.4. *Mesures, variables et statistiques*

Cette expérimentation repose sur plusieurs variables indépendantes. Tout d'abord la catégorie d'appartenance des sujets, nommée $VI_{expertise}$ et qui peut prendre deux valeurs, élèves ou experts. La deuxième variable indépendante correspond à la fiabilité du JTAC, nommée VI_{Fiab_JTAC} et qui peut prendre quatre valeurs : 20%, 50%, 70% et 90%. La troisième est le pendant pour l'aide IA, nommée VI_{Fiab_IA} et qui peut prendre elle aussi les quatre mêmes valeurs.

Cependant, pour les analyses, les variables indépendantes VI_{Fiab_JTAC} et VI_{Fiab_IA} seront combinées pour créer des catégories de conditions particulières. Les trois conditions principales sont : (C_{JTAC}) quand le JTAC a une fiabilité plus forte que l'aide IA, (C_{IA}) quand c'est l'aide IA qui est plus fiable que le JTAC, et enfin ($C_{=}$) quand les deux aides ont la même fiabilité. Dans chacune de ces trois catégories, des sous catégories permettront d'analyser plus finement l'effet de la fiabilité. La Figure 4 explicite ces différentes catégories. La catégorie $C_{=}$ est composée de quatre sous-catégories, $C_{=20}$, $C_{=50}$, $C_{=70}$ et $C_{=90}$, chacune associée à une des valeurs possibles pour les fiabilités. Les deux autres catégories sont divisées en trois, selon la différence de fiabilité entre les deux aides : Faible si la différence est égale à 20% (90%-70% et 70%-50%) ; Moyenne si elle est entre 30% et 40% (50%-20% et 90%-50%) ; Forte si elle est supérieure à 50% (70% - 20% et 90%-20%).

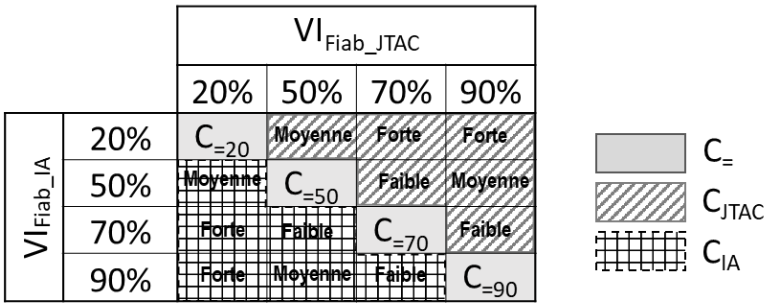

Figure 4. *Matrice des catégories de variables en fonctions du niveau de fiabilité associé à l'aide. En gris la catégorie de fiabilité égale C= et ses quatre sous-catégories. Hachurée, la catégorie CJTAC des essais correspondants à une fiabilité du JTAC plus forte que celle de l'IA et les trois sous-catégories (Faible, Moyenne, Forte). Quadrillée, la catégorie CIA des essais correspondant à une fiabilité de l'IA plus forte que celle du JTAC et les trois sous-catégories (Faible, Moyenne, Forte)*

Chacun des 64 essais réalisés par le participant produit une mesure catégorielle unique correspondant au choix de l'image qu'il a réalisé : soit celle proposée par le JTAC, soit celle proposée par l'aide IA.

Pour chaque participant, nous utilisons ces mesures catégorielles pour calculer la proportion de réponses en faveur du JTAC au sein d'une condition expérimentale. Cette variable dépendante sera donc, pour les essais de la condition expérimentale, le nombre de réponse du participant en faveur du JTAC sur le nombre total de réponse du participant pour cette condition expérimentale. Le nom de la variable dépendante sera indicé par la condition expérimentale.

Par exemple $VD_{C=}$ sera calculé pour chaque sujet en considérant uniquement les 16 essais de la condition expérimentale $C_{=}$ (4 essais pour chaque case de la diagonale, *Cf.* Figure 4) :

$$VD_{C=}(Participant\ n) = \frac{Nombre\ d'image\ du\ JTAC\ choisies\ par\ le\ Participant\ n\ lors\ des\ 16\ essais\ de\ C_{=}}{16}$$

Les tests statistiques sont réalisés en Python avec les bibliothèques Pingouin et Scipy. Le seuil de significativité retenu est 0.05. Pour les ANOVA, le test de sphéricité de Maulchy est utilisé pour déterminer si les valeurs de p doivent être corrigées. Si c'est le cas, la correction de Greenhouse-Geisser

est appliquée et la valeur corrigée est prise en compte pour l'analyse. Les tests pos-hoc utilisent des t-tests.

Pour les graphiques, les distributions sont représentées par des boites à moustache pour lesquelles la boite représente les quartiles de la distribution (1er en bas, 2ème=médiane, 3ème en haut), les cercles représentent les valeurs aberrantes, le triangle la moyenne, les moustaches représentent 5% et 95% de la distribution. Les seuils de significativité sont représentés avec la convention suivante : ns : $p>0.05$; * : $p<0.05$; ** : $p<0.01$; *** : $p<0.001$.

3. Résultats

Afin de tester la première hypothèse, qui stipule que dans une situation d'arbitrage, les individus choisissent l'aide présentant le plus fort taux de fiabilité, les 2 conditions expérimentales C_{JTAC} et C_{IA} sont analysées, de manière séparée.

Dans la condition expérimentale C_{JTAC} où le JTAC a une fiabilité plus forte que l'aide IA un t-test à un échantillon est réalisé pour chacune des populations (élèves et experts) afin de vérifier si la proportion de réponse en faveur du JTAC (VD_{CJTAC}) est significativement supérieure à 0.5 (0.5 correspondant au fait de choisir de manière égale les images proposées par le JTAC ou par l'aide IA). Pour les élèves, le test confirme un choix préférentiel pour les images proposées par le JTAC ($M=0.76$; $t(36)=4.78$, $p<0.001$). Un résultat équivalent et même un peu plus accentué est obtenu pour les experts ($M=0.89$; $t(36)=16,89$, $p<0.001$). La Figure 5 montre les deux distributions.

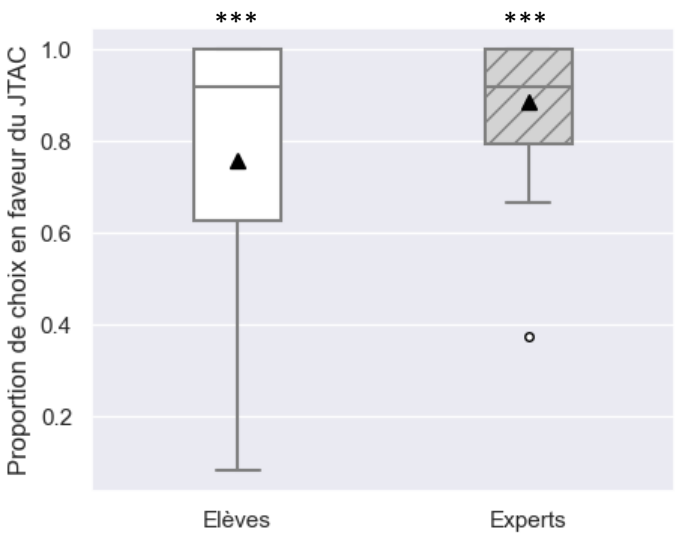

Figure 5. *Distributions des proportions de choix en faveur du JTAC quand la fiabilité du JTAC et plus forte que celle de l'IA, par population. Chaque distribution est significativement différente du choix aléatoire (0.5).*

Afin de vérifier si les deux populations (élèves et expert) sont comparables et si la différence de fiabilité entre les deux aides (Faible, Moyenne, ou Forte) a un impact sur la proportion de choix en faveur du JTAC pour chacune d'elles, une ANOVA mixte est réalisée (facteur externe : la population ; facteur interne : la différence de fiabilité entre les aides). Les résultats (*Cf.* Table 1) montrent qu'il n'y a pas d'effet d'interaction ($p=0.53$) et que la différence de fiabilité entre les deux aides ne conduit donc pas à des évolutions différentes de la proportion de choix en faveur du JTAC entre les deux populations. Il n'y a pas d'effet non plus de la différence de fiabilité sur la proportion de choix en faveur du JTAC (p-GG-corr=0.78 ; les deux populations confondues). En revanche il y a bien une différence entre les deux populations ($p=0.03$), les experts choisissant de manière plus significative les images proposées par le JTAC que ne le font les élèves.

Source	SS	DF1	DF2	MS	F	p-unc	p-GG-corr	n2	eps	spehricity	W-spher	p-spher
Population	0.93	1	72	0.93	4.91	**0.0299**		0.06				
Fiabilité	0.01	2	144	0.00	0.20	0.8152	**0.7797**	0,00	0.85	False	0.83	0.00
Interaction	0.02	2	144	0.01	0.64	**0.5281**		0.00				

Table 1. *Résultats de l'ANOVA mixte visant à vérifier s'il existe un effet d'interaction entre les deux populations (facteur 1) et la différence de fiabilité entre JTAC et IA (facteur 2) sur leur proportion de choix en faveur du JTAC, dans les conditions où le JTAC est plus fiable que l'IA.*

L'analyse de la catégorie C_{JTAC} va donc dans le sens de l'hypothèse 1 : dans une situation d'arbitrage, les individus choisissent l'aide présentant le plus fort taux de fiabilité. L'analyse montre que le résultat n'est pas lié dans cette expérience à la force de cette différence de fiabilité mais est modulé par la population étudiée.

Une analyse similaire est conduite dans la condition CIA dans laquelle c'est l'aide IA qui a une fiabilité plus grande que le JTAC. Pour les élèves, le test confirme un choix préférentiel pour les images proposées par l'aide IA, celles proposées par le JTAC étant peu choisies ($M=0.37$; $t(36)=-2.08$, $p<0.05$). Un résultat équivalent et même plus accentué est obtenu pour les experts qui choisissent encore moins l'aide du JTAC ($M=0.28$; $t(36)=-4,39$, $p<0.001$). La Figure 6 montre les deux distributions.

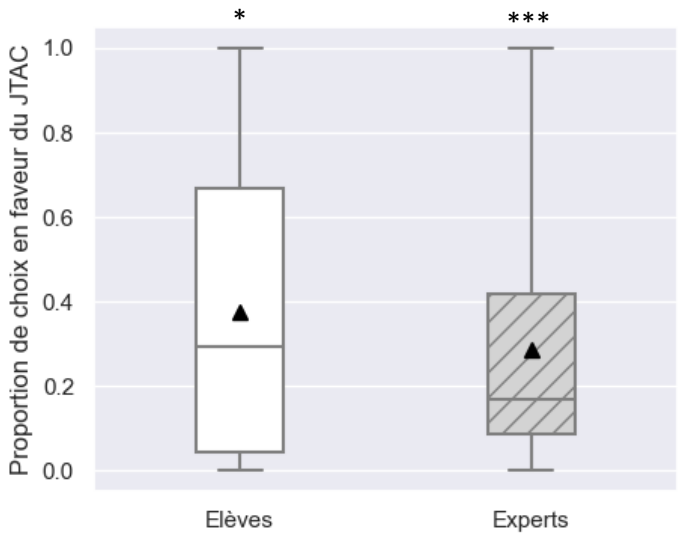

Figure 6. *Distributions des proportions de choix en faveur du JTAC quand la fiabilité de l'IA et plus forte que celle du JTAC, par population. Chaque distribution est significativement différente du choix aléatoire (0.5).*

Comme précédemment, une ANOVA mixte est réalisée pour analyser si la différence de fiabilité agit de la même manière sur les deux populations. Les résultats montrent qu'il n'y a pas d'effet d'interaction ($p=0.53$). En revanche la différence de fiabilité entre les aides a bien un impact sur la proportion de choix en faveur de l'image du JTAC ($p\text{-GG-corr}<0.001$).

Source	SS	DF1	DF2	MS	F	p-unc	p-GG-corr	n2	eps	spehricity	W-spher	p-spher
Population	0.45	1	72	0.45	1.33	**0.2519**		0.02				
Fiabilité	0.49	2	144	0.25	11.27	0.00	**0.0001**	0,02	0.79	False	0.73	0.00
Interaction	0.03	2	144	0.01	0.63	**0.5326**		0.00				

Table 2. *Résultats de l'ANOVA mixte visant à vérifier s'il existe un effet d'interaction entre les deux populations (facteur 1 : élèves, experts) et la différence de fiabilité entre IA et JTAC (facteur 2 : faible, moyenne, forte) sur leur proportion de choix en faveur du JTAC, dans les conditions où l'IA est plus fiable que le JTAC.*

Les tests post-hoc montrent une différence significative (les deux populations confondues) entre les catégories correspondant à une différence de fiabilité faible et moyenne ($p<0.001$) et entre faible et forte ($p<0.001$) (*Cf.* Table 2 & Figure 7). En revanche, les deux populations ne sont pas différentes entre elles ($p=0.25$) et choisissent donc dans des proportions semblables les images du JTAC dans ce cas-là.

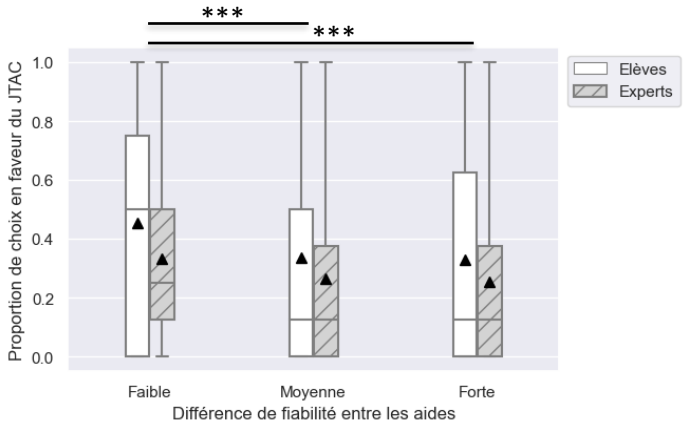

Figure 7. *Impact de la différence de fiabilité entre JTAC et IA sur la proportion de choix en faveur du JTAC dans la condition où l'IA est la plus fiable. Si les deux populations ne diffèrent pas entre elles, une différence de fiabilité forte (l'IA étant fortement plus fiable que le JTAC) amène une proportion de choix en faveur du JTAC significativement plus faible que quand la différence de fiabilité est plus réduite.*

L'analyse de la catégorie C_{IA} va donc elle aussi dans le sens de l'hypothèse 1 : dans une situation d'arbitrage, les individus choisissent l'aide présentant le plus fort taux de fiabilité. L'analyse montre que le résultat est ici modulé par la force de cette différence de fiabilité mais n'est en revanche pas sensible à la population étudiée.

L'analyse séparée des catégories C_{JTAC} et C_{IA} montrent un même effet global, à savoir le choix préférentiel pour l'aide la plus fiable mais des modulations différentes en fonction de la population ou de la différence de fiabilité entre les aides. Les résultats ne semblant pas symétriques, il est recherché si le type d'aide qui a la fiabilité la plus forte modifie la proportion de choix en fonction de cette aide. La variable considéré pour cette analyse n'est donc pas la proportion de choix en faveur du JTAC, mais la proportion de choix en fonction de l'aide la plus fiable. Une ANOVA à deux facteurs est réalisée pour étudier l'effet combiné de la population (facteur 1 : élèves ou experts) et du type d'aide qui a la fiabilité la plus forte (facteur 2 : JTAC ou IA). Les résultats (*Cf.* Table 3) montrent qu'il n'y a pas d'effet d'interaction ($p=0.69$) mais qu'il y a à la fois un effet de la population ($p<0.05$) et de l'aide qui à la plus

forte fiabilité ($p<0.01$). Les tests post-hoc valident la différence entre les élèves et les experts ($p<0.05$) et du type d'aide qui a la plus forte fiabilité ($p<0.01$).

Source	SS	DF	MS	F	p-unc	n2
Population	0.44	1	0.44	5.06	**0.0260**	0.0319
Aide	0.84	1	0.84	9.56	**0.0024**	0.0602
Population*Aide	0.01	1	0.01	0.16	**0.6853**	0.0010
Residual	12.65	144	0.09			

Table 3. *Résultats de l'ANOVA à deux facteurs (Population : élèves, experts ; et Type d'aide qui est la plus fiable : IA, JTAC) sur la proportion de choix en faveur de l'aide la plus fiable. S'il n'y a pas d'effet d'interaction, il y a bien un effet principal des deux facteurs.*

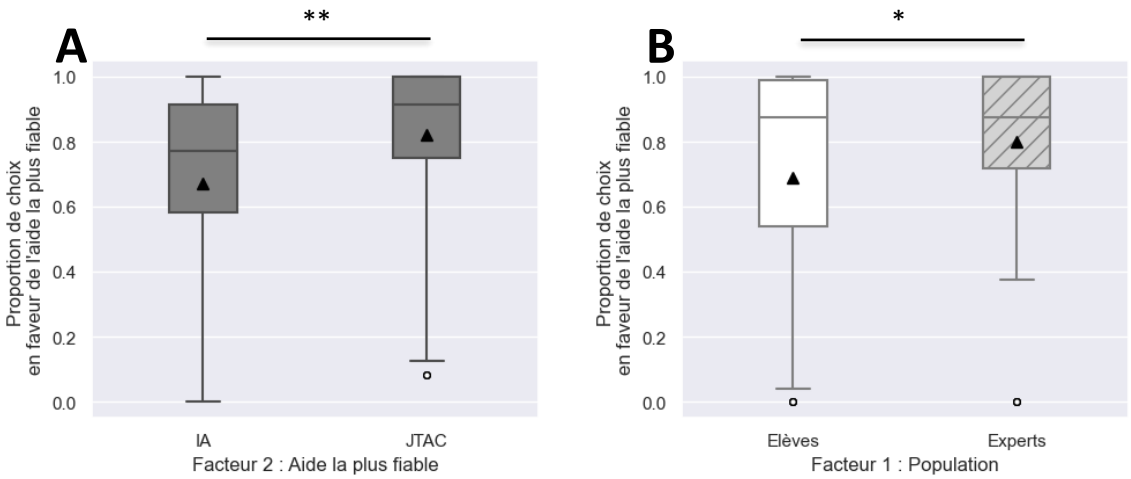

Figure 8. *Distribution des proportions de choix en faveur de l'aide la plus fiable selon l'aide la plus fiable (A) et selon la population (B).*

Cette analyse combinée des deux catégories confirme que la force du biais en faveur de l'aide la plus fiable dépend de la nature de l'aide la plus fiable : quand c'est le JTAC qui est le plus fiable, le biais est plus important (Cf. Figure 8, A). Mais la force du biais dépend aussi de la population étudiée, les experts étant plus enclins à suivre ce biais de fiabilité (*Cf.* Figure 8, B), quelle que soit l'aide la plus fiable.

Les hypothèses suivantes se focalisent sur le cas où les deux aides ont la même fiabilité et donc sur la catégorie $C_{=}$.

La deuxième hypothèse se concentre sur la population experte et stipule que dans une situation d'arbitrage et à fiabilité égale entre une aide automatisée basée sur l'IA et une aide humaine, les individus les plus âgés et les plus experts, choisissent plus souvent l'aide basée sur l'IA que l'aide humaine. La vérification de cette hypothèse repose sur les données de la condition $C_{=}$ pour laquelle les fiabilités des deux aides sont équivalentes. Un t-test à un échantillon est donc réalisé pour la population experte afin de vérifier si la proportion de réponse en faveur du JTAC ($VD_{C=}$) est significativement supérieure à 0.5 (0.5 correspondant au fait de choisir de manière équitable les images proposées par le JTAC ou par l'aide IA). Le test confirme un choix préférentiel pour les images proposées par le JTAC ($M=0.71$; $t(36)=5.96$, $p<0.001$) alors que les deux aides ont la même fiabilité. La figure 9 (*Cf.* partie B) montre la distribution des données.

La fiabilité équivalente des deux aides pouvant aller de 20% à 90% dans cette expérimentation, une ANOVA à mesure répétée est effectuée pour analyser si la fiabilité des aides modifie significativement la force de ce choix préférentiel pour les images du JTAC. Le résultat montre qu'il n'y a pas d'effet ($F(3,108)=0.64$, $p=0.59$).

La troisième hypothèse se concentre sur la population des élèves et prévoit que dans une situation d'arbitrage et à fiabilité égale entre une aide basée sur l'IA et une aide humaine, les individus ayant commencé à interagir avec des aides automatisées après 2000, choisissent aussi souvent l'aide humaine que l'aide basée sur l'IA. Un t-test à un échantillon est donc réalisé pour la population des élèves afin de vérifier si la proportion de réponse en faveur du JTAC ($VD_{C=}$) est significativement supérieure à 0.5 (0.5 correspondant au fait de choisir de manière équitable les images proposées par le JTAC ou par l'aide IA). Le test ($M=0.60$; $t(36)=1.89$, $p=0.07$) indique une tendance pour un choix préférentiel pour les images proposées par le JTAC quand les deux aides ont la même fiabilité. La Figure 9 (partie A) montre la distribution des données.

La fiabilité équivalente des deux aides pouvant aller de 20% à 90% dans cette expérimentation, une ANOVA à mesure répétée est effectuée pour analyser si la fiabilité des aides modifie significativement la force de ce choix préférentiel pour les images du JTAC. Le résultat montre qu'il n'y a pas d'effet ($F(3,108)=0.31$, $p=0.82$).

Il est alors nécessaire de se demander si ce comportement qui amène à choisir préférentiellement l'aide proposé par le JTAC est réellement différent entre les deux populations. Pour cela un t-test est réalisé entre les deux populations ($t(72)=1.78$, $p=0.08$) qui indique une tendance (*Cf.* Figure 9). Si les deux populations ne sont pas significativement différentes, il est bien observé une tendance des experts à choisir plus souvent l'aide du JTAC que les élèves.

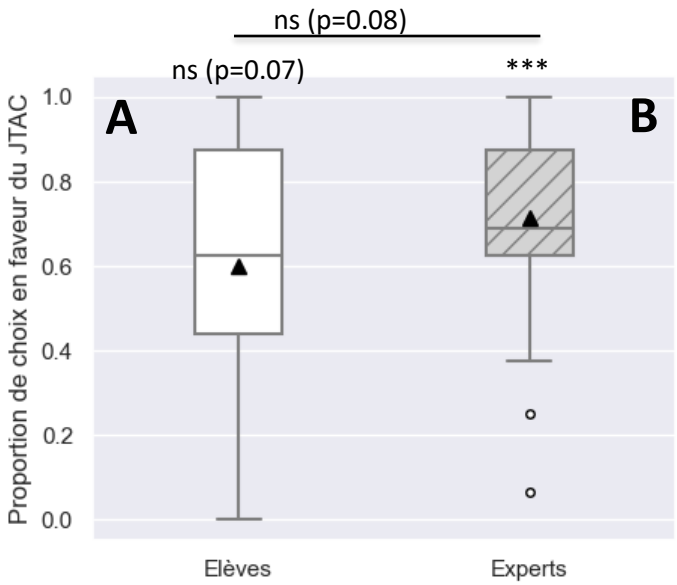

Figure 9. *Distributions des proportions de choix en faveur du JTAC quand la fiabilité des deux aides est égale, par population. La distribution B est significativement différente du choix aléatoire (0.5) mais pas la A (même si on observe une tendance p<0.1). Cependant les deux distributions ne sont pas statistiquement différentes entre elles bien qu'une tendance se dessine (p<0.1).*

Globalement, l'analyse de la catégorie $C_=$ pour laquelle les deux aides ont la même fiabilité montre un biais en faveur du JTAC pour la population des experts, allant donc à l'encontre de l'hypothèse 2. Cette préférence pour l'humain n'est dépendante de la fiabilité (équivalente) des deux aides. Ce résultat est beaucoup plus atténué pour la population des élèves mais ne confirme pas réellement l'hypothèse 3, une tendance à préférer l'aide humaine étant observée. Là aussi il n'y a pas d'effet du niveau de fiabilité des deux aides.

4. Discussion

Assez largement étudié dans la littérature, le biais d'automatisation a été éprouvé selon une approche expérimentale différentielle reposant sur un arbitrage entre deux d'aides informationnelles : une aide humaine, et une aide basée sur l'IA. Cette situation expérimentale a été conduite avec deux groupes de participants aux caractéristiques psychosociales différentes : un groupe de jeunes élèves officiers de moins de 29 ans, et un groupe de personnels navigants experts et âgés de plus de 29 ans.

Les résultats obtenus pour l'hypothèse 1 indiquent que pour les deux groupes de participants, l'aide présentant la fiabilité la plus importante est choisie. Ces résultats vont dans le sens de la littérature [MER 15] indiquant tout d'abord la bonne prise en compte par les participants des indices de fiabilité présentés. Ce premier résultat confirme que la fiabilité des aides est l'élément principal qui guide le choix lors d'un arbitrage entre deux aides. Cependant l'analyse plus détaillé montre une différence entre les deux populations : quand l'aide humaine est la plus fiable, les experts s'appuient encore plus sur cette aide humaine que les élèves. L'analyse de la distribution (*Cf.* Figure 4) indique que la population des experts est plus homogène dans sa réponse. La formation en escadron et la pratique réelle de ce type de mission pourraient donc être des éléments qui modifient le rapport des opérateurs aux aides disponibles et favorisent une standardisation du comportement. Par ailleurs, les résultats ont aussi mis en évidence une dissymétrie entre le cas où c'est l'IA qui est la plus fiable et le cas où c'est le JTAC qui est le plus fiable (*Cf.* Figure 7 A). L'aide la plus fiable est d'autant plus choisie que c'est l'aide humaine, indiquant une sorte de biais en faveur de l'aide humaine. Ceci confirme que la nature de l'aide est bien prise en compte par les participants, indiquant la validité du dispositif expérimental. Si la fiabilité des aides constitue bien le critère primordial qui guide le choix, les résultats montrent que la nature de l'aide (confiance apprise) et la population (confiance dispositionnelle) viennent tout de même moduler l'effet de ce premier critère. Ils prennent donc toute leur importance lorsque le critère de fiabilité des aides n'est plus discriminant, comme testé pour les hypothèses 2 et 3.

Si l'expérience confirme l'hypothèse 1, en revanche, les résultats vont à l'encontre de l'hypothèse 2 : ils montrent que dans une situation d'arbitrage avec une fiabilité équivalente indiquée pour chaque aide, les personnels navigants expérimentés préfèrent choisir l'aide humaine plutôt que l'aide basée sur l'IA. Cela indique que le biais d'automatisation ne se manifeste pas dans ce type de contexte, mais qu'il s'agit plutôt d'un biais de décision en faveur de l'aide humaine, qui renforce la préférence pour l'aide humaine déjà signalée au paragraphe précédent.

Parallèlement, les résultats liés à l'hypothèse 3 sont plus nuancés, c'est à dire que les participants plus jeunes préfèrent l'aide humaine mais de façon moins marquée que les experts. Autrement dit, en accord avec Pearson et al. [PEA 16], dans la mesure où la fiabilité est équivalente, les jeunes sont moins biaisés dans leur arbitrage entre une aide humaine et une IA.

En ce sens, il y a bien une différence entre les deux groupes sociaux, c'est-à-dire un impact de la confiance dispositionnelle sur le choix entre une aide humaine et une aide basée sur l'IA. Ces différences peuvent être effectivement dues à l'âge et à l'expérience avec les aides automatisées et avec un JTAC, mais elles peuvent également être dues à une appréhension du risque différente liée à la connaissance des missions de CAS ayant inspiré le protocole expérimental (confiance situationnelle). En effet, l'expérimentation menée ici induit une situation avec un niveau de risque élevé, compte tenu des enjeux liés à la mission. Or, il semble que plus une situation est risquée, moins l'individu attribue de confiance en l'automatisation [PEA 16] [PER 10] [RAJ08] [SAT 17]. Ainsi, l'absence d'observation du biais d'automatisation pourrait également être liée au niveau de risque élevé, cela signifiant que le risque influencerait considérablement la prise de décision dans les contextes de confiance dans l'automatisation. Les travaux de Liehner et al. [LIE 21] montrent que la perception du risque peut se subdiviser en plusieurs catégories, telles que les risques physiques, psychologiques et moraux, chacune ayant des implications distinctes sur la confiance. Il apparaît donc nécessaire de reproduire cette expérience avec un risque moral moins important par exemple pour comparer les résultats et mieux comprendre l'impact

potentiel du niveau de risque sur le biais d'automatisation, tout en intégrant une réflexion sur les implications des différents types de risque sur la confiance.

Par ailleurs, aucun des deux groupes n'a eu de comportement attendu pour un biais d'automatisation, c'est à dire une préférence marquée pour l'aide basée sur l'IA. Il semble donc pertinent de reconsidérer les conclusions tirées des études passées sur le biais d'automatisation dans les contextes technologiques contemporains basés sur des aides multiples à la décision. Ceci invite aussi à revisiter le paradigme initial, dans lequel l'opérateur garde à sa charge ou délègue une tâche, pour des contextes d'aide plus sophistiquées. Les prochaines études devront viser à ségréguer d'avantage les différentes dimensions de la confiance du modèle avant d'obtenir des résultats plus exploitables. Ainsi, elles pourraient par exemple faire varier la nature de l'aide automatisée (aide automatisée basique vs aide basée sur l'IA) pour tester spécifiquement la confiance apprise. Une telle approche permettrait de tester de manière plus précise le comportement des individus face à des systèmes proposant des aides automatisées plus ou moins évoluées, et d'apporter une compréhension plus approfondie des situations d'arbitrage et du biais d'automatisation.

Enfin, il convient de rappeler que cette étude se concentre sur la confiance *a priori*, ne visant pas l'étude de la dynamique de la confiance ; analyser comment la confiance évolue au fil de la collaboration avec une aide automatisée sur des périodes prolongées permettrait de mieux appréhender les interactions homme-système et le biais d'automatisation comme le propose Glikson et al. [GLI 20].

5. Conclusion & perspectives

En conclusion, les résultats de cette recherche indiquent l'absence de biais d'automatisation dans ce cadre expérimental. Ils suggèrent que dans les contextes socio-complexes contemporains où la gradation des capacités des différents systèmes automatisés varie continuellement, des automatismes primaires aux IA, la prégnance de ce biais tend à s'atténuer. Il semble donc que ce biais n'est pas invariant, mais plutôt influencé par des facteurs identifiés dans le modèle. Celui-ci illustre l'impact des paramètres environnementaux, des caractéristiques de l'aide automatisée et des variations inter-individuelles sur la confiance envers l'automatisation ; en se référant à ce modèle, cette étude démontre expérimentalement que les individus manifestent des niveaux de confiance et des comportements distincts en fonction de leurs caractéristiques psychosociales, autrement dit de leur confiance dispositionnelle. Autrement dit, la prise en compte d'éventuels biais d'automatisation, ou « d'humanisation » comme c'est le cas ici, doit être réalisée au regard des caractéristiques psychosociales des utilisateurs. En matière d'équipement, la confiance dispositionnelle des individus doit être mesurée afin de ne pas leur proposer d'interfaces favorisant un biais. Par ailleurs, les perspectives soulevées par ces résultats pourraient influencer les protocoles de formation des opérateurs, en les préparant à interagir de manière plus efficace avec l'automatisation en tenant compte des variations des risques environnementaux et des types d'automatisation.

Enfin cette étude pointe la nécessité d'être attentif aux caractéristiques psychosociales des participants des études menées sur le biais d'automatisation et sur la confiance plus globalement : leur profil doit être en adéquation avec celui des utilisateurs finaux lorsque ceux-ci sont connus. Plus encore, il devient nécessaire de considérer différemment les résultats d'études similaires menées avec des groupes de participants différents (e.g. Pearson et al. [PEA 16] *vs* Lyons et Stokes [LYO 12]), afin de dissiper une confusion contraire à la compréhension du champ.

Keypoint

- En situation d'arbitrage entre une aide humaine et une aide basée sur l'IA, les individus choisissent l'aide présentant la fiabilité la plus importante.
- En situation d'arbitrage entre une aide humaine et une aide basée sur l'IA, les individus choisissent plus souvent l'aide humaine.

- Cette préférence est plus marquée chez les participants expérimentés et ayant été confrontés à l'aide automatisée avant 2000.
- L'ensemble des résultats montrent une absence de biais d'automatisation.

Bibliographie

[CHA 16] CHAVAILLAZ A., WASTELL D., & SAUER J., "System reliability, performance and trust in adaptable automation", *Applied Ergonomics*, 52, p. 333–342, 2016.

[DZI 01] DZINDOLET M. T., PIERCE L. G., BECK H. P., DAWE L. A., & ANDERSON B. W., " Predicting misuse and disuse of combat identification systems", *Military Psychology*, *13*(3), p. 147-164, 2001.

[DZI 02] DZINDOLET M. T., PIERCE L. G., BECK H. P., & DAWE L. A., "The perceived utility of human and automated aids in a visual detection task", *Human Factors*, 44(1), p. 79–94, 2002.

[EZE 08] EZER N., FISK A. D., & ROGERS W. A., "Age-related differences in reliance behavior attributable to costs within a human-decision aid system", *Human Factors*, 50(6), p. 853–863, 2008.

[FRA 03] FRANTZ R., "Herbert Simon. Artificial intelligence as a framework for understanding intuition", *Journal of Economic Psychology*, *24*(2), p. 265–277, 2003.

[GLI 20] GLIKSON E., & WOOLLEY A. W., "Human trust in artificial intelligence: Review of empirical research", *Academy of Management Annals*, *14*(2), p. 627–660, 2020.

[HAN 11] HANCOCK P. A., BILLINGS D. R., SCHAEFER K. E., CHEN J. Y., DE VISSER E. J., & PARASURAMAN R., "A meta-analysis of factors affecting trust in human-robot interaction", *Human factors*, *53*(5), p. 517-527, 2011.

[HOF 15] HOFF K. A., & BASHIR M., " Trust in automation: Integrating empirical evidence on factors that influence trust", *Human Factors*, *57*(3), p. 407–434, 2015.

[KAP 21] KAPLAN A. D., KESSLER T. T., BRILL J. C., & HANCOCK P. A., " Trust in artificial intelligence: Meta-analytic findings", *Human factors*, *65*(2), p. 337-359, 2021.

[LAY 94] LAYTON C., SMITH P. J., & MC COY C. E., "Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation", *Human Factors*, *36*(1), p. 94–119, 1994.

[LEE 04] LEE J. D., & SEE K. A., "Trust in automation: Designing for appropriate reliance", *Human Factors*, *46*(1), p. 50–80. 2004.

[LER 97] LERCH F. J., PRIETULA M. J., & KULIK C. T., "The Turing effect: The nature of trust in expert systems advice", dans P. FELTOVICH, K. FORD (dir), *Expertise in context: Human and machine*, MIT Press, Cambridge, MA, USA, p. 417-448, 1997.

[LIE 21] LIEHNER G. L., BRAUNER P., SCHAAR A. K., & ZIEFLE M., " Delegation of moral tasks to automated agents—the impact of risk and context on trusting a machine to perform a task", *IEEE Transactions on Technology and Society*, *3*(1), p. 46-57, 2021.

[LYO 12] LYONS J. B., & STOKES C. K., "Human-human reliance in the context of automation", *Human Factors*, *54*(1), p. 112–121, 2012.

[MAY 95] MAYER R. C., DAVIS J. H. & SCHOORMAN F. D., " *An Integrative Model of Organizational Trust", Academy of Management Review*, 1995.

[MCG 06] MCGUIRL J. M., & SARTER N. B., " Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information", *Human Factors*, *48*(4), p. 656–665, 2006.

[MER 15] MERRITT S. M., SINHA R., CURRAN P. G., & ILGEN D. R., "Attitudinal predictors of relative reliance on human vs. automated advisors", *International Journal of Human Factors and Ergonomics*, *3*(3-4), p. 327-345, 2015.

[MOS 92] MOSIER K. L., PALMER E. A., & DEGANI A., " Electronic checklists: Implications for decision making", dans *Proceedings of the Human Factors Society Annual Meeting*, *36*(1), p. 7-11, Sage Publications, Los Angeles, CA, USA, 1992.

[MOS 98] MOSIER K. L., SKITKA L. J., HEERS S., & BURDICK M., "Automation bias: Decision making and performance in high-tech cockpits", *Decision Making in Aviation*, p. 271–287, 1998.

[MOS 01] MOSIER K. L., SKITKA L. J., DUNBAR M., & MCDONNELL L., « Aircrews and automation bias: The advantages of teamwork?", *International Journal of Aviation Psychology*, *11*(1), p. 1–14, 2001.

[PAR 97] PARASURAMAN R., & RILEY V., “Humans and automation: Use, misuse, disuse, abuse”, *Human factors*, *39*(2), p. 230-253, 1997.

[PAR 10] PARASURAMAN R., & MANZEY D. H., “Complacency and bias in human use of automation: An attentional integration”, *Human Factors*, *52*(3), p. 381–410, 2010.

[PEA 16] PEARSON C. J., WELK A. K., BOETTCHER W. A., MAYER R. C., STRECK S., SIMONS-RUDOLPH J. M., & MAYHORN C. B., “Differences in trust between human and automated decision aids”, dans *Proceedings of the Symposium and Bootcamp on the Science of Security*, p. 95–98, 2016.

[PEA 19] PEARSON C. J., GEDEN M., & MAYHORN C. B., “Who’s the real expert here? Pedigree’s unique bias on trust between human and automated advisers”, *Applied Ergonomics*, *81*, p. 102907, 2019.

[PER 10] PERKINS L. A., MILLER J. E., HASHEMI A., & BURNS G., “ Designing for human-centered systems: Situational risk as a factor of trust in automation”, dans *Proceedings of the Human Factors and Ergonomics Society*, *3*, p.2130–2134, 2010.

[RAH 19] RAHWAN I., CEBRIAN M., OBRADOVICH N., BONGARD J., BONNEFON J. F., BREAZEAL C., CRANDALL J. W., CHRISTAKIS N. A., COUZIN I. D., JACKSON M. O., JENNINGS N. R., KAMAR E., KLOUMANN I. M., LAROCHELLE H., LAZER D., MCELREATH R., MISLOVE A., PARKES D. C., PENTLAND A., … WELLMAN M., “Machine behaviour. Machine Learning and the City: Applications in Architecture and Urban Design*”*, *Nature*, *568*, p. 143–166, 2010.

[RAJ08] RAJAONAH B., TRICOT N., ANCEAUX F., & MILLOT P., « The role of intervening variables in driver-ACC cooperation”, *International Journal of Human Computer Studies*, *66*(3), p. 185–197, 2008.

[REM 85] REMPEL J. K., HOLMES J. G., & ZANNA M. P., “Trust in Close Relationships”, *Journal of Personality and Social Psychology*, *49*(1), p. 95–112, 1985.

[ROV 07] ROVIRA E., MCGARRY K., & PARASURAMAN R., “Effects of imperfect automation on decision making in a simulated command and control task”, *Human Factors*, *49*(1), p. 76–87, 2007.

[SAR 00] SARTER N. B., & ALEXANDER H. M., « Error types and related error detection mechanisms in the aviation domain: An analysis of aviation safety reporting system incident reports”, *The international journal of aviation psychology*, *10*(2), p. 189-206, 2000.

[SAR 01] SARTER N. B., & SCHROEDER B., “Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing”, *Human Factors*, *43*(4), p. 573–583, 2001.

[SAT 17] SATTERFIELD K., BALDWIN C., DE VISSER E., & SHAW T., “The influence of risky conditions in trust in autonomous systems”, dans *Proceedings of the Human Factors and Ergonomics Society*, p. 324–328, 2017.

[SHE 02] SHERIDAN T. B., “ *Humans and automation: System design and research issues”*, Wiley en cooperation avec the Human Factors and Ergonomics Society, Santa Monica, CA, USA, 2002.

[SKI 00] SKITKA L. J., MOSIER K. L., BURDICK M., & ROSENBLATT B., “Automation bias and errors: Are crews better than individuals?”, *International Journal of Aviation Psychology*, *10*(1), p. 85–97, 2000.

[WES 05] WESTBROOK J. I., COIERA E. W., & GOSLING A. S., “Do online information retrieval systems help experienced clinicians answer clinical questions?”, *Journal of the American Medical Informatics Association*, *12*(3), p. 315–321, 2005.

[WOH 16] WOHLEBER R. W., CALHOUN G. L., FUNKE G. J., RUFF H., CHIU C. Y. P., LIN J., & MATTHEWS G., “The impact of automation reliability and operator fatigue on performance and reliance, dans *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), p. 211-215, Sage Publications, Los Angeles, CA, USA, 2016.

[YIG 22] YIGITCANLAR T., DEGIRMENCI K., BUTLER L., & DESOUZA K. C., “What are the key factors affecting smart city transformation readiness? Evidence from Australian cities”, *Cities*, *120*, p. 103434, 2022.

[ZHA 21] ZHANG L., PENTINA I., & FAN Y., “Who do you choose? Comparing perceptions of human vs robo-advisor in the context of financial services”, *Journal of Services Marketing*, *35*(5), p. 634–646, 2021.

