# Effets d'ancrage liés à la pression temporelle et à la difficulté des tâches dans un contexte de décision humaine assistée par IA

Jules Leguy[1], Nicolas Soulié[2], Felipe Torres Figueroa[2], Andon Tchechmedjiev[1], Jacky Montmain[1], Sébastien Harispe[1]

[1] SyCoIA, IMT Mines Ales, Ales, France
[2] Université Paris-Saclay, Univ Evry, IMT-BS, LITEM, 91025, Evry-Courcouronnes, France

jules.leguy@mines-ales.fr

## Résumé

*Les systèmes de décision assistée par IA se multiplient dans les domaines à forts enjeux. Une calibration adéquate de la confiance et de la reliance (inclination à suivre les prédictions de l'IA) des opérateurs humains envers ces systèmes est nécessaire pour en permettre un usage approprié. Cette calibration peut être affectée par la pression temporelle et la difficulté des tâches, ainsi que par d'éventuels effets d'ancrage engendrés par les premières interactions avec le système. Nous menons une étude permettant d'examiner l'impact des conditions d'exposition initiale à un système de décision assistée par IA sur le comportement ultérieur des utilisateurs (N = 172). Nous montrons que les conditions de pression temporelle et de difficulté de la première tâche peut influencer la sur-reliance, la confiance, la charge de travail et les temps de décision sur l'ensemble de l'interaction avec le système prédictif. Une analyse plus fine montre que le facteur déterminant est la première rencontre avec une condition exigeante, qui déclenche un changement de comportement persistant.*

## Mots-clés

*IA centrée sur l'humain, Décision assistée par IA, IHM*

## Abstract

*AI-assisted decision systems are proliferating in high-stakes domains. Proper calibration of human operators' trust and reliance (inclination to follow AI predictions) on these systems is necessary to ensure appropriate use. This calibration may be affected by time pressure and task difficulty, as well as by potential anchoring effects arising from initial interactions with the system. We conduct a study examining the impact of initial exposure conditions to an AI-assisted decision system on subsequent user behavior (N = 172). We show that the time pressure and difficulty conditions of the first task can influence over-reliance, trust, workload, and decision times throughout the interaction with the predictive system. Further analysis shows that the determining factor is the first encounter with a demanding condition, which triggers a persistent behavioral shift.*

## Keywords

*Human-centered AI, AI-assisted decision making, HCI*

## 1 Introduction

Les systèmes d'aide à la décision basés sur l'intelligence artificielle (IA) connaissent une croissance rapide dans de nombreux domaines à forts enjeux. En médecine, des modèles prédictifs assistent des radiologues ou des dermatologistes dans la détection de lésions cancéreuses [10, 7]. Dans le domaine militaire, ces systèmes sont déployés pour la fusion de données de renseignement et la planification opérationnelle [20].

Un système d'aide à la décision par IA est un système qui fournit à un opérateur humain une recommandation ou une prédiction issue d'un modèle algorithmique, afin de l'assister dans l'élaboration de son jugement pour traiter des instances d'un problème donné. Dans ce paradigme, la responsabilité de la décision incombe toujours à l'humain. Il est donc essentiel que la confiance et la reliance de l'opérateur envers le système soient calibrées en accord avec les performances effectives de ce dernier [18]. Nous faisons la distinction entre la confiance qui est une attitude envers le système prédictif, et la reliance qui est le comportement visant à reproduire les mêmes décisions que celles qui sont suggérées par le système prédictif, ou des décisions similaires [25]. Une confiance trop faible conduit à rejeter des recommandations pertinentes, tandis qu'une confiance excessive pousse l'opérateur à adopter les recommandations du système sans jugement critique, même lorsqu'elles sont erronées [18]. Dans cet article, nous utilisons le terme de sur-reliance pour désigner ce dernier comportement. Par ailleurs, le contexte opérationnel dans lequel s'inscrivent ces interactions peut influencer significativement la qualité des décisions. La pression temporelle et la difficulté des tâches qui sont traitées par les opérateurs sont susceptibles d'avoir un impact direct sur l'adoption des recommandations du système [23, 31]. A titre d'exemple, les radiologues travaillent souvent dans un contexte de stress intense, où ils doivent analyser avec une extrême précision des images médicales aux enjeux potentiellement vitaux, sous une forte pression temporelle et dans un environnement fréquemment

perturbé par les urgences, les interruptions et la charge de travail élevée [3]. En contexte militaire, la pression temporelle et la complexité des tâches sont des caractéristiques intrinsèques des environnements opérationnels, où les décideurs doivent traiter des informations incomplètes dans des délais souvent très contraints [8], avec des répercussions également susceptibles d'être vitales.

Au-delà de ces facteurs contextuels, des mécanismes cognitifs comme l'effet d'ancrage – tendance à se fier excessivement à la première information reçue (appelée *ancre*) pour prendre une décision ou faire une estimation, même si cette information est arbitraire ou peu pertinente – peuvent également avoir un impact important sur la calibration de la confiance et de la reliance dans les systèmes prédictifs [12]. L'effet d'ancrage se manifeste dans les interactions humain-IA à deux niveaux distincts. Le premier niveau est celui des décisions individuelles, où la prédiction du modèle peut constituer une ancre cognitive dont il est difficile pour l'opérateur de s'affranchir, même lorsqu'il dispose d'indices contradictoires [22, 4]. La littérature propose à cet égard des *cognitive forcing functions* (interventions délibérément conçues pour interrompre le raisonnement automatique et promouvoir un traitement analytique) comme mécanismes permettant de limiter cet ancrage [4]. Le second niveau d'apparition des effets d'ancrage est celui des utilisations répétées d'un système prédictif. Un effet d'ancrage peut se former à partir de la première impression que l'opérateur se forge des performances globales du modèle, et persister lors de tâches ultérieures : des utilisateurs exposés à de bonnes performances initiales auront une plus forte tendance à la sur-reliance lors des interactions suivantes, tandis que ceux confrontés à des erreurs précoces auront tendance à sous-estimer les compétences du système [21, 28, 2, 11]. Swaroop et al. ont observé un lien entre pression temporelle et effet d'ancrage, selon lequel l'exposition à une condition de forte pression temporelle crée un ancrage qui augmente la reliance et rend les décisions des utilisateurs plus rapides [27]. Dans cet article, nous cherchons à confirmer et à étendre ces résultats en examinant conjointement les effets de la pression temporelle et de la difficulté des tâches.

**Notre question de recherche est** : quels sont les impacts de l'ordre d'exposition à différents niveaux de pression temporelle et de difficulté de tâche, sur le comportement d'opérateurs humains envers un système d'aide à la décision par IA ? Les données présentées dans cet article ont été obtenues en réalisant le protocole pré-enregistré à l'adresse `https://osf.io/56wnj/`, dont l'objet d'étude initial porte sur le comportement humain en présence de différentes techniques d'explicabilité de l'IA (XAI). Nous proposons une analyse post-hoc d'un sous-ensemble des données collectées, qui permet d'aborder la question de recherche considérée dans cet article. Nous reportons ainsi les résultats d'une étude expérimentale ($N = 172$) dans laquelle nous faisons varier le niveau de pression temporelle (faible ou forte) et de difficulté (faible ou élevée) pour la résolution d'une tâche de détection de motifs. L'étude utilise un plan mixte inter-intra sujets : une partie des participants résout les tâches de manière autonome, l'autre est assistée par un modèle prédictif. Tous les participants sont soumis dans un ordre aléatoire aux 4 couples de conditions combinant pression temporelle et difficulté. En isolant les effets des premières conditions expérimentales sur le comportement global des participants, nous étudions en particulier la reliance et la sur-reliance envers le modèle, la confiance déclarée, la charge de travail perçue et les temps de décisions des participants.

# 2 Contexte scientifique

## 2.1 Calibration de la confiance et reliance

L'un des enjeux majeurs de la collaboration humain-IA réside dans la calibration de la confiance et de la reliance de l'opérateur envers le système prédictif [30]. Il est important de distinguer la confiance envers le système prédictif, qui est une attitude ou une intention humaine et qui est mesurée de façon déclarative, de la reliance (comportement de s'appuyer sur les prédictions du modèle pour la prise de décision) qui est un comportement effectif mesuré par des mesures d'accord avec les décisions du modèle [18, 25]. Bien que ces deux concepts soient a priori liés, les humains peuvent avoir un comportement de reliance effective contradictoire avec la confiance qu'ils déclarent subjectivement [29]. Cette discordance peut être causée par des facteurs externes, telles que la volonté de gagner du temps ou de réduire la charge de travail [17]. Schaffer et al. observent que des sujets affichant une familiarité élevée avec la tâche réalisée déclarent également une confiance supérieure dans le système prédictif, tout en rejetant plus souvent les conseils de l'IA [24]. La calibration de la confiance et de la reliance est un processus complexe, pouvant être influencé par de nombreux facteurs. Parmi ceux-ci, les conditions de pression temporelle, la difficulté des tâches et les effets d'ancrage jouent un rôle important. Ces différents aspects sont examinés dans la suite de cette section.

## 2.2 Effets de la pression temporelle et de la difficulté des tâches

Lorsque les opérateurs humains doivent prendre une décision sous une pression temporelle forte, ils sont plus enclins à adopter les suggestions du système prédictif [26]. Cette tendance s'observe y compris lorsque les suggestions sont incorrectes, menant à un phénomène de sur-reliance [23]. De plus, cette pression peut engendrer une anxiété qui altère les capacités de résolution de la tâche, provoquant un phénomène de « suffocation » (choking) où l'utilisateur se précipite pour terminer le travail en déléguant aveuglément la décision à l'IA, alors même que le temps imparti aurait été suffisant en l'absence de stress [13]. Le second paramètre à considérer est la complexité de la tâche à réaliser. Lorsque les opérateurs humains considèrent une tâche comme difficile, ils sont également plus enclins à adopter les suggestions de l'IA [1]. Il a été démontré expérimentalement que la sur-reliance augmente de manière significative avec la difficulté de la décision [31]. La pression temporelle et la difficulté des tâches ont des effets similaires sur l'aug-

mentation de la sur-reliance, mais présentent des nuances théoriques dans leurs mécanismes. La difficulté correspond au niveau de ressources cognitives requis pour résoudre la tâche, alors que la pression temporelle réduit les capacités cognitives mobilisables en induisant un stress environnemental [5]. Peu de travaux ont étudié l'interaction entre la pression temporelle et la difficulté des tâches. Hermanns et Teubner observent que dans un contexte de pression temporelle forte, l'effet négatif d'une augmentation de la difficulté sur la performance est faible [16]. Ils expliquent ce résultat par le fait qu'en situation de forte pression temporelle, les utilisateurs ont déjà tendance à déléguer leur jugement aux prédictions du modèle, ce qui atténue l'effet additionnel du niveau de difficulté de la tâche.

## 2.3 Effets d'ancrage

L'effet d'ancrage est un biais cognitif fort qui se manifeste lorsqu'une valeur initiale, appelée ancre, exerce une influence disproportionnée sur des jugements ultérieurs [12]. Dans un contexte de décision assistée par IA, des effets d'ancrage peuvent se manifester à plusieurs niveaux. Ils peuvent apparaître au niveau des décisions individuelles, notamment lorsque les suggestions des modèles de décision sont soumises à l'opérateur humain en même temps que l'instance du problème à résoudre. La prédiction agit alors comme une ancre dont l'humain peut éprouver des difficultés à se défaire [22], d'autant plus que l'IA peut être perçue comme une figure d'autorité ou une source de données hautement objectives [6]. Des stratégies ont été proposées dans la littérature pour limiter l'effet d'ancrage lié aux prédictions individuelles, en augmentant délibérément le temps alloué à la décision [22] ou en intégrant des mécanismes dont le but est de forcer la réflexion humaine (*cognitive forcing functions*) avant de prendre connaissance de la prédiction du modèle [4]. Des effets d'ancrage peuvent aussi se manifester à une échelle temporelle plus large, entre plusieurs phases d'interactions avec le système prédictif. Une première impression positive du système, résultant d'une exposition initiale à des succès de l'IA, peut augmenter la reliance lors d'interactions ultérieures [28]. Biswas et al. observent un effet d'« inertie des croyances », par lequel une confiance élevée dans le système prédictif peut se transférer sous forme de confiance a priori pour une tâche ultérieure, y compris si les domaines des tâches sont différents [2]. La distribution temporelle des erreurs constitue également un facteur d'ancrage. Freel et al. observent que des erreurs précoces peuvent être plus facilement pardonnées par les utilisateurs que des erreurs tardives, car un mécanisme de réparation de la confiance peut avoir lieu si le système prédictif se comporte de manière fiable par la suite [11].

## 2.4 Positionnement de l'étude

Dans cet article, nous cherchons à mesurer les effets d'ancrage liés spécifiquement aux conditions de pression temporelle et de difficulté des tâches lors des premières interactions avec le modèle. Nous cherchons à étudier l'impact de ces effets sur la calibration de la confiance et de la reliance envers les modèles prédictifs. Les travaux de la littérature les plus proches sont ceux de Swaroop et al. [27], qui étudient le compromis entre efficacité (minimiser les temps de réponse) et performance (taux de bonnes réponses) pour la résolution d'une tâche dans un contexte de décision assistée par IA. Leur protocole fait varier la pression temporelle imposée aux participants. Ils montrent que la pression temporelle augmente la sur-reliance envers le système. Ils rapportent également (sans en présenter les données) que la première exposition à une tâche sous forte pression temporelle peut induire un comportement de sur-reliance persistant, qui se maintient lorsque la pression temporelle diminue par la suite. Notre protocole est similaire en ce que les participants sont exposés successivement à différents niveaux de pression temporelle, mais il s'en distingue sur plusieurs points. D'une part, nous étudions également l'effet de la difficulté des tâches [1] ; et d'autre part, nous analysons, en plus de la performance, du temps de décision et de la sur-reliance, qui sont les variables reportées par Swaroop et al., les variables subjectives que représentent la confiance déclarée envers le modèle et la charge de travail associée aux tâches. Enfin, notre analyse isole explicitement l'effet de la première condition expérimentale sur l'ensemble du comportement ultérieur des participants.

# 3 Méthodes

Cette section présente la méthodologie mise en œuvre pour étudier les effets d'ancrage liés à la pression temporelle et à la difficulté des tâches dans un contexte de décision assistée par IA. Comme indiqué dans l'introduction, les résultats présentés proviennent d'une analyse post-hoc de données collectées dans le cadre d'un protocole plus large incluant des techniques d'explicabilité (XAI), qui fera l'objet d'une communication dédiée. Nous restreignons notre analyse aux participants n'ayant pas eu accès à des explications, répartis en deux conditions : résolution sans assistance (groupe H) et avec une prédiction de l'IA (groupe H+IA).

## 3.1 Plan d'expériences

Le protocole expérimental a été soumis au Comité d'éthique de la recherche de l'Université de Montpellier (projet « Étude de l'impact de l'intelligence artificielle explicable sur la performance d'opérateurs humains pour une tâche de décision assistée par l'Intelligence Artificielle »), qui a rendu un avis favorable (Avis consultatif n°UM 2025-091). Il a aussi été pré-enregistré publiquement sur la plateforme Open Science Framework (OSF) [2] avant la collecte des résultats.

### 3.1.1 Tâche de décision

La tâche soumise aux participants consiste en une décision binaire : pour chaque couple (image, motif), le participant doit indiquer si le motif — ou l'une de ses rota-

1. Notons que Swaroop et al. considèrent également la difficulté comme une variable indépendante, mais uniquement pour la deuxième expérience reportée dans leur article. Seule leur première expérience traite de la question de l'ordre des conditions de pression temporelle.

2. Le protocole est accessible à l'adresse https://osf.io/56wnj/.

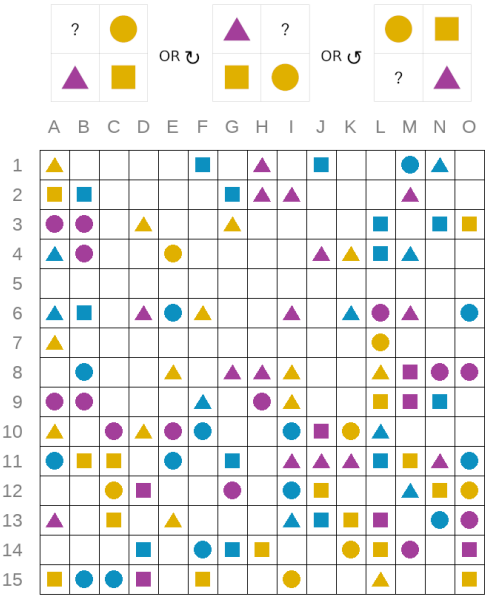


FIGURE 1 – Exemple de question soumise aux participants. Partie supérieure : motif ou ses rotations à identifier dans la grille. Partie inférieure : grille de symboles (de difficulté élevée). La question soumise explicitement est : le motif ou l'une de ses rotations est-il présent dans l'image ? Pour cet exemple, le motif est bien présent (la rotation vers la droite aux positions N11-N12-O12).

tions à 90° (gauche ou droite) — est présent dans l'image. Chaque image, générée aléatoirement, se présente sous la forme d'une grille de cellules contenant chacune un symbole coloré parmi un ensemble $S^-$ ($|S^-| = 9$) ou aucun symbole ($\emptyset$), de sorte qu'une grille de taille $n \times n$ est un élément de $S^{n \times n}$ avec $S = S^- \cup \emptyset$. Un motif correspond à un élément de $S^{2\times2}$, et chaque image contient au plus une occurrence du motif ou de ses rotations. Une tâche comprend une série de questions portant sur différentes images et un même motif. La Figure 1 illustre une instance de question soumise aux participants.

Différentes conditions expérimentales sont considérées en fonction de la difficulté des tâches et de la pression temporelle imposée aux participants pour les résoudre. La difficulté est caractérisée comme fonction de la taille de la grille associée à l'image présentée à l'opérateur :

1. Facile : grille dans $S^{9\times9}$, motif dans $S^{2\times2}$.
2. Difficile : grille dans $S^{15\times15}$, motif dans $S^{2\times2}$.

La pression temporelle est caractérisée par le temps alloué aux participants pour prendre chaque décision. Un minuteur permet aux participants de savoir à tout moment combien de temps il leur reste pour répondre à une question. Lorsque le temps est imparti, l'interface passe à la question suivante. L'absence de réponse dans le temps imparti est considérée comme une mauvaise réponse. Deux niveaux de pression temporelle sont distingués :

1. Faible : 25 secondes maximum par image.
2. Forte : 10 secondes maximum par image.

Les différentes conditions expérimentales sont ainsi croisées de manière à proposer des tâches faciles ou difficiles avec une pression temporelle faible ou forte, au sens des définitions proposées ci-dessus. Nous comparons les performances et les comportement observés des participants ayant accès (groupe H+IA) ou non (groupe H) à l'assistance d'un modèle prédictif pour résoudre les tâches. L'assistance se présente sous la forme d'un encart "L'IA prédit : Oui/Non" visible dans l'interface de résolution (en anglais). Les participants ont dans tous les cas la responsabilité de la décision finale et sont libres de considérer ou d'ignorer les recommandations du modèle dans leurs processus décisionnels. Les participants ne sont pas informés de la performance des modèles, mais ils sont informés à la fin de chaque tâche de décision de leur score (pourcentage de réponses correctes). Ce choix vise à reproduire des conditions proches d'un usage réel, dans lequel les opérateurs construisent progressivement leur estimation des performances du système à partir de leur expérience accumulée lors d'utilisations répétées, et sans retour immédiat à chaque décision.

### 3.1.2 Variables dépendantes

Les variables dépendantes mesurées pour chaque tâche de l'expérience principale sont :

- Le score (proportion de réponses correctes, i.e. *accuracy*).
- La reliance (proportion de réponses identiques à celles de l'IA).
- La sur-reliance (proportion de réponses identiques à celles de l'IA lorsque celle-ci produit une mauvaise réponse).
- Le temps de réponse moyen.
- La confiance dans le système d'assistance (échelle de Likert à 7 points pour la question « Je fais confiance à l'IA pour détecter le motif et ses rotations. »).
- Charge de travail : somme de l'enquête standard NASA-TLX [14] sur une échelle à 7 points par question. La question concernant la charge physique est exclue de la somme car elle n'est pas considérée pertinente ici. Cette question a préalablement été exclue dans d'autres travaux de la littérature [28].

Les questionnaires pour mesurer la confiance et la charge de travail sont soumis aux participants immédiatement après les tâches de l'expérience principale, et avant de leur communiquer leur score pour la tâche courante, afin de ne pas biaiser la mesure de la charge de travail — dont l'une des questions porte sur l'estimation de la performance individuelle lors de la tâche.

### 3.1.3 Procédure

La procédure suivie par chaque participant comprend deux phases principales, illustrées en Figure 2. Une phase introductive permet d'abord aux participants de se familiariser avec l'interface et les tâches, afin de se préparer pour la phase d'expérience principale, pendant laquelle les données seront collectées. L'introduction comprend une première tâche de familiarisation comprenant 4 questions, sans

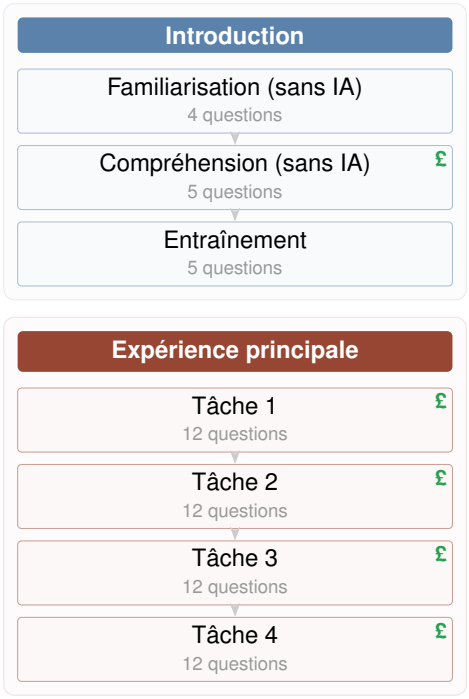


FIGURE 2 – Déroulement du protocole. Le symbole £ précise les tâches considérées pour la rémunération bonus.

limite de temps de réponse et sans assistance de l'IA, quel que soit le groupe. Elle comprend ensuite une tâche de compréhension (5 questions, sans limite de temps et sans assistance de l'IA), qui sera utilisée pour exclure de l'analyse statistique les participants n'ayant pas démontré une compréhension suffisante de la tâche. La tâche de compréhension est la seule tâche de l'introduction qui est prise en compte pour la rémunération bonus des participants, afin de les inciter à répondre du mieux possible. L'introduction se clôt avec une tâche d'entraînement comprenant 5 questions et dans laquelle l'assistance IA et ses éventuelles explications sont introduites pour les groupes concernés. Une limite de temps de réponse de 45 secondes est également introduite pour cette tâche. Les grilles de l'introduction sont de taille $12 \times 12$ (difficulté intermédiaire entre Facile et Difficile), sauf pour le test de compréhension pour lequel les grilles sont de taille $5 \times 5$. À l'issue de cette phase débute la phase expérimentale proprement dite. Elle se compose de quatre tâches de 12 questions chacune. Chaque tâche correspond à une combinaison unique des facteurs pression temporelle et difficulté de tâche, avec Pression $\in \{\text{Faible}, \text{Élevée}\}$ et Difficulté $\in \{\text{Facile}, \text{Difficile}\}$. L'ensemble des conditions expérimentales est ainsi $\{\text{Faible}, \text{Élevée}\} \times \{\text{Facile}, \text{Difficile}\}$, et l'ordre de présentation de ces conditions est tiré aléatoirement pour chaque participant. Chaque condition est associée à un motif spécifique et à un ensemble prédéterminé de 12 grilles. L'ordre de présentation des grilles pour chaque participant est déterminé aléatoirement. Pour chaque ensemble de 12 grilles, le modèle prédictif commet exactement deux erreurs (un faux positif et un faux négatif), identiques pour l'ensemble des participants. Cela correspond donc à un système d'assistance présentant une *accuracy* très proche de 85%. À l'issue de chaque tâche de l'expérience principale, les participants répondent à des questionnaires évaluant leur confiance et leur charge de travail. Leur score pour la tâche qu'ils viennent d'effectuer leur est ensuite communiqué. L'expérience principale inclut également deux tests d'attention[3], placés après la première et après la quatrième tâche. L'expérience a été conçue pour être réalisée de manière autonome, en ligne via le Web, pour une durée totale d'environ 20 minutes.

#### 3.1.4 Participants

L'obtention des résultats a mobilisé 201 participants recrutés via Prolific[4]. Après exclusions pour problèmes techniques (5), échecs à un test d'attention (16) ou au test de compréhension (8), la cohorte retenue comprend 172 participants : 85 dans le groupe H et 87 dans le groupe H+IA. Tous sont adultes résidant au Royaume-Uni (55,8% de femmes, âge moyen de 44,0 ans). Les personnes présentant des troubles de vision des couleurs étaient exclues dès le recrutement, la tâche reposant sur la discrimination de symboles colorés. Les participants ont été rémunérés à un taux fixe de 6£/h, complété par un bonus de 0,04£ par bonne réponse aux 53 questions évaluées (soit un taux horaire effectif entre 6£ et 12,12£). L'assignation aux groupes H ou H+IA est aléatoire.

### 3.2 Modèles d'IA

La tâche de prédiction est une classification binaire (présence ou absence du motif) à partir d'une image RGB ($256 \times 256 \times 3$). Un modèle ResNet-18 [15] est entraîné de manière supervisée pour chaque couple (dimension de la grille, motif), à l'aide de jeux de données synthétiques (49k images d'entraînement, 1k de test). L'architecture atteignant aisément 100% de précision sur cette tâche, l'entraînement est volontairement interrompu dès que le seuil de 85% est atteint sur le jeu de test, afin de simuler un système d'assistance imparfait. Toutes les données incluses dans le protocole sont issues du jeu de test.

### 3.3 Implémentation et hébergement

Le protocole a été déployé sur une instance de la plateforme web WebXAII [19], hébergée sur des serveurs d'IMT Mines Alès. L'ensemble des pages nécessaires à l'expérience (informations, instructions, tâches de décision et questionnaires) y ont été implémentées.

### 3.4 Analyses statistiques

Deux analyses complémentaires sont menées pour étudier l'influence de l'ordre des tâches sur les variables dépendantes. Les hypothèses de normalité et d'homoscédasticité n'étant pas garanties avec des sous-groupes de petite taille, des tests non paramétriques U de Mann-Whitney bilatéraux

3. Chaque test comporte deux questions : l'une demande de sélectionner une réponse précise sur une échelle de Likert (ex. « Somewhat Disagree »); l'autre propose un choix parmi trois options, leur demandant la couleur du ciel ou le principe des tâches qu'ils réalisent.

4. https://www.prolific.com/

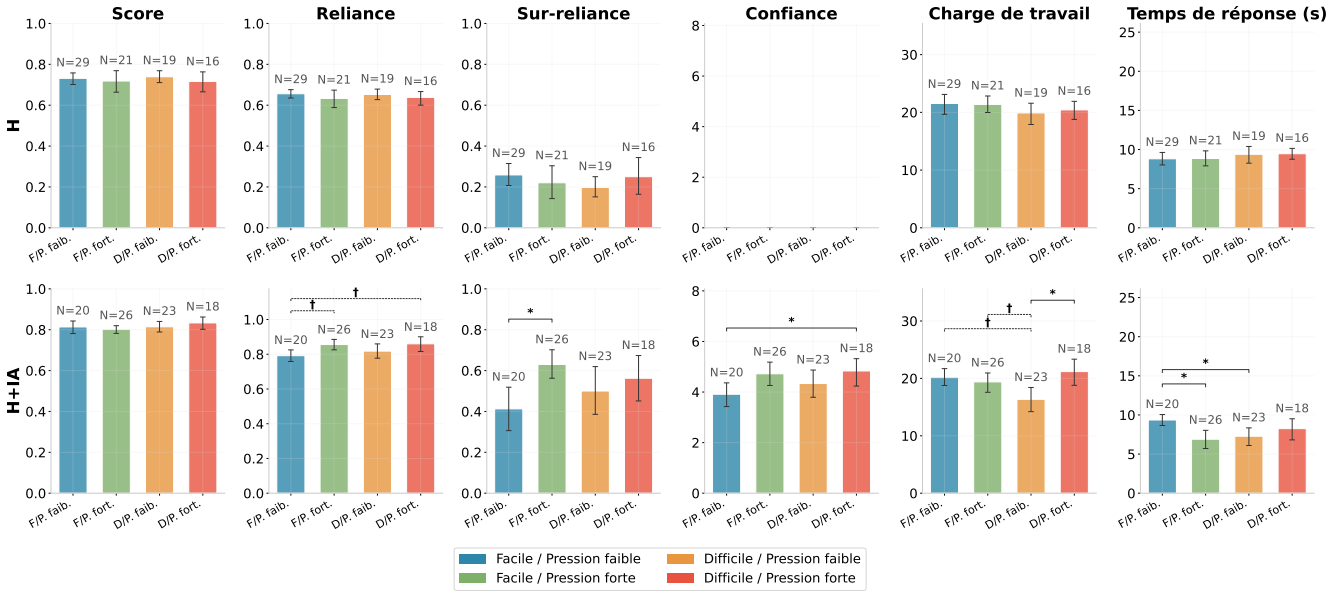


FIGURE 3 – Valeurs des variables dépendantes agrégées sur l'ensemble des tâches pour les conditions H et H+IA, avec intervalles de confiance à 95% obtenus par bootstrap, en fonction de la tâche initiale (combinaison difficulté × pression temporelle). Les différences significatives sont indiquées par des astérisques : * pour $p_{\text{BH}} < .05$, ** pour $p_{\text{BH}} < .01$ et *** pour $p_{\text{BH}} < .001$. Les différences marginales ($p_{\text{BH}} < .10$) sont signalées par le symbole †.

sont utilisés pour l'ensemble des comparaisons. La taille d'effet est estimée par le $d$ de Cohen et les intervalles de confiance à 95% sont obtenus par bootstrap (10 000 rééchantillonnages). Les $p$-values sont corrigées par la procédure de Benjamini-Hochberg (BH-FDR), qui contrôle le taux de fausses découvertes tout en offrant une puissance supérieure à des corrections plus conservatrices telles que Bonferroni.

La première analyse porte sur l'effet de la tâche initiale. Les participants sont répartis en quatre groupes selon la première tâche effectuée, et les variables dépendantes sont agrégées sur l'ensemble des quatre tâches. Des tests de Mann-Whitney sont réalisés pour chacune des six paires de groupes, séparément par condition expérimentale (H, H+IA) et par variable dépendante, cette combinaison définissant la famille de correction BH-FDR.

La seconde analyse porte sur l'effet de l'exposition antérieure à un type de tâche donné, sur la condition H+IA uniquement. Pour chaque dimension (pression temporelle, difficulté) et chaque direction d'exposition, les observations sont réparties en trois groupes selon l'historique des tâches précédentes : aucune exposition au type étudié, exposition aux deux types, ou exposition uniquement au type étudié. Par exemple, pour l'analyse de l'effet d'un antécédent à pression forte, les trois groupes correspondent aux observations sans pression forte antérieure, avec pression forte et faible antérieures, et avec uniquement pression forte antérieure. Quatre séries de comparaisons sont ainsi effectuées, et les trois tests par paires au sein de chaque combinaison série × variable dépendante constituent la famille de correction. Cette analyse vise à tester si l'exposition antérieure à une condition exigeante a un impact persistant sur le comportement lors des tâches suivantes, en lien avec l'observation par Swaroop et al. selon laquelle la première exposition à une forte pression temporelle déclencherait un comportement de sur-reliance durable [27]. Nous cherchons à vérifier cet effet sur nos données et à déterminer si un effet similaire apparaît avec la première exposition à une tâche difficile.

Nous effectuons finalement une analyse des effets directs du genre et de l'âge sur les variables dépendantes, dans les deux groupes H et H+IA, respectivement par un test de Mann-Whitney bilatéral et une corrélation de Spearman, avec correction BH-FDR appliquée séparément pour chaque famille de tests (six tests par famille).

# 4 Résultats et discussions

Dans cette section, nous présentons les principaux résultats issus de l'expérimentation. Nous analysons dans un premier temps l'influence de la première tâche expérimentale sur le comportement global des participants, puis nous examinons plus finement les effets de la première exposition à une condition exigeante (forte pression temporelle ou difficulté élevée) ou peu exigeante (faible pression temporelle ou difficulté faible) sur le comportement lors des tâches ultérieures.

## 4.1 Effets de la tâche initiale sur le comportement global

La Figure 3 présente les variables dépendantes agrégées sur l'ensemble des quatre tâches, en fonction de la première condition expérimentale rencontrée par chaque participant, pour les groupes H (sans assistance) et H+IA. La Table 1 détaille les comparaisons par paires dont la p-value est inférieure à 0,1 – ce qui inclut l'ensemble des comparaisons

TABLE 1 – Comparaisons par paires selon la première condition ($p_{\text{BH}} < .10$).

| Cond. | Variable | Comparaison | $M_1$ | $M_2$ | $d$ | $p_{\text{MW}}$ | $p_{\text{BH}}$ |
|---|---|---|---|---|---|---|---|
| H+IA | Charge de travail | Diff./Press. faible vs Diff./Press. forte | 16.348 | 21.181 | -0.93 | 0.0058 | **0.0346** |
| H+IA | Charge de travail | Fac./Press. faible vs Diff./Press. faible | 20.200 | 16.348 | +0.86 | 0.0213 | 0.0639 |
| H+IA | Charge de travail | Fac./Press. forte vs Diff./Press. faible | 19.385 | 16.348 | +0.62 | 0.0495 | 0.0990 |
| H+IA | Confiance | Fac./Press. faible vs Diff./Press. forte | 3.913 | 4.833 | -0.78 | 0.0029 | **0.0173** |
| H+IA | Reliance | Fac./Press. faible vs Fac./Press. forte | 0.792 | 0.856 | -0.81 | 0.0100 | 0.0546 |
| H+IA | Reliance | Fac./Press. faible vs Diff./Press. forte | 0.792 | 0.860 | -0.80 | 0.0182 | 0.0546 |
| H+IA | Sur-reliance | Fac./Press. faible vs Fac./Press. forte | 0.412 | 0.630 | -1.01 | 0.0057 | **0.0341** |
| H+IA | Temps de réponse | Fac./Press. faible vs Fac./Press. forte | 9.353 | 6.894 | +0.97 | 0.0071 | **0.0247** |
| H+IA | Temps de réponse | Fac./Press. faible vs Diff./Press. faible | 9.353 | 7.278 | +0.89 | 0.0082 | **0.0247** |

significatives au seuil de 0,05.

Un premier résultat notable concerne le groupe H : aucune différence significative n'est observée entre les quatre groupes définis par la tâche initiale, que ce soit en termes de score, de charge de travail ou de temps de réponse. L'absence d'effet d'ordre dans la condition sans assistance constitue un résultat de contrôle important : il indique que les effets observés dans la condition H+IA ne sont pas imputables à des différences intrinsèques entre les groupes de participants ou à un simple effet d'apprentissage de la tâche, mais relèvent bien d'une interaction avec le système prédictif. Pour le groupe H+IA, les résultats révèlent des différences significatives en fonction de la tâche initiale. Les participants ayant débuté par la condition Facile/Pression faible se distinguent des autres sur plusieurs variables. Leur sur-reliance est significativement plus faible que celle des participants ayant débuté par Facile/Pression forte ($d = -1{,}01$, $p_{\text{BH}} = {,}034$). Leur confiance dans le modèle est significativement inférieure à celle des participants ayant débuté par Difficile/Pression forte ($d = -0{,}78$, $p_{\text{BH}} = {,}017$). Enfin, leurs temps de réponse sont significativement plus longs que ceux des participants ayant débuté par Facile/Pression forte ($d = +0{,}97$, $p_{\text{BH}} = {,}025$) ou par Difficile/Pression faible ($d = +0{,}89$, $p_{\text{BH}} = {,}025$). On observe également une tendance marginale pour la reliance, qui est plus faible chez les participants ayant débuté par Facile/Pression faible par rapport à ceux ayant débuté par Facile/Pression forte ou Difficile/Pression forte. La charge de travail présente un profil différent : les participants ayant débuté par Difficile/Pression faible rapportent une charge significativement inférieure à celle des participants ayant débuté par Difficile/Pression forte ($d = -0{,}93$, $p_{\text{BH}} = {,}035$), avec des tendances marginales par rapport aux deux groupes ayant débuté par une tâche facile. Ces résultats suggèrent que débuter par une tâche facile sous faible pression temporelle offre aux participants l'opportunité de développer une stratégie de résolution autonome. Il est probablement plus facile pour les participants de constater que le modèle prédictif commet des erreurs dans cette condition. Cette première expérience d'un travail effectif de résolution semble ancrer un comportement d'évaluation critique des prédictions qui perdure sur l'ensemble de l'expérience, se traduisant par une confiance et une sur-reliance plus faibles ainsi que des temps de réponse plus longs. À l'inverse, les participants confrontés d'emblée à une tâche exigeante semblent contraints de recourir rapidement à une stratégie de délégation au modèle, faute de ressources cognitives ou temporelles suffisantes pour résoudre la tâche de manière autonome. Toutefois, ces conclusions doivent être nuancées : les effets ne sont pas systématiques sur l'ensemble des combinaisons de facteurs. Par exemple, la comparaison entre les participants ayant débuté par Facile/Pression faible et ceux ayant débuté par Difficile/Pression forte n'atteint pas le seuil de significativité pour les temps de réponse, alors que les comparaisons impliquant un seul facteur sont significatives. Cela suggère que les mécanismes par lesquels la pression temporelle et la difficulté influencent le comportement ne sont pas strictement additifs.

## 4.2 Effets de l'exposition antérieure à une condition exigeante

La Figure 4 et la Table 2 détaillent les effets de l'exposition antérieure aux différentes conditions, en distinguant quatre analyses selon la dimension étudiée (pression temporelle ou difficulté) et la direction de l'effet. Les observations sont réparties en trois groupes selon l'historique des tâches précédentes : aucune exposition au type étudié, exposition aux deux types, ou exposition uniquement au type étudié.

### 4.2.1 Effets de l'exposition antérieure à une pression forte

Pour les observations réalisées sur des tâches à pression faible, les participants ayant déjà rencontré au moins une tâche à pression forte présentent une confiance significativement plus élevée que ceux n'ayant jamais été exposés à la pression forte. Cet effet est particulièrement marqué pour le groupe n'ayant connu que de la pression forte auparavant ($d = -0{,}80$, $p_{\text{BH}} = {,}001$) et reste significatif lorsque les deux types de pression ont été rencontrés ($d = -0{,}56$, $p_{\text{BH}} = {,}003$). Une tendance marginale dans la même direction est observée pour la sur-reliance. Les temps de réponse sont significativement plus courts pour les participants ayant été exposés aux deux types de pression ($d = +0{,}57$, $p_{\text{BH}} = {,}007$), avec une tendance marginale pour ceux n'ayant connu que la pression forte. Ces résultats indiquent que l'exposition préalable à une pression temporelle forte déclenche un comportement de délégation accrue

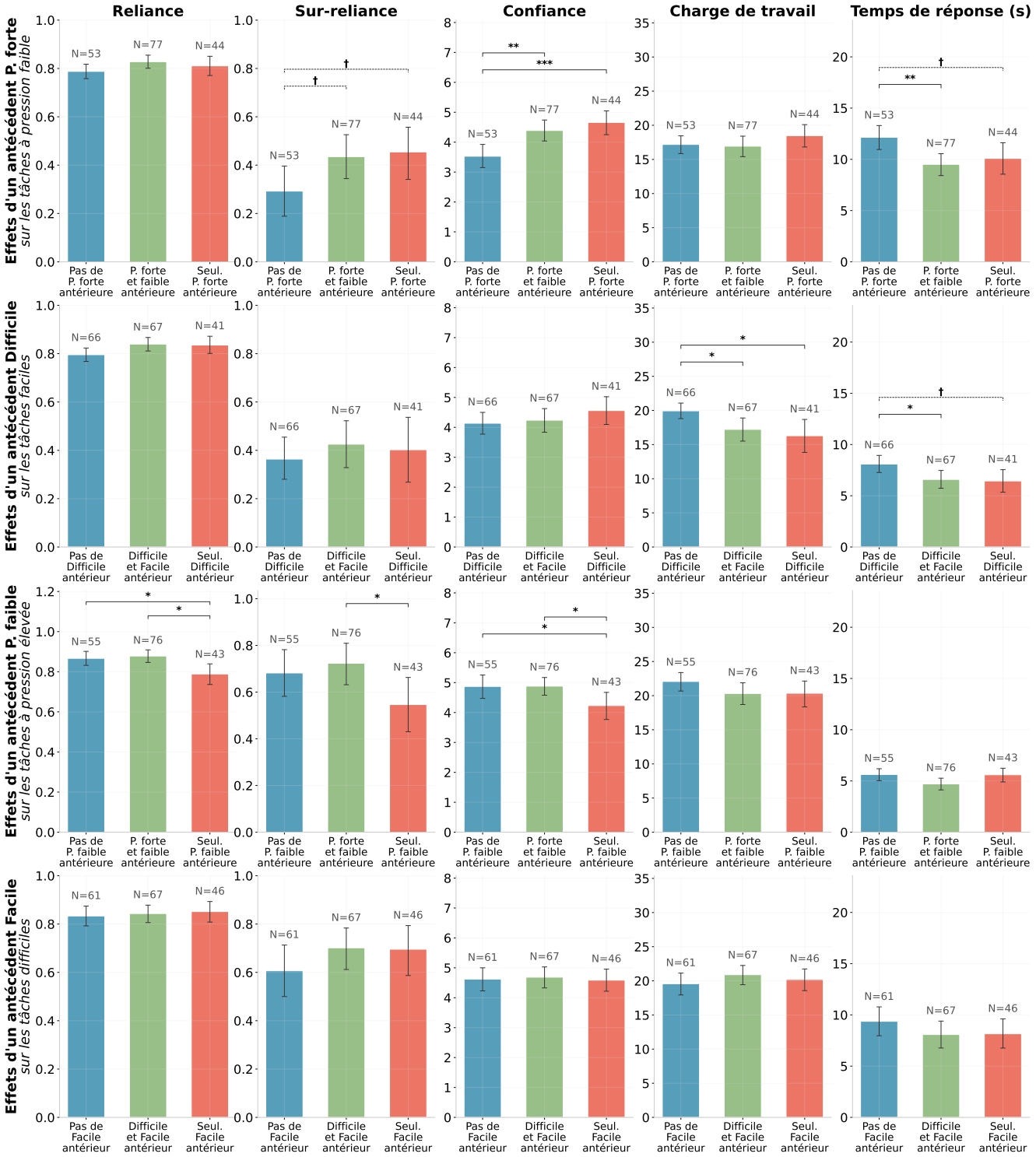


FIGURE 4 – Analyse des variables selon l'exposition antérieure aux conditions opposées, avec intervalles de confiance à 95% obtenus par bootstrap. Pour chaque type de tâche, les observations sont réparties en trois groupes selon l'historique des tâches précédentes : aucune exposition au type opposé, exposition aux deux types, ou exposition uniquement au type opposé. Les comparaisons par paires sont corrigées par famille par la procédure de Benjamini-Hochberg (BH-FDR). Les différences significatives sont indiquées par des astérisques : * pour $p_{\text{BH}} < .05$, ** pour $p_{\text{BH}} < .01$ et *** pour $p_{\text{BH}} < .001$. Les différences marginales ($p_{\text{BH}} < .10$) sont signalées par le symbole †.

TABLE 2 – Effet de l'exposition antérieure — Mann-Whitney U, correction BH-FDR, $p_{\text{BH}} < .10$ — H+IA

| Analyse | VD | Comparaison | $M_i$ | $M_j$ | $d$ | $p_{\text{BH}}$ |
|---|---|---|---|---|---|---|
| Tâches à pression faible | Sur-reliance | Pas de P. forte ant. vs P. forte et faible ant. | 0.292 | 0.435 | −0.36 | 0.0640 |
| | | Pas de P. forte ant. vs Seul. P. forte ant. | 0.292 | 0.455 | −0.43 | 0.0640 |
| | Confiance | Pas de P. forte ant. vs Seul. P. forte ant. | 3.528 | 4.659 | −0.80 | **0.0005** |
| | | Pas de P. forte ant. vs P. forte et faible ant. | 3.528 | 4.390 | −0.56 | **0.0029** |
| | Tps réponse (s) | Pas de P. forte ant. vs P. forte et faible ant. | 12.149 | 9.512 | +0.57 | **0.0072** |
| | | Pas de P. forte ant. vs Seul. P. forte ant. | 12.149 | 10.093 | +0.43 | 0.0589 |
| Tâches faciles | Charge de travail | Pas de Diff. ant. vs Diff. et Facile ant. | 19.939 | 17.224 | +0.45 | **0.0469** |
| | | Pas de Diff. ant. vs Seul. Diff. ant. | 19.939 | 16.293 | +0.58 | **0.0469** |
| | Tps réponse (s) | Pas de Diff. ant. vs Diff. et Facile ant. | 8.094 | 6.592 | +0.42 | **0.0208** |
| | | Pas de Diff. ant. vs Seul. Diff. ant. | 8.094 | 6.443 | +0.47 | 0.0705 |
| Tâches à pression forte | Reliance | P. forte et faible ant. vs Seul. P. faible ant. | 0.878 | 0.789 | +0.59 | **0.0114** |
| | | Pas de P. faible ant. vs Seul. P. faible ant. | 0.867 | 0.789 | +0.51 | **0.0391** |
| | Sur-reliance | P. forte et faible ant. vs Seul. P. faible ant. | 0.724 | 0.547 | +0.44 | **0.0418** |
| | Confiance | Pas de P. faible ant. vs Seul. P. faible ant. | 4.873 | 4.233 | +0.42 | **0.0376** |
| | | P. forte et faible ant. vs Seul. P. faible ant. | 4.882 | 4.233 | +0.46 | **0.0376** |

au modèle prédictif — se traduisant par une confiance plus élevée et des temps de réponse plus courts — qui persiste même lorsque la pression temporelle diminue. Ce résultat est cohérent avec les observations de Swaroop et al. [27], ce qui tend à confirmer l'existence d'un effet d'ancrage lié à la pression temporelle.

#### 4.2.2 Effets de l'exposition antérieure à une tâche difficile

Pour les observations réalisées sur des tâches faciles, les participants ayant déjà rencontré au moins une tâche difficile rapportent une charge de travail significativement plus faible, aussi bien pour le groupe ayant connu les deux types de difficulté ($d = +0{,}45$, $p_{\text{BH}} = {,}047$) que pour celui n'ayant connu que des tâches difficiles ($d = +0{,}58$, $p_{\text{BH}} = {,}047$). Les temps de réponse sont également significativement plus courts pour le groupe ayant été exposé aux deux types ($d = +0{,}42$, $p_{\text{BH}} = {,}021$), avec une tendance marginale pour celui n'ayant connu que des tâches difficiles. L'exposition antérieure à la difficulté ne semble pas affecter directement la confiance ou la sur-reliance, contrairement à l'exposition à la pression temporelle. Son effet se manifeste plutôt par une réduction de la charge de travail perçue et des temps de décision. Ce résultat suggère que les participants ayant d'abord effectué des tâches difficiles tendent à évaluer les tâches faciles suivantes comme nécessitant moins d'effort, ce qui se traduit par des décisions plus rapides. Il pourrait s'agir d'un effet de contraste plutôt que d'un ancrage sur la reliance, les participants ajustant leur investissement cognitif à la baisse après avoir expérimenté un niveau de difficulté supérieur.

#### 4.2.3 Effets de l'exposition antérieure à une pression faible et à une tâche facile

L'examen de l'impact d'une exposition antérieure à des conditions peu exigeantes sur le comportement lors de tâches exigeantes complète les analyses précédentes. Pour les tâches à pression forte, les différences significatives apparaissent systématiquement en comparaison au groupe n'ayant connu que de la pression faible auparavant : ce groupe présente une reliance, une sur-reliance et une confiance significativement plus faibles que les participants ayant déjà rencontré de la pression forte (que ce soit exclusivement ou en combinaison avec de la pression faible). Ces résultats confirment que c'est bien la rencontre préalable avec une tâche à pression forte qui déclenche le changement de comportement : les participants exclusivement exposés à la pression faible maintiennent un usage plus modéré du système d'assistance, tandis que dès lors qu'une tâche à pression forte a été rencontrée, le comportement de délégation s'installe indépendamment de l'historique ultérieur. En revanche, pour les tâches difficiles précédées de tâches faciles, aucune différence significative n'est observée sur l'ensemble des variables étudiées. L'exposition antérieure à des tâches faciles ne semble donc pas modifier le comportement des participants lors de tâches difficiles.

#### 4.2.4 Asymétrie des effets d'ancrage

L'ensemble de ces résultats met en évidence une asymétrie dans les effets d'ancrage liés aux conditions expérimentales. Les conditions exigeantes exercent un effet d'ancrage sur le comportement lors des tâches ultérieures moins exigeantes : la pression temporelle forte ancre une confiance et une reliance élevées qui persistent, tandis que la difficulté élevée ancre une réduction de la charge de travail perçue et des temps de réponse. L'analyse des effets en fonction des antécédents à pression ou à difficulté faible confirme cette asymétrie : pour la pression temporelle, c'est la rencontre avec une tâche à pression forte qui constitue le facteur déterminant du changement de comportement, puisque les différences significatives impliquent systématiquement le

groupe n'ayant connu que de la pression faible. Pour la difficulté, l'exposition antérieure à des tâches faciles n'exerce aucun effet détectable sur le comportement lors de tâches difficiles. Cette asymétrie suggère que les conditions exigeantes constituent des ancres cognitives plus puissantes que les conditions peu exigeantes. D'un point de vue pratique, ce résultat implique que l'ordre dans lequel les opérateurs sont exposés aux différentes conditions de travail avec un système d'assistance par IA n'est pas anodin : une exposition progressive, débutant par des conditions peu contraignantes, pourrait favoriser une calibration plus appropriée de la confiance et de la reliance.

### 4.3 Variabilité interindividuelle

Dans le groupe H+IA, le genre et l'âge ne sont associés à aucune variable comportementale de manière significative (tous les $p_{\text{BH}} > ,23$), à l'exception de la charge de travail reportée, qui est significativement plus élevée chez les femmes que chez les hommes ($M_{\text{femmes}} = 21{,}2$ vs $M_{\text{hommes}} = 17{,}2$, $d = +0{,}89$, $p_{\text{BH}} < ,001$). Il s'agit d'un effet qui a déjà été observé dans la littérature [9]. Dans le groupe H, le genre n'est associé à aucune variable dépendante (tous les $p_{\text{BH}} > ,44$). En revanche, l'âge prédit négativement le score ($\rho = -0{,}28$, $p_{\text{BH}} = ,024$), et positivement la charge de travail ($\rho = +0{,}30$, $p_{\text{BH}} = ,024$). Ces effets de l'âge, absents dans le groupe H+IA, suggèrent que l'assistance par IA pourrait atténuer l'influence de l'âge sur la performance et la charge de travail.

## 5 Limites

Notre étude présente plusieurs limites qu'il convient de mentionner. Premièrement, les résultats sont issus d'une analyse post-hoc de données obtenues pour un protocole ayant pour but initial d'examiner l'impact de techniques d'explicabilité de l'IA. L'assignation des participants aux conditions de tâche initiale résulte d'une randomisation de l'ordre des tâches, ce qui engendre des groupes de tailles inégales. Un protocole dédié, avec un contrôle explicite de la tâche initiale et des groupes plus équilibrés, serait nécessaire pour confirmer les effets observés avec une puissance statistique accrue.

Deuxièmement, la tâche proposée constitue une abstraction qui ne reflète pas directement les conditions réelles d'utilisation des systèmes d'aide à la décision. Si cette simplification permet un contrôle rigoureux de la difficulté et de la pression temporelle, elle limite la généralisation des résultats à des contextes professionnels spécifiques. Des travaux futurs pourraient reproduire ces analyses avec des tâches plus proches de situations réelles, dans des domaines où les systèmes d'aide à la décision par IA sont déjà déployés, tels que la radiologie ou le domaine militaire. L'expertise métier des opérateurs spécialisés de ces domaines pourrait moduler la résistance aux effets d'ancrage.

Troisièmement, le système d'assistance présente un taux d'erreur fixe d'environ $15\%$, identique pour toutes les conditions. Dans un contexte réel, les performances d'un modèle prédictif pourraient varier selon la difficulté des instances. Il serait intéressant, dans des travaux futurs, de faire varier indépendamment la difficulté des instances et le taux d'erreur du modèle, afin de mesurer la contribution propre de chacun de ces facteurs sur la reliance des opérateurs humains.

## 6 Conclusion

Cette étude contribue à montrer que la construction de la relation de confiance entre un opérateur humain et un système prédictif d'IA peut dépendre des premières interactions avec ce système. En analysant le comportement de $172$ participants dans une tâche de détection de motifs assistée par IA, nous mettons en évidence des effets d'ancrage liés aux conditions de pression temporelle et de difficulté rencontrées lors des premières tâches expérimentales. Nos résultats montrent que les participants confrontés d'emblée à des conditions exigeantes adoptent rapidement une stratégie de délégation au modèle prédictif, qui persiste même lorsque les conditions deviennent plus favorables. À l'inverse, les participants débutant par une tâche facile sous faible pression tendent à avoir une reliance et sur-reliance plus faible sur l'ensemble de l'expérience. L'analyse des effets d'exposition antérieure révèle une asymétrie : les conditions exigeantes constituent des ancres cognitives plus puissantes, la pression temporelle agissant sur la confiance et la sur-reliance, la difficulté sur la charge de travail et les temps de décision. Ces résultats impliquent que l'ordre d'exposition aux conditions de travail avec un système d'assistance par IA n'est pas anodin. Une introduction progressive, débutant par des conditions peu contraignantes, pourrait permettre aux opérateurs de calibrer leur confiance à sa juste valeur avant d'être confrontés à des situations exigeantes. Cette observation pourrait être pertinente tant dans un cadre éducatif que pour le déploiement opérationnel de systèmes d'assistance par IA. Par ailleurs, une analyse de la variabilité interindividuelle montre que dans le groupe n'ayant pas accès à l'assistance de l'IA, l'âge est corrélé négativement à la performance et positivement à la charge de travail. Dans le groupe avec l'assistance (H+IA), ces effets ne sont pas observés, ce qui suggère que l'IA permet de gommer une baisse de performance liée à l'âge. En revanche, nous observons dans le groupe ayant accès à l'assistance une charge de travail reportée plus élevée chez les femmes. Des travaux futurs devront confirmer ces résultats à l'aide de protocoles dédiés et de tâches plus représentatives de contextes métiers réels.

## Remerciements

Ce travail est soutenu par le programme de recherche et d'innovation HORIZON de l'Union européenne, convention de subvention n° 101120657, projet ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI). Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources AD011011309R5 attribuée par GENCI.

# Références


[1] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas : Evidence From a Large, Dynamic Experiment. In *Proceedings of the ACM Collective Intelligence Conference*, CI '25, pages 198–213, New York, NY, USA, August 2025. Association for Computing Machinery.

[2] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. Belief Updating and Delegation in Multi-Task Human-AI Interaction : Evidence from Controlled Simulations, February 2026. arXiv :2602.01986 [cs].

[3] R. J. M. Bruls and R. M. Kwee. Workload for radiologists during on-call hours : dramatic increase in the past 15 years. *Insights into Imaging*, 11(1) :121, November 2020.

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think : Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1) :1–21, April 2021. arXiv :2102.09692 [cs].

[5] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2) :277 :1–277 :26, October 2023.

[6] Lemuria Carter and Dapeng Liu. How was my performance ? Exploring the role of anchoring bias in AI-assisted decision making. *International Journal of Information Management*, 82 :102875, June 2025.

[7] Tirtha Chanda, Katja Hauser, Sarah Hobelsberger, Tabea-Clara Bucher, Carina Nogueira Garcia, Christoph Wies, Harald Kittler, Philipp Tschandl, Cristian Navarrete-Dechent, Sebastian Podlipnik, Emmanouil Chousakos, Iva Crnaric, Jovana Majstorovic, Linda Alhajwan, Tanya Foreman, Sandra Peternel, Sergei Sarap, İrem Özdemir, Raymond L. Barnhill, Mar Llamas-Velasco, Gabriela Poch, Sören Korsing, Wiebke Sondermann, Frank Friedrich Gellrich, Markus V. Heppt, Michael Erdmann, Sebastian Haferkamp, Konstantin Drexler, Matthias Goebeler, Bastian Schilling, Jochen S. Utikal, Kamran Ghoreschi, Stefan Fröhling, Eva Krieghoff-Henning, and Titus J. Brinker. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, 15(1) :524, January 2024.

[8] M L Cummings. Automation Bias in Intelligent Time Critical Decision Support Systems. *Decision making in aviation*, 2004.

[9] Mary Lynne Dittmar, Joel S. Warm, William N. Dember, and David F. Ricks. Sex Differences in Vigilance Performance and Perceived Workload. *The Journal of General Psychology*, 120(3) :309–322, July 1993. _eprint : https ://doi.org/10.1080/00221309.1993.9711150.

[10] Nuša Farič, Sue Hinder, Robin Williams, Rishi Ramaesh, Miguel O Bernabeu, Edwin Van Beek, and Kathrin Cresswell. Early experiences of integrating an artificial intelligence-based diagnostic decision support system into radiology settings : a qualitative study. *Journal of the American Medical Informatics Association*, 31(1) :24–34, December 2023.

[11] Alicia Freel, Sabid Bin Habib Pias, Selma Šabanović, and Apu Kapadia. How Misclassification Severity and Timing Influence User Trust in AI Image Classification : User Perceptions of High- and Low-Stakes Contexts. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 2906–2923, New York, NY, USA, June 2025. Association for Computing Machinery.

[12] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1) :35–42, February 2011.

[13] Nikita Haduong and Noah A. Smith. How Performance Pressure Influences AI-Assisted Decision Making, February 2025. arXiv :2410.16560 [cs].

[14] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index) : Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati, editors, *Advances in Psychology*, volume 52 of *Human Mental Workload*, pages 139–183. North-Holland, January 1988.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[16] Lukas Hermanns and Timm Teubner. Under pressure : how time constraints, task complexity, and AI reliability shape human-AI interaction. *Behaviour & Information Technology*, 0(0) :1–25, December 2025. _eprint : https ://doi.org/10.1080/0144929X.2025.2587732.

[17] Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Ilia Sucholutsky, Andrew Strait, Q. Vera Liao, and Umang Bhatt. Measuring and mitigating overreliance is necessary for building human-compatible AI, September 2025. arXiv :2509.08010 [cs].

[18] John D. Lee and Katrina A. See. Trust in Automation : Designing for Appropriate Reliance. *Human Factors*, 46(1) :50–80, March 2004.

[19] Jules Leguy, Pierre-Antoine Jean, Felipe Torres Figueroa, and Sébastien Harispe. WebXAII : an open-source web framework to study human-XAI interaction, May 2025.

[20] Michael Mayer. Trusting machine intelligence : artificial intelligence and human-autonomy teaming in military operations. *Defense & Security Analysis*, 39(4) :521–538, October 2023. _eprint : https ://doi.org/10.1080/14751798.2023.2264070.

[21] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, pages 340–350, New York, NY, USA, April 2021. Association for Computing Machinery.

[22] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding Fast and Slow : The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1) :83 :1–83 :22, April 2022.

[23] Stephen Rice, David Keller, Gayle Hunt, and David Trafimow. Automation Dependency Under Time Pressure. *2009 International Symposium on Aviation Psychology*, pages 611–616, January 2009.

[24] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your AI : expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 240–251, Marina del Ray California, March 2019. ACM.

[25] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. Trust and Reliance in XAI – Distinguishing Between Attitudinal and Behavioral Measures, March 2022. arXiv :2203.12318 [cs].

[26] Sonia Jawaid Shaikh and Ignacio F. Cruz. AI in human teams : effects on technology use, members' interactions, and creative performance under time scarcity. *AI & SOCIETY*, 38(4) :1587–1600, August 2023.

[27] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 138–154, New York, NY, USA, April 2024. Association for Computing Machinery.

[28] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The effects of explanations on automation bias. *Artificial Intelligence*, 322 :103952, September 2023.

[29] Ruoxin Yang, Sisheng Li, Yawei Qi, Jiali Liu, Qinghua He, and Haichao Zhao. Unveiling users' algorithm trust : The role of task objectivity, time pressure, and cognitive load. *Computers in Human Behavior Reports*, 18 :100667, May 2025.

[30] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, New York, NY, USA, January 2020. Association for Computing Machinery.

[31] Zelun Tony Zhang, Felicitas Buchner, Yuanting Liu, and Andreas Butz. You Can Only Verify When You Know the Answer : Feature-Based Explanations Reduce Overreliance on AI for Easy Decisions, but Not for Hard Ones. In *Proceedings of Mensch und Computer 2024*, pages 156–170, Karlsruhe Germany, September 2024. ACM.