Acceptabilité de l'Intelligence Artificielle en contexte professionnel : facteurs d'influence et méthodologies d'évaluation

Alexandre Agossah

COLLEGE DOCTORAL PAYS DE LA LOIRE — ED MATHEMATIQUES STIC

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Alexandre AGOSSAH

Acceptabilité de l'Intelligence Artificielle en contexte professionnel

Facteurs d'influence et méthodologies d'évaluation

Thèse présentée et soutenue à Nantes, le 09 octobre 2024
Unité de recherche : UMR 6004 - LS2N

Composition du jury :

Rapporteurs avant soutenance :

Pascal SALEMBIER	Professeur des universités, Université de Technologie de Troyes
Marco WINCKLER	Professeur des universités, Université Côte d'Azur

Examinateurs :

Président :	Marc-Eric BOBILLIER CHAUMON	Professeur des universités, CRTD CNAM Paris
Examinateurs :	Yannick PRIÉ	Professeur des universités, Nantes Université
	Pierre THEROUANNE	Maître de Conférences, Université Côte d'Azur

Equipe encadrante :

Dir. de thèse :	Patrick LE CALLET	Professeur des universités, Nantes Université
Co-encadrant :	Matthieu PERREIRA DA SILVA	Maître de conférences, Nantes Université
Co-encadrante :	Frédérique KRUPA	Enseignante-Chercheuse, L'École de Design Nantes Atlantique

Invité :

Frédéric TAESCH	Responsable technique, SIGMA Informatique

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à mon directeur de thèse, Patrick, et à mes co-encadrants, Frédérique et Matthieu, pour leur encadrement, leurs précieux conseils et leur soutien tout au long de ces trois années de recherche. Je les remercie également d'avoir cru en moi, même quand je doutais le plus.
Je remercie chaleureusement l'ensemble des membres de mon jury, dont Marc-Eric, ainsi que mes rapporteurs, Marco et Pascal, et mes membres de CSI, Pierre et Yannick.
Un grand merci à mes collègues de l'équipe IPI et de l'EDNA pour les nombreux échanges emplis de bienveillance.
Merci à SIGMA Informatique de m'avoir accueilli pendant ces trois années et permis de développer mes recherches sur un sujet si passionnant. Une mention toute particulière à Guillaume, qui m'a accompagné à merveille, à en devenir un ami et un modèle dont j'espère pouvoir suivre les traces.
Á mes amis, en particulier JP, Sharon, Aurélien, Mathilde et Marius. Si j'en suis là, c'est aussi parce qu'ils ont décelé chez moi ce potentiel d'accomplir des choses que je n'aurais jamais cru possibles. Á Léa, ces quelques mots pour lui dire merci de croire en moi, de me pousser à donner le meilleur et surtout de me supporter.
Je remercie enfin ma famille, en particulier mes parents, pour leur amour, leur patience, et leur encouragement sans faille tout au long de mes études. Leur soutien a été essentiel à la réalisation de ce projet. Ils m'ont toujours donné tout ce qu'il me fallait et même plus, j'espère donc que l'aboutissement de mon travail les rendra fiers. Á ma sœur Sabine, qui a su m'écouter, me réconforter dans les moments les plus difficiles. Et à Cédric, mon frère, mon modèle, sans qui je n'aurai sûrement jamais accompli toutes ces choses, car il m'a toujours inspiré et poussé à rêver plus haut et plus loin.
Je conclurai en rappelant que cette thèse est l'aboutissement de mon travail certes, mais aussi du soutien de toutes ces personnes. Á l'issue de ces trois années et demi de doctorat, je leur adresse donc mes plus sincères et chaleureux remerciements.

TABLE DES MATIÈRES

GLOSSAIRE

Acceptabilité Degré auquel des utilisateurs identifiés désirent utiliser un artefact. Souvent représentée par l'intention d'usage.

Acceptabilité à priori Jugement anticipé par lequel une personne accepte un artefact avant d'avoir une expérience directe d'usage avec ce dernier.

Acceptabilité pratique Capacité d'un artefact à être désirable par des utilisateurs cibles dans un contexte d'interaction donné, sur la base de ses caractéristiques.

Acceptabilité sociale Intention d'usage d'un artefact conditionné par des facteurs psychosociaux et contextuels.

Acceptation Perceptions et attitudes des utilisateurs cibles à l'égard d'un artefact à la suite d'une expérience d'usage.

Adoption Utilisation effective d'un artefact par les utilisateurs cibles, qui intègrent l'artefact dans leurs activités.

Aide à la décision Outils techniques utilisés pour évaluer des options complexes afin d'aider l'utilisateur à prendre une décision finale. Ils peuvent inclure des systèmes informatisés, des modèles mathématiques et des analyses statistiques.

Anthropomorphisme Attribution de caractéristiques humaines à des entités qui ne le sont pas.

Appropriation Adoption volontaire d'un artefact, conduisant à adapter son usage et à l'élargir à d'autres environnements.

Blackbox Un modèle IA dont les mécanismes internes ne sont pas compréhensibles ou perceptibles par sa complexité, ou encore accessible en raison d'un droit de propriété, etc.

Biais d'un modèle IA Apparition de préjugés dans les données d'entraînement qui affectent le comportement de l'algorithme.h

Chatbot Un programme informatique conçu pour simuler une conversation avec des utilisateurs humains.

Confiance Sentiment ou croyance qu'une personne ou un artefact est fiable, bon, honnête et efficace.

Confiance raisonnée Confiance justifiée par des arguments tangibles ou des expériences passées démontrant la fiabilité d'un individu, d'un artefact.

Confiance non raisonnée Confiance non justifiée factuellement, basée sur l'intuition ou l'émotion plutôt que sur la logique ou l'expérience.

Deep Learning Sous-ensemble du Machine Learning, mobilise des réseaux de neurones profonds pour créer des architectures informatiques complexes qui apprennent à partir de grandes quantités de données.

Expérience utilisateur (UX) Perception et réaction d'une personne en réponse à l'interaction avec un artefact. L'UX évolue de manière longitudinale par rapport à l'interaction.

Ergonomie Science de la conception de l'environnement de travail, des produits et des systèmes pour qu'ils s'adaptent aux personnes qui les utilisent.

Ethnographie Méthode d'étude qui implique une enquête de terrain à l'aide de méthodes de collecte d'informations et de description de faits humains pertinents dans leur environnement naturel.

Éthique Système de principes moraux qui guide le comportement humain. En informatique, cela peut se référer aux implications morales et sociales de la technologie.

Facteurs Humains Étude des interactions Humains-Machine dans un environnement donné. Les Facteurs Humains désignent aussi bien les comportements, que les capacités et les caractéristiques individuelles en lien avec l'interaction avec un artefact.

Interactions Humain-Machine (IHM) Discipline qui étudie l'ensemble des dispositifs et technologies servant à un humain à interagir avec un artefact, tels qu'un système informatique. L'objectif principal des IHM est de rendre ces interactions les plus intuitives, efficaces et agréables possible pour l'utilisateur en impactant l'artefact par son interface utilisateur, son ergonomie, son accessibilité ou encore son expérience utilisateur.

Interprétabilité Capacité d'un modèle IA à présenter de manière claire et compréhensible les raisons qui sous-tendent ses décisions ou ses prédictions.

Intelligence artificielle Discipline de l'informatique qui vise à créer des machines capables de réaliser des tâches normalement attribuées à un humain, telles que l'apprentissage, la planification et la compréhension du langage naturel.

Intelligence Artificielle Explicable (XAI) Sous-domaine de l'IA qui se concentre sur la création de systèmes d'IA qui peuvent expliquer leur fonctionnement et leurs décisions de manière compréhensible pour les humains.

Innovation Diffusion d'un artefact nouveau qui devient accessible pour permettre de résoudre un problème jusque là non résolu ou de le traiter d'une meilleure manière.

Influence sociale Degré auquel les croyances et les comportements de l'entourage d'une personne l'influence dans son interaction avec un artefact.

Machine Learning, ou apprentissage machine Sous-discipline de l'IA consistant à créer des systèmes entraîné à réaliser des tâches et résoudre des problèmes par des moyens normalement attribués à l'humain et sans instructions explicites.

Grand modèle de langage (LLM) Modèles d'apprentissage automatique spécialisés dans le traitement du langage naturel qui sont entraînés sur de grandes quantités de données textuelles pour générer une multitude de contenu de manière cohérente.

Réseaux de neurones Inspiré du cerveau humain, ce type d'algorithme correspond en IA à un ensemble inter-connecté de neurones artificiels visant à la résolution de problèmes complexes.

Traitement de langage naturel (NLP) Sous-domaine de l'intelligence artificielle qui se concentre sur la compréhension et l'expression du langage humain par des machines, en particulier pour que les machines traitent et analysent de grandes quantités de données linguistiques.

Utilisabilité Degré d'utilisation possible d'un outil par des utilisateurs cibles dans un contexte spécifique pour une tâche donnée.

INTRODUCTION

L'intelligence Artificielle (IA) est un domaine en plein essor depuis maintenant plusieurs années. Elle occupe un place de plus en plus importante dans notre vie personnelle, ainsi que dans notre vie professionnelle. A la différence d'un usage personnel où nous choisissons librement nos outils, le contexte d'entreprise fait que le choix de nos outils professionnels revient le plus souvent à la hiérarchie pour accomplir une tâche spécifique. Lors du choix, de la conception et/ou du déploiement des outils professionnels, les employés semblent peu consultés et inclus dans ces étapes alors qu'ils sont les premiers destinataires de ces solutions. Les solutions informatique intégrant des modèles IA n'échappent pas à la règle, et amènent des questionnements à la fois similaires à tous types d'Interactions Humain-Machine (IHM), mais aussi des préoccupations plus spécifiques. L'ingéniosité des algorithmes et la disponibilité des données permettent de créer des outils très performants, mais des freins à l'usage se font sentir lorsque la solution n'a pas suffisamment pris en compte les besoins spécifiques, les représentations et les préférences de l'utilisateur. Ce sont pourtant des critères nécessaires à prendre en compte pour s'assurer de l'acceptabilité de la solution.

L'acceptabilité désigne, dans nos travaux, le degré d'intention d'usage d'un dispositif pour un utilisateur-cible. Elle renvoie autant aux facteurs psychosociaux qui conduisent l'utilisateur à vouloir utiliser la solution, qu'aux processus organisationnels mis en place pour s'assurer qu'elle intéressera l'utilisateur, ainsi qu'aux caractéristiques intrinsèques à l'outil qui assure son utilisabilité, son utilité ou encore sa fiabilité.

Pour la communauté IHM, l'acceptabilité est la première étape du processus d'adoption d'une technologique. C'est une attitude qui se construit à partir de la représentation que nous nous faisons d'une solution, mais a priori du premier usage ou de manière précoce. L'idée que nous nous faisons de l'outil influe sur notre intention de nous en servir ou non. Travailler sur l'acceptabilité de solutions IA en amont du déploiement de l'outil est indispensable pour limiter 1) un manque de confiance des utilisateurs envers la solution, 2) une résistance à l'adoption face à des habitudes de travail ou à la crainte de perdre le

contrôle sur la réalisation de tâche, et 3) un mauvais usage de la solution qui pourrait entraîner des erreurs.

Pour l'industrie, rendre une solution IA acceptable est également une préoccupation réelle. Pour autant une vision récurrente dans la conception de solutions IA semble être que si l'outil est performant alors il sera forcément acceptable. Cette pensée et les contraintes financières qui l'entoure poussent à privilégier une approche technocentrée, c'est-à-dire de privilégier l'efficacité et l'efficience d'un socle technique parfois au détriment de la prise en compte des exigences des utilisateurs.

Contexte de la thèse CIFRE

Cette thèse a été réalisée en collaboration avec SIGMA Informatique, une ESN [1] spécialisée dans 1) l'édition de logiciels, 2) l'intégration de solutions digitales sur mesure, 3) l'externalisation de systèmes d'information et 4) les solutions *cloud*. Dans ses domaines d'expertise, l'IA s'impose comme une problématique stratégique, dont il faut s'emparer pour en faire axe de développement qui accroît la performance grâce à l'exploitation de la donnée. Au-delà des aspects novateurs et techniques que l'IA permettrait à SIGMA Informatique de proposer à ses clients, l'ESN s'interroge également sur le type de transformation entraînée par ces outils sur l'activité des collaborateurs. Le déploiement de solutions intégrant de l'IA serait susceptible d'entraîner des transformations dans le poste des employés qui voient arriver ces solutions. Cet axe de réflexion a conduit au lancement de ce projet de thèse CIFRE [2] pour étudier les vecteurs d'adoption de l'IA et les intégrer dans les processus de conception internes.

L'objectif initial est d'évaluer l'impact des solutions IA sur l'activité, au travers d'une méthodologie généralisable et facilement applicable. Et ce, afin de mieux orienter la conception et de permettre une intégration plus sereine de ces outils au sein des métiers. Les besoins exprimés par SIGMA Informatique au lancement de la thèse sont orientés sur deux axes. Le premier est de bénéficier d'éléments permettant de mieux comprendre les utilisateurs finaux de ces systèmes et l'impact estimé de leur acceptabilité sur la pérennité de l'outil. Le second axe consiste à proposer une méthodologie de conception et d'évaluation

1. Entreprise de Services du Numérique
2. Convention Industrielle de Formation par la REcherche

de l'acceptabilité de solutions IA qui soit compréhensible et exploitable opérationnellement.

Le projet de thèse est également le point de départ d'une collaboration plus large entre SIGMA Informatique, le LS2N [3] de Nantes Université et le Digital Design Lab de l'EDNA [4] : le PII [5] – IA & Acceptabilité. Avec le soutien du RFI-OIC [6], ce programme, lancé en septembre 2021, incarne un consortium recherche-innovation pluridisciplinaire qui réfléchit à trois principales questions de recherche :

1. Jusqu'où l'IA peut-elle aider l'humain dans ses décisions en contexte professionnel ? Jusqu'où faut-il pousser la part de charge de travail confié à l'IA ?
2. Quels modes de collaboration humain / machine sont envisageables avec l'IA ?
3. Qu'est-ce qui est acceptable individuellement et collectivement en termes d'IA sur les situations de travail ?

Objectifs de la thèse

L'objectif de cette thèse est d'identifier les déterminants de l'acceptabilité des solutions IA en contexte professionnel. Ces travaux de thèse se concentrent sur certains des défis actuels tels que la prise en compte des utilisateurs dans la conception de solutions IA qui, nous supposons, vont venir bousculer les habitudes de travail. Plus précisément, nous avons défini et exploré quatre questions de recherche qui contribuent à mieux comprendre et mesurer l'acceptabilité des solutions IA par les collaborateurs :

1. Quelle est la perception de l'IA en contexte professionnel ?
2. Quelles méthodes sont employées et employables pour mesurer de la confiance en des solutions IA ?
3. Quel est l'impact de la transparence sur la collaboration Humain-IA ?
4. Quelles sont les typologies de collaboration Humain-IA en contexte professionnel ?

3. Laboratoire des Sciences du Numérique de Nantes
4. L'École de Design Nantes Atlantique
5. Programme Interdisciplinaire d'Innovation
6. Recherche, Formation et Innovation en Pays de la Loire – Ouest Industries Créatives

Positionnement

Dans ces travaux, nous considérons l'acceptabilité des solutions IA en contexte professionnel comme le degré d'intention d'usage d'outils intégrant des modèles IA dans les environnements professionnels. Comme nous allons le voir au fur et à mesure de cette thèse, cette intention d'usage semble grandement impactée par la perception que les opérateurs humains ont de la solution IA, ainsi que par la confiance qu'ils lui accordent. C'est en tout cas le constat que nous faisons de l'exploration des terrains mis à disposition par SIGMA Informatique, desquels nous présentons une méthodologie de mesure de la confiance dans les solutions IA. Une partie des méthodes mobilisées sont initialement utilisées dans le domaine de l'évaluation subjective de la qualité d'expérience (QoE). Notre méthodologie est portée sur l'évaluation de la confiance dans les solutions IA. Avec l'apparition des outils génératifs, qui bouleverse la collaboration Humain-IA telle que nous la percevions initialement, cette méthodologie prend également sens pour sélectionner de manière plus adéquate un socle technique qui permettrait de concevoir un outil génératif acceptable pour les utilisateurs finaux de ces systèmes.

Plan de la thèse

Le document se décompose en trois parties dont les lectures sont complémentaires. La partie I se focalise sur un état de l'art autour de l'acceptabilité des solutions IA en contexte professionnel et la prise en compte de l'expérience utilisateur dans leur conception. La partie II retrace nos contributions terrains que ce soit en explorant ceux mis à disposition par SIGMA Informatique et ses partenaires ou en expérimentant en laboratoire sur de potentiels déterminants de l'acceptabilité pour l'aide à la décision. La partie III s'intéresse à l'émergence des outils génératifs. Dans celle-ci laquelle nous explorons comment ces outils sont susceptibles d'impacter la manière dont nous travaillons. Pour résumer, cette thèse présente des travaux autour de l'interaction Humain-IA en contexte professionnel. Les contributions de ces travaux sont :

— 1) La mise en avant de la confiance de l'opérateur humain comme principal déterminant d'acceptabilité de l'IA ;
— 2) Une méthodologie de sélection de socle technique basée sur les préférences des utilisateurs.

Maintenant que nous avons une vue globale de l'organisation de cette thèse (figure 1),

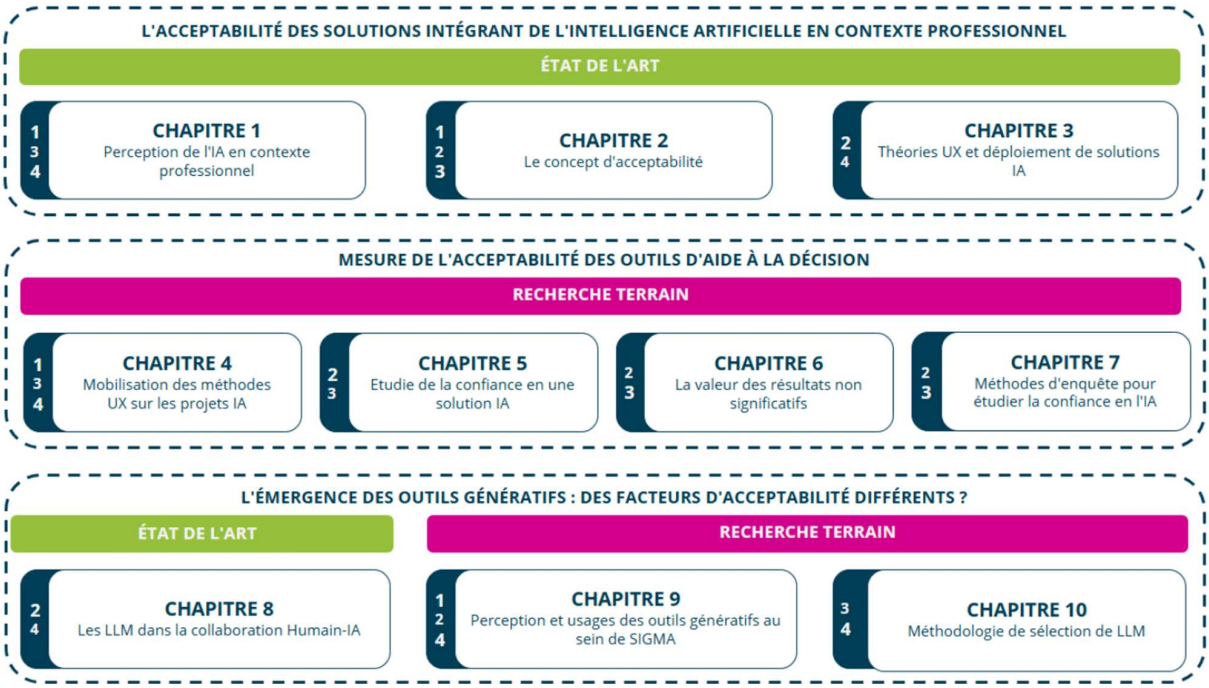


FIGURE 1 – Plan de la thèse.
Les intitulés sont des résumés des titres des chapitres et les chiffres à côté des chapitres correspondent au numéro des questions de recherche auxquels ils répondent (voir introduction page 17).

nous pouvons explorer les détails de chacun des chapitres :

- Le **chapitre 1** explore notre état de l'art sur la perception de l'IA en contexte professionnel. Nous y abordons également comment l'IA vient bouleverser la manière dont les employés perçoivent leur travail.
- Le **chapitre 2** se focalise sur le concept d'acceptabilité, comment elle est définie et modélisée dans la littérature. Nous nous intéressons également à comment elle évolue dans le cas particulier de l'IA.
- Le **chapitre 3** s'intéresse à la notion d'expérience utilisateur et la place qu'elle occupe dans la conception de solutions IA.
- Le **chapitre 4** montre comment nous avons mis en application les connaissances acquises dans les chapitres précédents pour explorer les terrains mis à disposition par SIGMA Informatique et ses partenaires au moyen d'une démarche ethnographique.
- Le **chapitre 5** retrace une étude réalisée en laboratoire qui s'intéresse à l'effet de la confiance déclarée par un outil prédictif sur la confiance que l'utilisateur lui accorde. Ce chapitre est également l'occasion d'explorer des méthodes alternatives

d'analyse de données pour étudier l'effet de la transparence sur la confiance en milieu contrôlé.

— Le **chapitre 6** est complémentaire au chapitre 5, puisqu'il présente les données et méthodes qui ont été utilisées, dont les résultats statistiques n'étaient pas significatifs. Mes ces résultats restent utiles pour comprendre comment s'intéresser à la confiance des opérateurs humains envers des solutions IA.

— Le **chapitre 7** finalise la présentation de l'étude amorcée dans les chapitres 5 et 6, en mettant en avant des méthodes d'enquête pour compléter les connaissances acquises.

— Dans le **chapitre 8**, nous explorons un nouvel état de l'art autour des LLM [7] et la manière dont ils sont perçus en contexte professionnel. Nous examinons notamment l'impact potentiel de ces outils sur le travail.

— Dans le **chapitre 9**, nous présentons une enquête qui s'intéresse à la perception des employés à l'égard des outils génératifs dans leur travail. Ce chapitre est complété par une étude sur l'apport des outils génératifs sur la performance et le bien-être de développeurs Java.

— Le **dixième et dernier chapitre** retrace la mise en place d'une méthodologie de sélection de LLM, respectant un cadre de sécurisation de la données, et basée sur les préférences des utilisateurs finaux.

7. Large Language Models, voir chapitre 8

Première partie

L'acceptabilité de solutions intégrant de l'Intelligence Artificielle en contexte professionnel

---

Chapitre 1

Perception de l'Intelligence Artificielle en contexte professionnel

Dans ce chapitre

Ce chapitre présente un état de l'art sur la perception de l'IA en contexte professionnel. Nous nous intéressons à l'arrivée des solutions IA auprès des postes de travail, ainsi qu'aux enjeux individuels et collectifs liés à leur conception. Pour y faire face, il est nécessaire de faciliter la compréhension du fonctionnement des solutions IA en les rendant plus transparentes. Plus une solution IA est transparente, plus l'opérateur la comprend et est susceptible de lui faire confiance.

1.1 Introduction

Offrant la possibilité d'automatiser une partie ou l'entièreté de certaines tâches, l'IA devient de plus en plus présente en contexte professionnel. Pourtant son arrivée en entreprise est freinée par de multiples inquiétudes. En termes de conception, un vrai défi se dresse dans le respect de bonnes conduites en termes d'exploitation des données pour alimenter une démarche de conception jugée éthique. Notamment face à des sujets de controverse comme l'exploitation massive de données dans l'affaire *Cambridge Analytica* [87]. Un autre sujet dominant est la responsabilité à attribuer aux décisions prises par ou avec ces outils. De plus, les employés, qui sont les principaux concernées par le déploiement des solutions IA, voient leurs perceptions et craintes souvent négligées.

Bien que l'IA soit perçue par les organisations comme un moyen de gagner en productivité, il arrive qu'elle rebute notamment pour la place que ces technologies vont occuper par rapport aux opérateurs humains en poste. Dans ce chapitre, nous proposerons une analyse détaillée des typologies de perception de l'IA, ainsi que de l'impact et des défis que l'IA représente pour l'avenir du travail.

1.2 Les promesses de l'IA auprès du monde professionnel

Les investissements en matière d'IA opérés en milieu professionnel visent majoritairement à rehausser la performance, à enrichir la qualité du rendement, et à adresser des besoins sociaux émergents [59]. Toutefois, l'adoption parfois partielle, voire le rejet, de ces outils, sont souvent attribués à une négligence des besoins, contraintes et représentations des utilisateurs. Ce qui témoigne d'une dissonance entre ce que l'organisation veut faire et ce dont les collaborateurs ont besoin. Cela soulève divers questionnements sur la manière d'appréhender l'univers de l'IA, dont la tendance de ces dernières années est à l'optimisation des tâches.

Initialement, les solutions IA étaient présentées comme des technologies permettant aux systèmes informatiques de répliquer le comportement humain [51]. Et plus impressionnant encore, à apprendre à exécuter une tâche sans nécessiter de programmation explicite [150]. Cependant, l'IA se redirige aujourd'hui vers la création de systèmes automatisés aptes à résoudre des problématiques complexes, et pas nécessairement comme les aurait réalisées un opérateur humain [81]. Dans le contexte professionnel, les solutions IA continuent d'incarner des systèmes informatiques permettant de résoudre des tâches coûteuses en temps et en effort pour l'humain. La valeur des tâches que ces outils doivent accomplir est donc de plus en plus discutée, bien qu'ils soient très prisés pour des tâches à faible plus-value. L'IA est donc davantage présente dans les discussions sur la prise en charge des tâches laborieuses, automatisables et répétitives, libérant ainsi les employés pour des missions plus enrichissantes et porteuses de sens pour eux [82] [59] [66].

Cette tendance à l'automatisation semble être alimentée par la perception qu'entretiennent les entreprises à l'égard de l'IA. Ces perceptions sont majoritairement positives, bien qu'elles fluctuent en fonction des besoins et inquiétudes de chacunes. Elles y voient

notamment un levier pour booster leur performance, accélérer et fiabiliser leur prise de décision, ainsi qu'à revisiter leur activité et adopter une gestion du travail et du management des compétences de manière plus innovante. Cette tendance se reflète bien dans l'étude commanditée par Forbes en 2020, qui démontre que la plupart des grandes entreprises et ETI [1] sont enclines à élever leurs investissements dans l'IA, y percevant un vecteur significatif de croissance [123]. Cette étude décrit aussi ces solutions comme un moyen efficace de mieux optimiser les performances globales et de diminuer les erreurs humaines.

Le développement de l'IA en contexte professionnel vise également à proposer un certain confort d'assistance pour les opérateurs humains face aux tâches coûteuses [27]. La cohorte AI4People, qui rassemble différentes parties prenantes [2], estime que l'IA représente une niche d'opportunités pour les employés que nous pouvons répartir en 4 axes [63] :

— Un soutien à l'épanouissement personnel, en accompagnant le collaborateur en fonction de ses caractéristiques, ses intérêts, ses compétences et ses projets de vie.

— Un moyen d'améliorer l'agentivité du collaborateur, c'est-à-dire la mise en place d'une collaboration Humain-IA qui lui permet de faire plus, mieux et plus vite dans son environnement de travail.

— Un moyen d'améliorer les capacités sociales individuelles et collectives dans leur ensemble.

— Une culture de la cohésion sociale en amenant les outils à contribuer à résoudre des problèmes complexes de coordination entre les collaborateurs.

Il nous reste à préciser que malgré ces promesses plutôt positives pour les employés, des appréhensions subsistent. D'un côté, les entreprises s'inquiètent de conserver la souveraineté de leurs données et de veiller à la responsabilité des décisions prises par ou avec l'IA. De l'autre côté, les employés semblent manifester davantage d'inquiétudes quant à l'impact de l'IA sur leur emploi, redoutant la disparition de certains métiers au profit d'une automatisation croissante [59]. De plus, des défis demeurent en termes de formation et de compétences nécessaires pour manœuvrer les solutions IA.

1. Entreprises de Taille Intermédiaire
2. Des organisations académiques, industrielles et de la société civile

1.3 Entre opportunités et défis, la perception évolutive des employés à l'égard de l'IA en entreprise

Depuis 2018, de multiples études prospectives sur la transformation des emplois par l'IA ont été réalisées. Au fur et mesure des années, elles indiquent des chiffres en hausse quant aux nombres d'emplois concernés. L'IA semble être devenue un vrai sujet de questionnement, là où l'ingéniosité technologique surplombe l'actuelle réglementation en matière d'usage et de protection des données. En 2019, l'OCDE [3] annonçait que l'IA et la robotique seraient responsables de la mutation d'environ 14% des emplois de ses pays membres dans les années qui allaient suivre (toutes professions confondues) [135]. Ce chiffre serait passé à 18% depuis 2023 avec l'arrivée des outils génératifs sur le marché du travail, d'après une étude de Goldman Sachs, qui estime également qu'un quart des emplois aux États-Unis et en Europe pourraient être automatisés [75]. Toujours d'après cette étude, certains secteurs seront plus touchés que d'autres : les professions administratives et juridiques seront les plus impactées avec des suppressions de postes pouvant atteindre respectivement 46% et 44% contre environ 6% à 4% pour les métiers les plus physiques. Ces prévisions ne sont pas sans impact, accentuant, chez certains, la crainte de se voir concernés par ces remplacements.

Si ces chiffres sont un premier indice des facteurs d'influence de la perception de l'IA au travail, nous pouvons également nous référer aux travaux de l'association Impact IA qui agit autour de deux objectifs : 1) la prise en compte d'enjeux éthiques et sociétaux dans la conception de solutions IA et 2) le soutien de projets technologiques innovants. L'association a commandé pendant trois années de suite (de 2018 à 2020 inclus) une enquête sur la perception de l'IA en entreprise. Ce travail d'enquête a été réalisé auprès d'un peu plus de 500 salariés sur chacune des trois années, offrant une vision panoramique de la perception de l'IA dans le monde professionnel français. Les principaux résultats de cette étude (voir figure 1.2) montrent que la perception de l'IA et la confiance que lui accordent les salariés français sont relativement élevées en 2018 et légèrement en hausse depuis 2019. Pourtant seulement près d'un cinquième des personnes interrogées ont déjà utilisé une solution IA au moins une fois au travail.

Plus de la moitié des interrogés attribuent la bonne image de l'IA à son utilité perçue, y trouvant une place nécessaire dans l'évolution du travail. La dernière enquête de Impact

3. Organisme de Coopération et de Développement Economiques

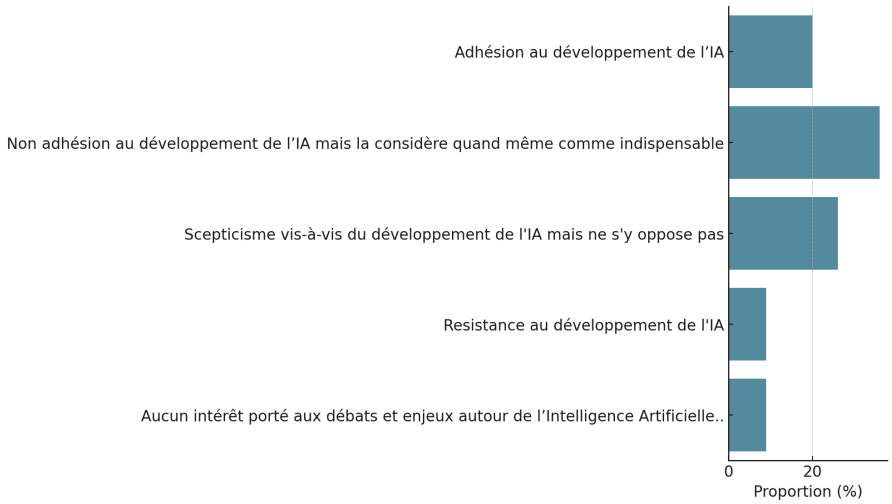


FIGURE 1.1 – Opinion globale des salariés français à l'égard du développement de l'IA en 2020, d'après Impact IA [86]

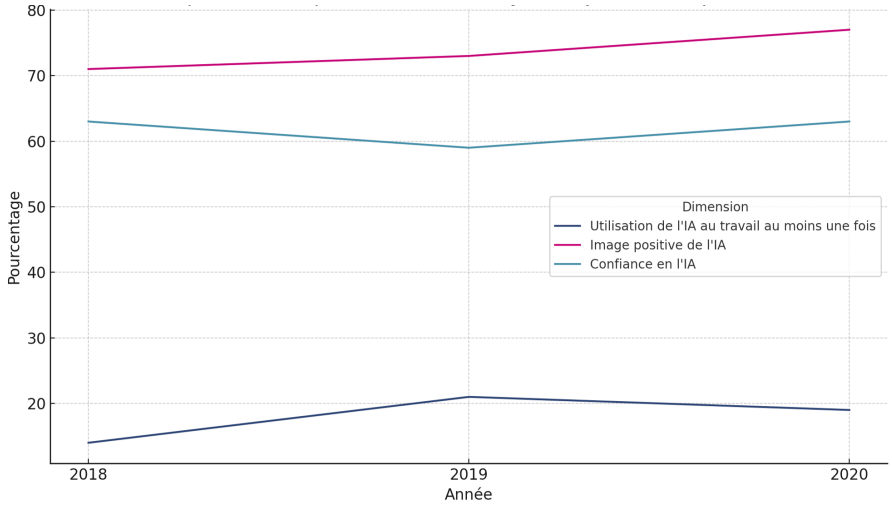


FIGURE 1.2 – Évolution de l'utilisation, de la perception et de la confiance envers l'IA par les salariés français, d'après Impact IA [86]

IA remonte à 2020, et cette année-là une exploration des secteurs d'activité plus fine a été réalisée. Le secteur de la Recherche et Développement (R&D) semble se démarquer avec près de deux tiers des répondants de ce secteur d'activité qui déclarent se servir activement de solutions IA dans leur travail. La R&D est suivie par les secteurs commerciaux, de services, de l'informatique et des nouvelles technologies et de support qui atteignent respectivement les 50%. Les secteurs associés à des métiers artisanaux, manuels ou encore à faible niveau de qualification ont cependant des scores drastiquement plus bas

en termes d'utilisation de l'IA dans leur travail. La pensée selon laquelle l'IA va améliorer la productivité professionnel est pourtant en constante hausse.

Trouver une utilité à ces outils n'ôte cependant pas la crainte de destruction des emplois. Ceci est démontré par une perception de l'impact positif de l'IA sur les performances au travail qui reste mitigée (voir figure 1.1). Seul une minorité de répondants pensent que l'IA contribuera à leur bien-être au travail ou assurera la pérennité de leur emploi. Une majorité des répondants expriment plutôt la volonté d'être davantage sensibilisés aux conséquences et applications de l'IA au travail. Ce qui accentue le besoin de communiquer sur cette discipline avec 80% des répondants, en 2020, qui ont exprimé un besoin de mieux réglementer l'IA.

Cette enquête révèle une perception nuancée de l'IA parmi les salariés français. Alors que la confiance et l'optimisme augmentent, des inquiétudes persistent quant aux risques et aux impacts sur l'emploi et la vie privée. Il en ressort un besoin croissant d'éducation et de réglementation pour naviguer dans le paysage évolutif de l'IA en entreprise. Pour autant, une majorité croissante des répondants (70% en 2020) pense que l'IA va changer le monde du travail.

Parallèlement à cette étude, des recherches ont également été menées pour comprendre la perception des employés face à cette révolution. Le sociologue Yann Ferguson étudie la représentation qu'ont les employés des solutions IA et leur futur possible dans le monde du travail [59]. Dans ce cadre, il identifie quatre principales représentations des relations que l'on peut entretenir avec les outils intelligents au travail :

— L'augmentation par l'IA : Les employés voient une opportunité d'acquérir des compétences sur des tâches à plus forte valeur ajoutée en confiant les tâches simples à l'IA.
— La ré-humanisation par l'IA : L'IA permet aux employés de se focaliser sur des tâches où ils gardent un avantage conséquent, comme la créativité ou l'intelligence émotionnelle.
— La soumission à l'IA : Certains employés, percevant l'IA comme supérieure et/ou moins sujette à l'erreur que l'humain, estiment n'avoir d'autre choix que de suivre ses prédictions.
— Le remplacement par l'IA : Cette vision traduit une crainte d'être totalement

substitué par l'IA.

Ces représentations jouent un rôle crucial dans la manière dont l'IA est acceptée et intégrée dans l'environnement professionnel. Les employés qui se voient "augmentés" par l'IA ont tendance à avoir une perception plus positive de la technologie, s'attendant à des améliorations dans leurs performances professionnelles. En revanche, ceux qui se sentent menacés par l'IA risquent d'avoir une perception négative, associée à un sentiment d'exclusion ou de danger pour leur poste.

Le déploiement de solutions IA en situation de travail aurait pourtant souvent été victime d'échecs [4] [163]. Malgré la plus-value identifiée pour les entreprises et les employés, ces échecs sont souvent attribués à un déficit d'intégration des employés dans le processus d'innovation, une perte de maîtrise du travail par les employés qui sont aliénés par la solution IA ou encore des positions contestataires [5] des employés par crainte pour leur poste de travail [26]. C'est en tout cas ce qu'a mis en avant dans le rapport d'enquête du laboratoire LaborIA [6], en 2024, qui se focalise sur l'étude des impacts de l'IA sur le travail, l'emploi, et les compétences, afin de mieux comprendre et anticiper ses effets [159]. Dans la poursuite des travaux de Yann Ferguson [59], la synthèse générale de l'enquête de LaborIA montrent qu'une réelle dissonance peut apparaître entre ce qui est prévu par l'organisation et de l'appréciation des employés qui auront un certain degré d'acceptabilité ou de rejet. Selon LaborIA, le rejet/la mise à distance des solutions IA résultent principalement de configurations humain-machine jugée "aliénantes" à cause de :

— Un excès ou important manque de confiance en la solution IA ;
— Coûts d'optimisation trop élevés ;
— Perte de compétences en se reposant trop sur la solution IA
— Perte de conscience du travail au profit d'une dépendance à l'IA pour effectuer ses tâches.

4. En 2020, les chercheurs du *Capgemini Research Institute* expliquaient que seul 13% des organisations, qui investissaient dans l'IA, réussissaient à déployer avec succès des cas d'usage en production et à continuer de déployer davantage
5. Résistance à l'utilisation/rejet de la solution IA
6. Structure initiée par le Ministère du Travail, de la Santé et des Solidarités et l'INRIA

1.4 Enjeux sociétaux : robustesse, légalité et éthique

Au cours de la dernière décennie, des communautés telles que AI4People et Impact IA se sont formées pour réfléchir à l'éthique et à l'impact sociétal de l'IA, cherchant à définir les caractéristiques d'une « bonne IA ». Ces discussions mettent en lumière l'importance de la transparence et de la compréhension des outils IA, facilitant ainsi leur acceptation et leur utilisation responsables.

En 2018, le premier forum organisé par la commission européenne pour échanger sur les impact sociétaux de l'IA, AI4People, s'est tenu en présence d'acteurs publiques et privés avec le soutien du AI HLEG [7], avec pour objectif d'apporter une compréhension plus éclairée des problématiques éthiques et juridiques liées à l'IA. Bien que des règles de recueil, de gestion et d'exploitation de la donnée soient déjà existantes notamment au travers du RGPD [8], il réside toujours des interrogations quant aux responsabilités personnelles (face à l'usage de l'outil) et à la responsabilité de l'outil-même [63]. AI4People était donc l'occasion de réunir différents acteurs du paysage de l'IA pour collaborer sur la mise en œuvre d'une stratégie commune dans la gestion de l'IA. Les recommandations qui découlent de ce forum répondent à 3 besoins autour de l'IA :

- De la robustesse - l'IA doit être techniquement et socialement sûre, sécurisée et fiable ;
- Le caractère légal de l'IA - l'IA doit respecter les législations et réglementations applicables ;
- Le caractère éthique de l'IA - l'IA doit adhérer à des principes et valeurs éthiques.

Selon AI4People, la bonne tenue des recommandations de ces trois axes permet de valoriser une IA sociétalement de confiance [86].

Les recommandations, issues du forum, ont ensuite été raffinées, toujours en respectant les trois axes. Ceci a permis de proposer une grille de bonnes pratiques, sur la base de sept dimensions, qui guide les organisations dans leur stratégie de conception de solutions IA. Nous avons répertorié ces dernières dans le tableau 1.1

Cette grille stratégique incite les acteurs à adopter, volontairement une démarche de conception qui soutient le développement de technologies IA, favorables à générer de la confiance en minimisant les risques liés à leur usage [63] [86]. Ce type de démarche est plus que nécessaire face aux défis liés à l'expansion de l'IA dans nos vies personnelles et

7. High Level Expert Group
8. Règlement Général de Protection des Données

Action et contrôle humain	L'IA doit favoriser l'équité entre les humains, sans restreindre l'autonomie humaine.
Robustesse technique et sécurité	Les algorithmes servant à concevoir les modèles d'IA doivent être suffisamment sûrs et fiables pour gérer efficacement les éventuelles erreurs.
Respect de la vie privée et gouvernance des données	Les citoyens doivent posséder un contrôle complet sur leurs données personnelles, garantissant que ces dernières ne soient pas exploitées à leur désavantage ou pour des objectifs discriminatoires.
Transparence	les solutions IA doivent permettre une compréhension de leur fonctionnement et une traçabilité des informations à disposition du système.
Diversité, non-discrimination et équité	les solutions IA doivent prendre en compte la diversité de caractéristiques et capacités chez les utilisateurs-cibles.
Bien-être social et environnemental	les solutions IA doivent soutenir des "évolutions sociales positives et renforcer la durabilité et la responsabilité écologique".
Responsabilité	Des mécanismes doivent être établis pour assurer la responsabilité liée à l'utilisation de solutions IA et à leurs résultats.

TABLE 1.1 – Recommandations de conception IA robuste, légale et éthique par AI4People

professionnelles. Ces défis, en matière de responsabilité, de respect de la vie privée ou encore d'équité, sont des sujets abordés dans des travaux tels que la thèse de Jouis sur l'explicabilité de l'IA qui apporte des contributions significatives, en proposant des applications concrètes pour expliquer le prise de décision de modèles IA complexes, renforçant ainsi la confiance des utilisateurs [95].

L'évolution rapide des technologies IA a également mis en lumière le besoin de mieux comprendre la discipline, face à des modèles IA complexes et/ou incompréhensibles, souvent appelés *blackbox* (pour boîtes noires). Ces modèles sont généralement inintelligibles, car trop complexes pour que les opérateurs humains comprennent le cheminement ayant mené à leur prise de décision, ou alors ils sont protégés par les entreprises conceptrices qui ne veulent pas laisser fuiter le fonctionnement de leur création. Il semble que ces modèles sont de moins en moins appréciés au profit de modèles plus compréhensibles et dont il est plus facile de retracer le cheminement qui a conduit à leurs résultats, on parle de modèles transparents [148] [20] [21]. Les recherches sur cette thématique sont essentielles pour naviguer dans ce paysage en constante évolution pour s'assurer que les solutions IA

développées sont bénéfiques à la société. De multiples études dans ce secteur s'axent sur l'impact de la transparence de l'IA au profit de la confiance, soulignant l'importance de l'explicabilité dans l'acceptation des outils IA [189] [56] [186].

1.5 L'enjeu de la transparence : les IA explicables (XAI)

Selon Hoc (2000), l'absence de modèle préétabli du fonctionnement d'une machine crée une dépendance vis-à-vis de la confiance ou de la méfiance envers cette dernière, pouvant mener à une utilisation aveugle ou à un contournement de la machine [82]. C'est-à-dire que si l'utilisateur a une méconnaissance du fonctionnement de l'outil, il risque de s'y fier ou de s'en méfier sur la base d'arguments illusoires. Pour pallier ce problème, Hoc suggère d'informer davantage l'opérateur humain du comportement de la machine, favorisant ainsi une meilleure compréhension et coopération humain-machine.

En IA, accroître la transparence de l'outil pour favoriser sa compréhension par l'humain passe principalement par l'explicabilité. Les modèles IA qui sont capables d'expliquer leur fonctionnement et/ou résultat sont nommés les IA explicables (XAI). Les XAI forment un domaine pluridisciplinaire qui englobe un ensemble de techniques et outils permettant d'accroître la transparence des modèles visés [21] [148] [95]. Il existe une variété de techniques en XAI, qui sont classées en deux catégories :

- Les techniques d'explications locales, qui fournissent une explication pour une décision donnée.
- Les techniques d'explications globales, qui fournissent une explication pour l'ensemble du modèle.

L'explicabilité est devenue un enjeu clé dans le domaine de l'IA, particulièrement pour les modèles IA réputés pour leur complexité. Dans sa thèse sur l'évaluation de l'explicabilité des modèles profonds, Gaelle Jouis (2023) distingue deux dimensions de l'explicabilité. D'abord, la transparence, le degré de compréhension permis par le modèle pour des personnes cibles, et ensuite la pertinence, le degré de cohérence entre les explications fournies par le modèles et le besoin de compréhension par rapport au contexte. Elle affirme que ces deux critères sont essentiels pour améliorer la compréhension et la confiance des utilisateurs [95].

Dans le même sens, Barredo Arrietta et ses collaborateurs (2020) proposent une revue de la littérature sur les XAI [21]. Ils présentent une taxonomie des XAI se basant sur deux dimensions : 1) La direction, qui mesure si l'explication est fournie du point de vue du modèle ou de l'utilisateur et 2) la granularité, qui rappelle la pertinence décrite par Jouis et qui réfère au niveau de détail de l'explication fournie. Cette taxonomie permet aux auteurs de classer les différentes techniques d'explications existantes. Ils identifient, eux, trois grandes catégories de techniques :

— Les explications locales, donc sur une décision donnée du modèle.
— Les explications globales sur l'ensemble du modèle.
— Les explications basées sur la compréhension humaine, qui fournissent des explications qui sont facilement compréhensibles par les humains.

Les XAI sont décrites dans la littérature comme offrant de nombreux avantages, tels qu'une meilleure compréhension des modèles, qui conduit à améliorer la confiance accordée par les utilisateurs et aussi à réduire les biais des modèles. Cependant, les XAI présentent également des défis. Car il n'est pas évident d'expliquer justement des modèles considérés comme très complexes avec des architectures conséquentes. De plus, pour répondre efficacement aux besoins d'explication des utilisateurs, il faut un travail supplémentaire de compréhension des besoins des personnes visées par ces explications. Et enfin, il faut prendre en compte les limitations des techniques d'explication actuelles. Barredo Arrieta et ses collaborateurs soulignent enfin que les XAI sont un domaine en plein développement, qui nécessitent des recherches supplémentaires pour en exploiter pleinement les opportunités [21].

Les recherches sur les XAI offrent des perspectives prometteuses pour le développement de systèmes IA transparents et éthiques, adaptés aux besoins de la société et capables de renforcer la confiance des utilisateurs.

1.6 La confiance comme déterminant de l'adoption

La confiance que les humains accordent aux solutions IA est un sujet devenu central dans les discussions académiques et industrielles. Établir une confiance robuste en l'IA semble complexe, en raison des incertitudes entourant les capacités et les limitations de cette technologie, ainsi que des préoccupations concernant l'exploitation des données [21] [148]. La littérature sur les interactions Humain-Machine révèle que la confiance est un mo-

dérateur crucial de l'interactivité entre les agents [90] [187]. Cependant, définir la confiance est un défi en soi, car ce concept est perçu différemment selon les domaines d'expertise. Plusieurs recherches ont noté l'absence d'une définition universelle de la confiance, mais elles s'accordent à dire qu'il existe deux composantes principales : la composante affective et la composante cognitive. L'aspect affectif de la confiance concerne les émotions et la foi en quelque chose/quelqu'un sans raison factuelle, contrairement à l'aspect cognitif qui s'ancre dans la rationalité, se focalisant sur la logique et les faits [48] [83]. Sur la base de ces définitions, même si les caractéristiques d'une solution IA indiquent qu'elle est fiable, il est toujours essentiel d'adopter une approche qui garantit sa fiabilité aussi bien sur les plans social, émotionnel et expérientiel. Ainsi l'impact des facteurs psycho-sociaux, comme l'influence de l'entourage, est crucial dans la façon dont l'outil va être perçu [182] [187]. Des chercheurs, tels que Gillath et ses collaborateurs, iront jusqu'à dire que l'aspect affectif de la confiance peut avoir un impact significativement plus important que l'aspect cognitif sur l'intention d'usage de solutions IA [67]. La perte de confiance dans les solutions IA peut avoir des conséquences profondes, affectant la perception de ces technologies mais aussi l'intention d'usage des utilisateurs. Une confiance dégradée peut entraîner de fortes réticences à utiliser l'IA, limitant son intégration efficace dans le monde professionnel. De plus, cette méfiance peut influencer les décisions d'investissement des entreprises, les amenant à se détourner de projets impliquant l'IA. Cette dynamique crée un cercle vicieux où le manque de confiance freine l'innovation, entravant ainsi le développement et l'amélioration des technologies d'IA, ce qui pourrait potentiellement renforcer cette méfiance.

Selon une enquête menée par Impact IA en 2020, les motifs de confiance et de non-confiance des salariés français vis-à-vis de l'IA révèlent des perspectives intéressantes [86]. Plus de 35% des salariés font confiance à l'IA en raison de sa capacité à résoudre des problèmes complexes, tandis qu'environ 20% attribuent leur confiance au professionnalisme des concepteurs et à la réduction des erreurs humaines. Un peu plus de 15% valorisent la manière dont les algorithmes sont conçus (voir figure 1.3). A contrario, la non-confiance se manifeste principalement à travers la crainte d'une utilisation malveillante (près de 25%), le remplacement humain (20%), les inquiétudes en matière de sécurité (20%), la peur de perdre le contrôle sur l'IA (plus de 15%), et la crainte que l'IA dépasse l'intelligence humaine (près de 10%) (voir figure 1.4). Cette dichotomie souligne l'importance d'adresser les préoccupations éthiques et sécuritaires pour renforcer la confiance en l'IA

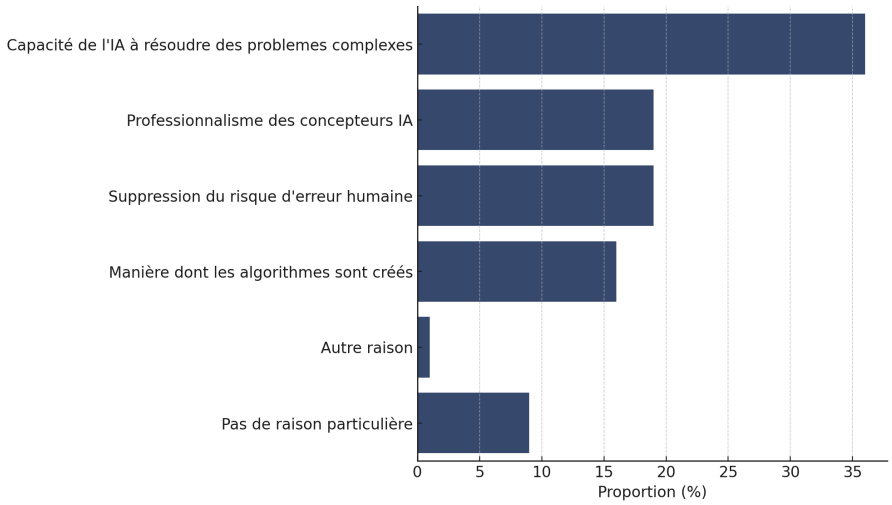


FIGURE 1.3 – Motifs de confiance en l'IA d'après les salariés français, d'après Impact IA [86]

Dans le milieu industriel, les questionnements semblent centrés sur comment assurer que la solution finale sera digne de confiance pour les utilisateurs, mais aussi pour les différentes parties prenantes au projet. Lors du webinar "IA & Confiance, quels enjeux ?" organisé par la société Sopra-Steria, Poujol et Lesaffre (2020) indiquent attribuer une place particulière à la confiance envers les projets intégrant des solutions IA. Ils considèrent la confiance comme essentielle, mais dépendante de six aspects distincts du projet :

— L'aspect technique ;
— La gestion de données ;
— L'aspect éthique, c'est-à-dire tout ce qui est relatif aux respect des droits fondamentaux, de la place de l'homme dans le processus de conception ;
— L'aspect juridique, donc tout ce qui est du ressort de la responsabilité ;
— La normalisation, qui est associée aux certifications par rapport aux lois et normes qui évoluent ;
— L'aspect sociétal, correspondant ici à l'acceptabilité de l'outil.

Poujol et Lesaffre indiquent que prendre en compte la confiance, en cherchant un équilibre entre ses différents aspects technico-sociaux, est nécessaire pour éviter l'apparition de verrous opérationnels dans l'étape de conception de solutions IA.

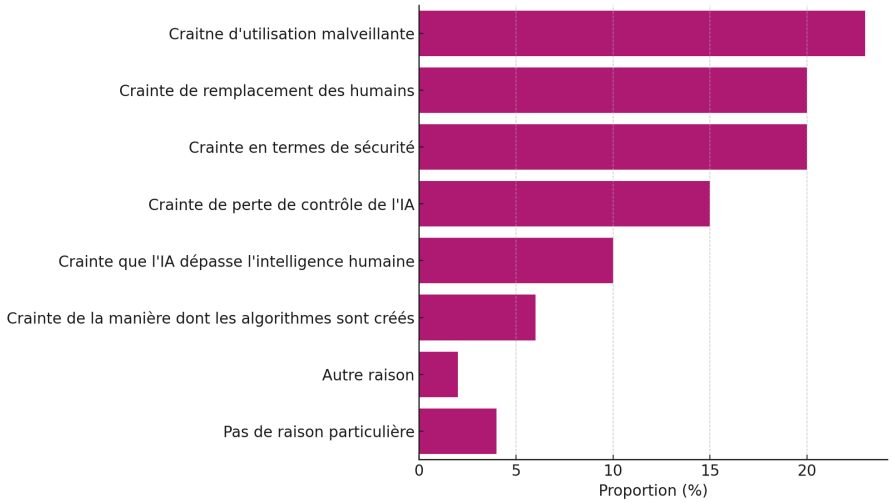


FIGURE 1.4 – Motifs de non-confiance en l'IA d'après les salariés français, d'après Impact IA [86]

1.7 Conclusion

L'IA semble ainsi présentée aux organisations comme un vecteur de productivité et d'amélioration de la qualité du travail. Pour autant, les représentations des salariés à son égard sont très disparates et certaines conduisent à un manque flagrant de confiance envers les technologies qui en découlent. C'est de ce constat que nous estimons la confiance dans les solutions IA comme un enjeu majeur, surtout que ce sujet semble en plein essor dans les milieux industriels et académiques. Sa prise en compte est variable selon le contexte mais tout porte à croire qu'elle est déterminante dans l'adoption et l'utilisation de ces technologies. La confiance semble principalement prise en compte dans la recherche académique par des prismes d'utilisabilité et d'expérience (par exemple, l'outil doit prouver qu'il peut accomplir une tâche donnée dans un contexte spécifique). Dans le contexte professionnel, les organisations s'intéressent à la confiance en l'IA comme un but à atteindre via des processus de conception orientés qui prouvent que la solution est éthique. Le point commun entre ces deux sphères est qu'elles s'accordent à penser une certaine implication de la confiance dans l'acceptabilité de l'IA. Depuis que les discussions sur la conception de l'IA de confiance se sont intensifiées, il est de plus en plus reconnu que promouvoir la confiance est essentiel pour garantir l'acceptabilité de l'IA. [158]. Ce besoin de favoriser la confiance s'est constitué parce que l'IA génère encore aujourd'hui beaucoup de craintes entre les cas de mésusages récurrents et la difficulté toujours présente de clairement définir ce qu'elle est [187] et ce qu'elle peut et ne peut pas faire [148].

Chapitre 2

Le concept d'acceptabilité dans le domaine IHM

Dans ce chapitre

Ce chapitre présente le concept d'acceptabilité, définie comme le degré d'intention d'usage d'un dispositif par les utilisateurs cibles. Selon les représentations que les utilisateurs se font d'un outil, ils le trouveront plus ou mois acceptable pour accomplir une tâche définie. Des modèles théoriques ont été élaborés pour expliquer et mesurer ce concept, le plus répandu reste le *Technology Acceptance Model* de Davis (1989). Avec l'arrivée de l'IA, il semble y avoir un besoin de reconsidérer la manière dont l'acceptabilité est abordée avec des interactions qui reposent davantage sur une relation de confiance et qui sont de plus en plus similaires à des interactions naturelles entre humains.

2.1 Introduction

Dans le domaine des IHM, l'acceptabilité joue un rôle déterminant. Elle décrit le degré auquel un artefact [1] est précocement apprécié et prêt à être adopté par l'utilisateur final [154] [129] [165]. C'est un facteur déterminant du succès du déploiement d'un artefact pour répondre à un besoin donné. De manière générale, l'acceptabilité désigne l'intention globale d'usage d'un outil ou service par les utilisateurs qui évaluent les avantages et les inconvénients potentiels à l'usage de ce dernier [52]. Au cours des années, plusieurs théories ont été formulées pour saisir et expliquer l'acceptabilité en IHM, mettant en

1. Élément artificiel permettant à l'utilisateur d'interagir avec un système ou un service

lumière différents facteurs qui peuvent influencer la manière dont un artefact est perçu et adopté par les utilisateurs.

L'acceptabilité joue également un rôle crucial dans les environnements professionnels, où selon le degré d'acceptation d'un système, les conséquences sur la productivité et l'efficacité peuvent être majeures [167]. Malgré une compréhension approfondie de l'acceptabilité pour concevoir des systèmes interactifs qui répondent aux besoins et aux attentes des utilisateurs, des défis inhérents à la mesure de ce concept et aux limites des modèles théoriques actuels restent en suspens [14].

Au sein de ce chapitre, nous discuterons d'abord de l'évolution du concept d'acceptabilité dans la littérature, puis nous nous intéresserons au modèle le plus mis en avant dans la littérature : le *Technology Acceptance Model* (ou modèle de l'acceptation de la technologie) (TAM) [44] [43] et de ses principales extensions qui ont apporté de nouvelles dimensions à la compréhension du concept. De nombreux autres modèles empiriques ont également été proposés, nous nous pencherons donc sur les travaux d'unification de ces derniers, qui sont représentés au sein du modèle *Unified Theories of Acceptance and Use of Technology* (ou modèle des Théories Unifiés de l'Acceptation et l'Utilisation de la Technologie) (UTAUT) [169] et son extension. Cette modélisation se présente comme une fusion des modèles existants, visant à définir une approche plus complète pour comprendre le processus d'acceptation d'une technologie. Nous examinerons ensuite les outils de mesure, qui sont aujourd'hui déployés pour étudier l'acceptabilité d'artefacts spécifiques : les solutions intégrant de l'IA. L'objectif est d'identifier si, avec les nombreuses spécificités de ces technologies, leur acceptabilité s'en retrouve également impactée. Enfin, nous terminerons par une critique du concept d'acceptabilité, en examinant ses limites et en discutant des défis et des possibilités pour la recherche future dans ce domaine. L'objectif de cette discussion est de fournir une compréhension approfondie de l'acceptabilité en IHM et de souligner son importance pour la conception et le développement de systèmes interactifs réussis.

2.2 L'évolution de la notion d'acceptabilité

En 1991, Brian Shackel oriente ses recherches en ergonomie vers l'amélioration des IHM, notamment en étudiant le concept d'acceptabilité qu'il associe à l'utilisabilité [154].

Shackel détermine d'abord que l'utilisabilité d'un système réside dans sa capacité à être utilisé facilement et efficacement par certains utilisateurs, à condition qu'ils aient reçu une formation et une assistance adéquate pour accomplir une tâche définie par des scénarios d'utilisation. Shackel suggère que la balance entre l'utilité, l'utilisabilité et la satisfaction procurée par un système permettent de compenser les éventuels coûts humains et financiers de son utilisation, ce qui rend le système acceptable.

Toujours au début des années 90, Jacob Nielsen travaille sur le principe de conception centrée sur l'utilisateur pour maximiser l'adéquation entre les besoins et les contraintes des utilisateurs et les systèmes conçus. Dans cette dynamique, il formalise également un lien entre acceptabilité et utilisabilité [129]. Il soutient que plus un système est utilisable, plus il est susceptible d'être accepté. Nielsen distingue deux types d'acceptabilité (voir figure 2.1) : l'acceptabilité pratique, liée aux aspects matériels du dispositif, aux fonctionnalités proposées, à la facilité d'utilisation, et à la notion de coût introduite par Shackel [154], et l'acceptabilité sociale, qui se focalise sur les influences sociales et normatives impactant la perception du dispositif [137]. Cette représentation n'est pas sans rappeler la confiance, prise en compte sous deux aspects : cognitif et affectif (voir figure 1.6).

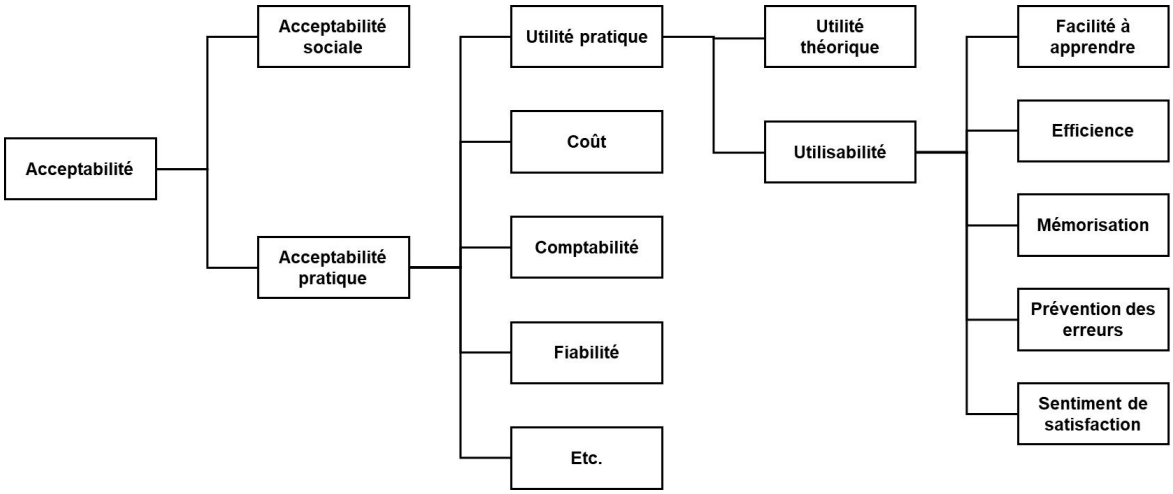


FIGURE 2.1 – Modèle d'acceptabilité d'après J. Nielsen (1993)

Le modèle d'acceptabilité de Nielsen est un outil précieux pour concevoir et évaluer des systèmes interactifs qui répondent aux besoins des utilisateurs en termes de performance, de leur satisfaction et du degré auquel ils se fidélisent à l'outil. Il prend également en compte une dimension expérientielle de l'usage. En 1996, Dillon et Morris conceptualisent

l'acceptabilité comme l'intention d'utiliser un dispositif, qui est influencée par l'utilité, l'utilisabilité, et les perceptions de l'utilisateur [52] (voir figure 2.2), en accord avec le modèle de Nielsen [129].

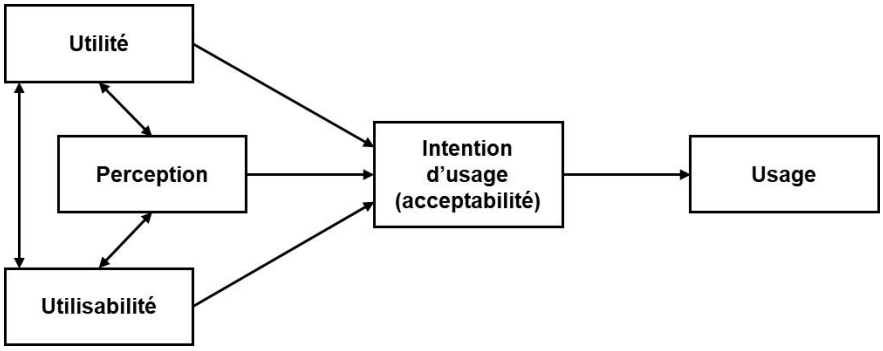


FIGURE 2.2 – Modèle d'acceptabilité d'après Dillon et Moris (1996)

En 2003, Schade et Schlag travaillent sur les stratégies de tarification pour les transports urbains, explorant la notion d'acceptabilité pour améliorer l'acceptation de leurs services. Ils définissent l'acceptabilité comme le degré auquel une solution est jugée juste et raisonnable par le public cible, en tenant compte de facteurs tels que les normes sociales, les attentes personnelles et l'efficacité perçue de la solution proposée [153].

La même année, André Tricot et ses collaborateurs s'interrogent sur la relation entre l'intention d'utiliser un dispositif, son utilité et son utilisabilité [165]. Ils clarifient que l'intention d'usage provient de l'accumulation de représentations mentales (attitudes, opinions, etc.) envers un dispositif. Tricot reconnaît également que l'acceptabilité, l'utilisabilité et l'utilité sont intrinsèquement liées et contribuent à l'utilisation du dispositif. De plus, ils soutiennent que ces trois concepts ne peuvent pas être hiérarchisés [47].

En 2008, l'Organisation Internationale de Normalisation (ISO, pour *International Standard Organization*) propose une définition opérationnelle de l'utilisabilité dans la norme 9241-11 [88]. Elle est définie comme le degré auquel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des objectifs spécifiés avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifique. Selon cette modélisation, l'acceptabilité est une composante de la satisfaction, qui est elle-même une composante de l'utilisabilité (voir figure 2.3).

En 2009, Barcenilla et ses collaborateurs tentent d'établir un lien entre les concepts

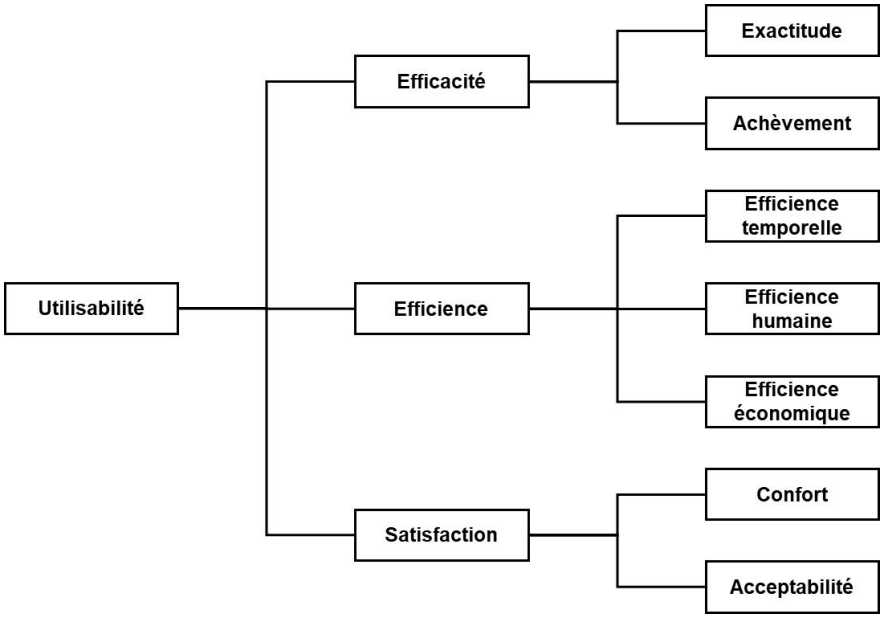


FIGURE 2.3 – Modèle d'utilisabilité d'après la norme ISO 9241-11

d'acceptabilité, d'ergonomie, d'utilisabilité et d'expérience utilisateur (UX). Ils définissent l'acceptabilité comme un "degré d'intégration et d'appropriation d'un objet dans un contexte d'utilisation" [18]. Au sein de l'acceptabilité, ils distinguent l'intégration du dispositif aux activités de l'utilisateur et la manière dont ce dernier s'approprie le dispositif. La même année, Bobilier-Chaumon et Dubois étudient les contextes et les conditions d'adoption des technologies numériques à travers les notions d'acceptabilité et d'acceptation. Ils présentent l'acceptabilité comme une évaluation anticipée des coûts-bénéfices de l'utilisation par l'utilisateur-cible sur un certain nombre de critères, tandis que l'acceptation renvoie à une analyse des conséquences de l'utilisation sur l'activité [25]. Ils soulignent le besoin d'intégrer ces deux concepts dans la conception et l'appropriation de la technologie par les organisations.

Souhaitant définir une démarche de conception, d'évaluation et d'implémentation de technologies innovantes afin de répondre aux attentes des utilisateurs, Kim (2015) s'intéressera également au concept d'acceptabilité [97]. En se basant sur les travaux de Shackel [154] et de Nielsen [129], Kim situe l'acceptabilité comme un concept de plus haut niveau que l'utilisabilité, car elle implique des aspects sociaux, organisationnels et financiers plus complexes (voir figure 2.4).

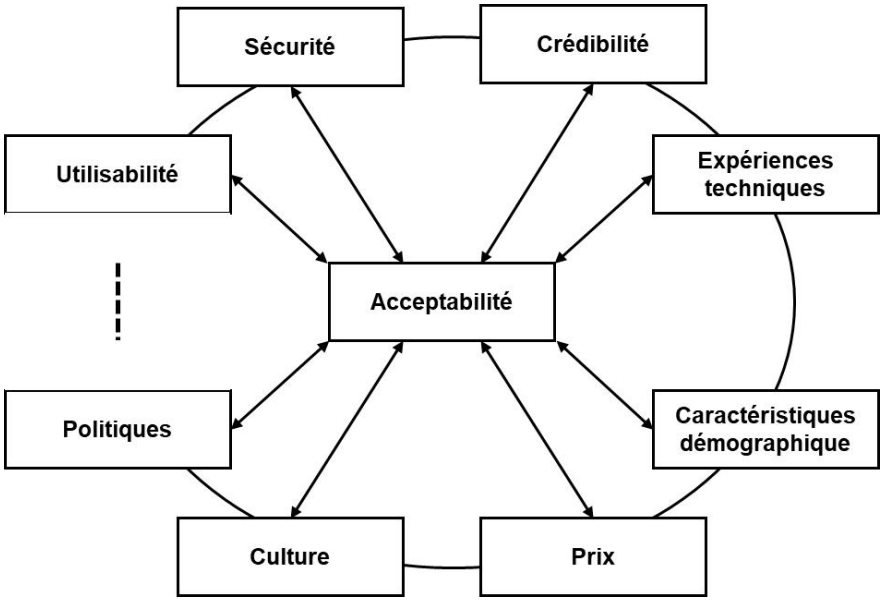


FIGURE 2.4 – Modèle d'acceptabilité d'après Kim (2015)

Il identifie également l'émergence d'un nouveau domaine de recherche : l'ingénierie de l'acceptabilité, qui se focalise sur l'étude des théories et méthodes pour créer des technologies innovantes qui sont acceptables pour les utilisateurs. Kim distingue cette discipline des IHM, car il estime que l'évaluation future des usages des technologies innovantes n'est pas suffisamment prise en compte dans les IHM. Le but de l'ingénierie de l'acceptabilité serait d'évaluer systématiquement l'impact des futures technologies innovantes du point de vue de l'acceptation des utilisateurs. Kim souligne l'importance de prendre en compte le point de vue des utilisateurs dès le début de la conception pour s'assurer que les solutions proposées répondent à leurs besoins et attentes.

Bauchet et ses collaborateurs (2020) se penchent également sur le concept d'acceptabilité, avec une approche plus similaire à celles de Schade et Schlag (2003) [153], et de Bobilier-Chaumon et Dubois (2009) [25]. Les chercheurs ont étudié le processus d'acceptation des outils numériques dans le système éducatif [22]. Dans leurs travaux, ils ont proposé le Modèle de 4A qui offre une description du processus général d'intégration du numérique au sein d'une structure institutionnelle en termes d'acceptation par les utilisateurs cibles. Le modèle de 4A montre qu'il y a une différence significative entre l'adoption consentie ou imposée d'une solution au sein d'une institution (voir figure 2.5). Les auteurs définissent l'acceptabilité comme étant les attitudes prédictives (intention d'utilisation et croyances

concernant la solution) de l'utilisation du dispositif. À la suite de l'implémentation de la solution, ils positionnent la notion d'acceptation, décrite comme les perceptions et attitudes que les utilisateurs peuvent développer après une première interaction avec la solution, tout comme le suggèrent Bobillier-Chaumon et Dubois (2009) [25]. La troisième étape est l'adoption de la solution, c'est-à-dire son utilisation effective qui s'intègre aux activités des utilisateurs cibles. Ils soulignent cependant qu'une technologie peut ne pas être acceptée par les employés et pourtant être utilisée, en particulier lorsque la solution est mise en œuvre à la suite des directives de l'organisation - ce qu'ils appellent l'adoption imposée. Dans ce scénario, les utilisateurs sont réticents mais contraints à un usage prescrit, ils n'adopteront donc pas pleinement la solution. À l'inverse, si la solution est acceptée, donc que les perceptions et attitudes sont positives après un premier usage, l'adoption est consentie et volontaire. Dans ce second cas, ils considèrent que les utilisations vont s'élargir et pourront se généraliser à d'autres environnements, ce qu'ils appellent l'appropriation.

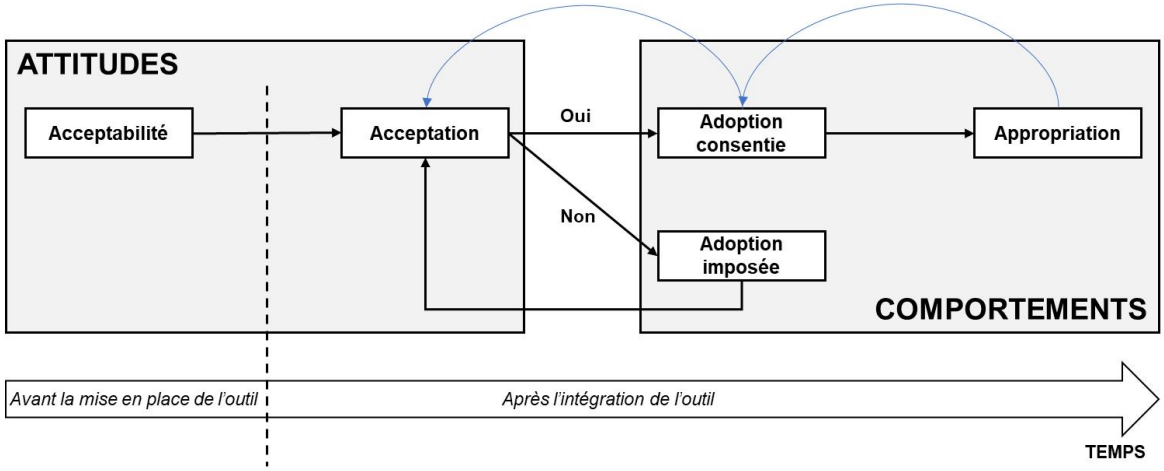


FIGURE 2.5 – Modèle de 4A, de Bauchet et al. (2020)

La notion d'acceptation se rapproche énormément d'une notion mise en avant dans la discipline de l'informatique : la qualité d'expérience (QoE pour *Quality of Experience*). La QoE est définie comme la qualité, finement mesurée, d'un service offert. Cette notion semble donc renvoyer au degré de contentement/satisfaction de l'utilisateur vis-à-vis du service utilisé, au même titre que l'acceptation [10] [38]. La QoE est une notion particulièrement mise en avant par Qualinet, le réseau européen de la qualité d'expérience dans les systèmes et services multimédias. Au sein de Qualinet, une large liste de contributeurs ont permis d'avoir une perception plus éclairée de ce qu'est la QoE au travers d'un livre blanc [28]. Les

chercheurs y définissent la QoE comme le degré de satisfaction ou d'insatisfaction ressenti par l'utilisateur lors de l'utilisation d'une application. Cette satisfaction est influencée par la capacité du service à répondre aux attentes des utilisateurs en termes d'utilité et de plaisir, tout en prenant en compte la personnalité et l'état émotionnel de l'utilisateur. Les chercheurs distinguent également la QoE de l'acceptabilité, qu'ils définissent comme le résultat d'une décision basée en partie sur la QoE. L'acceptabilité est présentée comme une conséquence de la QoE, reflétant la facilité avec laquelle un utilisateur peut choisir d'utiliser un service. En contraste, la QoE est vue comme une mesure plus large que la satisfaction qui intègre des aspects subjectifs tels que les émotions et les attentes. Contrairement à ce que nous avons exploré jusque-là, cette communauté scientifique positionne l'acceptabilité comme étant un concept plus global au sein duquel se trouve la qualité d'expérience, que nous associons à l'acceptation.

En 2021, Alexandre et ses collaborateurs ont étudié les relations entre l'acceptabilité et l'acceptation d'un outil, tout comme Bobillier-Chaumon et Dubois, pour déterminer s'il existe une hiérarchie entre les critères qui amènent un utilisateur à utiliser un outil particulier (comme Tricot) [165] [25] [11]. Les chercheurs démontrent que l'exposition préalable à l'outil influence le jugement de l'utilisateur, indiquant une différence entre acceptabilité, présenté comme anticipatrice et relative à l'observation de l'outil, et acceptation, liée à l'utilisation effective de l'outil. Dans la phase d'acceptabilité, les jugements des utilisateurs seraient guidés par la facilité d'utilisation perçue. Alors, pendant la phase d'acceptation, l'utilité perçue a une plus grande influence, que ce soit en termes de préférence d'outil ou de temps passé à utiliser des outils.

2.2.1 Le Technology Acceptance Model (TAM) et ses extensions

En 1986, Davis se penche sur le paradigme du coût-bénéfice, influençant l'acceptation d'une technologie de l'information et de la communication (TIC). Son approche est basée sur une stratégie de prise de décision où l'utilisateur potentiel équilibre cognitivement l'effort requis pour utiliser l'outil (facilité d'utilisation perçue) et le résultat escompté (utilité perçue), ce qui aboutit à l'intention d'utiliser l'outil, définissant ainsi l'acceptation de l'utilisateur [46] [43]. De ces principes, Davis élabore un modèle empirique initial : le TAM (*Technology Acceptance Model*). Le but de ce modèle est de prédire l'adoption d'une technologie en se basant sur des facteurs perçus qui influencent l'intention d'utiliser la TIC cible. Ces facteurs comprennent l'utilité perçue du système, sa facilité d'utilisation

perçue, et les attitudes envers son utilisation.

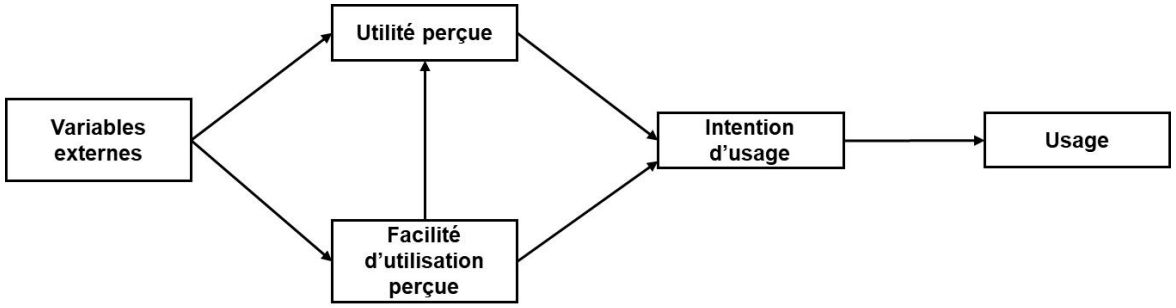


FIGURE 2.6 – Technology Acceptance Model (TAM), de F. Davis (1989)

Au départ, Davis a postulé que les attitudes envers l'utilisation étaient affectées par l'utilité perçue et la facilité d'utilisation perçue, et que ces attitudes influençaient à leur tour l'intention d'utilisation. Inspiré par le modèle TRA (*Theory of Raisonned Action*) [7], le TAM a été révisé en 1989. Davis a retiré le facteur des attitudes après avoir identifié que l'utilité perçue et la facilité d'utilisation perçue avaient un impact direct sur l'intention d'utilisation, un effet qu'il avait auparavant envisagé comme étant modéré par les attitudes.

L'année 2000 voit Venkatesh et Davis revisiter le TAM pour mieux comprendre les conditions d'adoption d'une technologie par des individus au sein d'une organisation [168]. Ils introduisent une version révisée du TAM, le TAM 2, qui prend en considération que l'utilité perçue est influencée par des processus cognitifs instrumentaux (comme la pertinence de l'emploi, la qualité des résultats, et la démonstrabilité des résultats) et des processus d'influence sociale (norme subjective, volontariat et image) (voir figure 2.7. Le TAM 2 se distingue de son prédécesseur, en intégrant des processus d'influence sociale comme impactant l'utilité perçue de l'artefact [96].

En 2008, Venkatesh et Davis retravaillent le TAM 2 et intègrent les déterminants de la facilité d'utilisation perçue, aboutissant au TAM 3 [167]. Ils identifient trois limites au TAM 3, toutes liées à la prise en compte de la dimension sociale : 1) l'exclusion d'autres dimensions sociales, 2) une insuffisance de facteurs modérateurs, et 3) une définition floue de la notion de norme (voir figure 2.8).

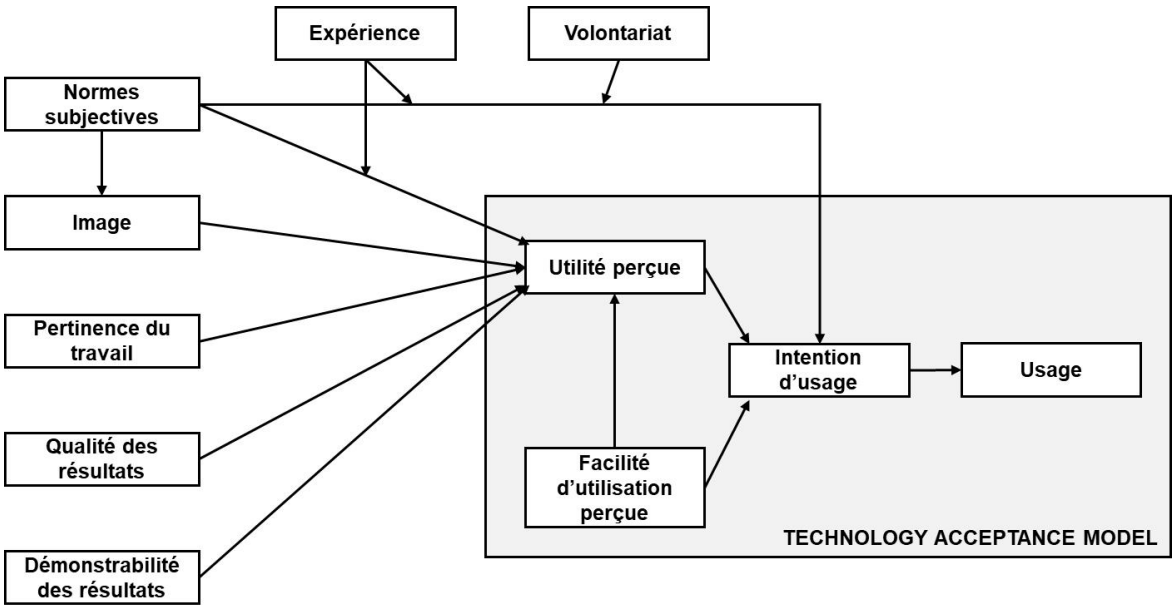


FIGURE 2.7 – Technology Acceptance Model première extension (TAM 2), de Venkatsh et Davis (2000)

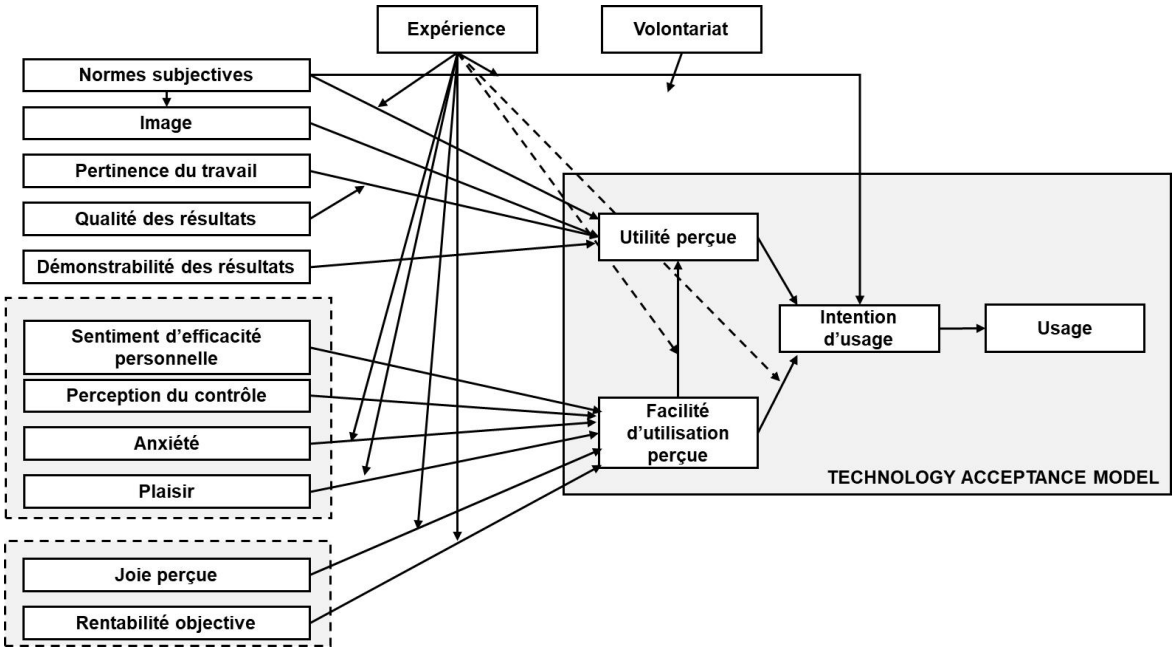


FIGURE 2.8 – Technology Acceptance Model seconde extension (TAM 3), de Venkatsh et Davis (2008)

2.2.2 Unified Theory of Acceptance and Use of Technology (UTAUT) et son extension

En raison de l'augmentation du nombre de recherches sur l'acceptation des TIC au début des années 2000, une multitude de modèles théoriques ont émergé. En 2003, Venkatesh et ses collaborateurs ont donc effectué une revue de la littérature sur l'acceptation des technologies et ont identifié huit modèles théoriques majeurs [166], dont le modèle de la Théorie du Comportement Planifié de Ajzen [5], la Théorie de l'Action Raisonnée de Ajzen et Fishbein [6], le TAM et le modèle motivationnel de Davis et al. [43] [45]. Ces modèles expliquaient entre 17% et 53% de la variance de l'intention d'utilisation. Ils ont donc proposé une théorie unifiée de ces modèles, l'UTAUT (*Unified Theory of Acceptance and Use of Technology*), qui comporte quatre déterminants principaux de l'intention d'utilisation et de l'utilisation réelle : la performance espérée, l'effort attendu, l'influence sociale, et les conditions facilitatrices (voir figure 2.9).

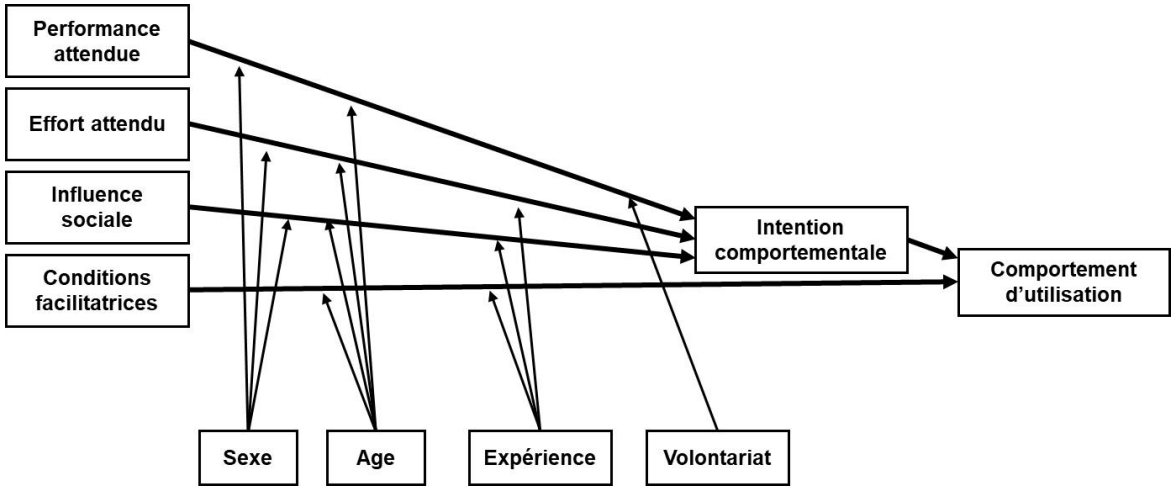


FIGURE 2.9 – Unified Theory of Acceptance and Use of Technology (UTAUT), de Davis et al. (2003)

L'impact de chaque déterminant est modéré par les caractéristiques propres à l'utilisateur cible (âge, sexe, expérience et utilisation volontaire). L'UTAUT permet une meilleure compréhension des facteurs de l'acceptation, et sert à concevoir de manière proactive des interventions (comme la formation et le marketing) sur des populations d'utilisateurs cibles susceptibles d'être moins enclines à adopter et utiliser de nouveaux systèmes.

En 2012, Venkatesh et ses collaborateurs révisent leur modèle UTAUT pour proposer

l'UTAUT 2 [170]. De nouveaux déterminants clés sont ajoutés pour expliquer l'intention d'utiliser une technologie. Ce modèle est destiné au marché de la consommation, et non au déploiement au sein d'une organisation, car les nouveaux déterminants sont axés sur le coût motivationnel, monétaire et social de l'utilisation, où l'utilisateur est le seul décideur (voir figure 2.10.)

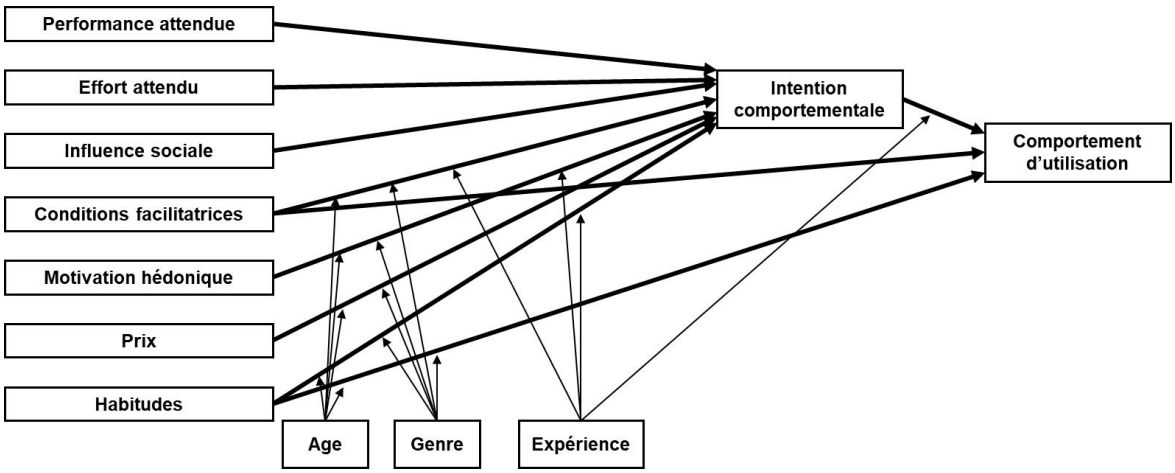


FIGURE 2.10 – Unified Theory of Acceptance and Use of Technology avec son extension (UTAUT 2), de Venkatesh et al. (2012)

Ces déterminants sont 1) la motivation hédonique, le degré auquel l'utilisation est perçue comme agréable et plaisante par l'utilisateur, 2) le prix, qui évalue si le coût de la solution est justifié par l'utilisation qui en est faite, et 3) les habitudes, qui examinent à quel point les habitudes d'utilisation des utilisateurs sont perturbées.

2.3 L'évaluation de l'acceptabilité en contexte professionnel

L'évaluation de l'acceptabilité des technologies dans un contexte professionnel est une tâche cruciale qui peut entraîner des conséquences profondes sur l'efficacité et la productivité d'une organisation. L'approche la plus couramment utilisée et présente dans la littérature est l'usage de questionnaires, pour mesurer quantitativement des aspects de l'acceptabilité, comme la facilité d'utilisation perçue et l'utilité perçue. Le TAM et le modèle UTAUT sont des exemples célèbres de cadres théoriques qui ont inspiré le développement de ces outils de mesure. Ces modèles sont souvent utilisés comme base pour

la création de questionnaires spécifiques au contexte qui peuvent évaluer l'acceptabilité d'une technologie particulière dans un environnement de travail.

Dans leur étude sur la perception des robots de service dans la restauration, Lee et ses collaborateurs (2018) ont mesuré l'acceptabilité de ces derniers auprès d'une population-cible. Pour cela, les auteurs ont utilisé la version initiale du TAM [44] pour mesurer l'utilité perçue, la facilité d'utilisation perçue, les attitudes et l'intention d'usage. En complément de cette approche, ils ont mesuré et cherché à faire le lien avec la confiance, l'interactivité et la qualité des résultats du robot. A l'issu de leur recherche, les chercheurs ont démontré un effet statistique de la confiance sur l'utilité perçue et un autre effet de l'interactivité et la qualité des résultats sur la facilité d'utilisation perçue.

De manière similaire, Pillai et ses collaborateurs (2019) ont utilisé la même approche mais avec la version du TAM de 1989 [43] (donc sans les attitudes). En complément, ils ont également mesuré et cherché à faire le lien avec les dimensions suivantes : l'anxiété technologique, la confiance, l'anthropomorphisme, l'intelligence perçue et l'attachement aux agents humains présents initialement. Leur étude démontre également un lien en la confiance et l'intention d'usage, dans leur contexte.

Durant sa thèse sur les liens entre utilisabilité et acceptabilité d'un dispositif de saisie et de reconnaissance de l'écriture manuscrite, Guillaume Deconde (2011) a traduit le TAM de Davis pour mesurer l'acceptabilité de dispositifs mobiles (voir figure 3) [47]. Parallèlement, pour sa thèse sur la définition de l'acceptabilité sociale, Hélène Marie Louise Pasquier (2012) a mesuré l'effet de valorisation des comportements dans l'explication de l'acceptabilité [137]. Pour cela, elle a élaboré un questionnaire mesurant les attitudes, les normes subjectives, le contrôle comportemental perçu, l'acceptabilité (intention comportementale), l'image et l'identité personnelle.

2.4 Mesurer l'acceptabilité de l'IA

Notre exploration de la littérature autour de l'acceptabilité des solutions IA en contexte professionnel nous montre que cette thématique prend une place importante dans le domaine des IHM. Parmi nos références bibliographiques traitant du processus d'acceptation d'une technologie, nous constatons un essor de publications s'y intéressant depuis le milieu

des années 2010. Cet accroissement est d'autant plus fort depuis 2020 avec l'émergence de nouvelles solutions IA telles que les outils génératifs, qui s'intègrent en entreprise. Ici, les opérateurs humains questionnent leur manière d'accomplir une tâche, avec des outils susceptibles d'occuper une partie de cette activité. A l'heure actuelle, ces sujets semblent principalement traités par les sciences du management et les sciences de l'information & communication.

Le TAM incarne le principal outil utilisé pour mesurer l'acceptabilité en contexte professionnel. Parmi ses différentes révisions qui ont été proposées dans la littérature [43], de nouvelles déclinaisons émergent. Notamment, le AI-TAM (pour *Artificial Intelligence - Technology Acceptance Model*) qui est un modèle proposé pour mesurer spécifiquement l'intention d'usage de solutions IA [19] (voir figure 2.11). Cette révision, proposée par Baroni et ses collaborateurs, intègre des facteurs associés à l'explicabilité de l'IA (XAI) et à la collaboration dans la conception. Les facteurs associés aux XAI comprennent 1) la confiance de l'utilisateur dans l'IA, c'est-à-dire la confiance que les utilisateurs ont dans les capacités de l'IA à produire des résultats fiables et précis, 2) la qualité perçue de la sortie de l'IA, qui désigne la qualité du résultat proposé et le niveau de soutien de la réponse. Quant à l'intention de collaborer avec le modèle, elle fait référence à la volonté des utilisateurs de contribuer à l'amélioration de l'IA en fournissant leurs propres données et/ou en validant les résultats du modèle pour l'entraînement. Les résultats de leur étude montrent que les facteurs liés à l'IA explicable (XAI) ont un impact positif et significatif sur l'utilité perçue, la facilité d'utilisation perçue, et l'intention d'utiliser l'IA. De plus, ils ont mis en évidence l'impact de l'intention d'usage sur l'intention de collaborer avec le modèle.

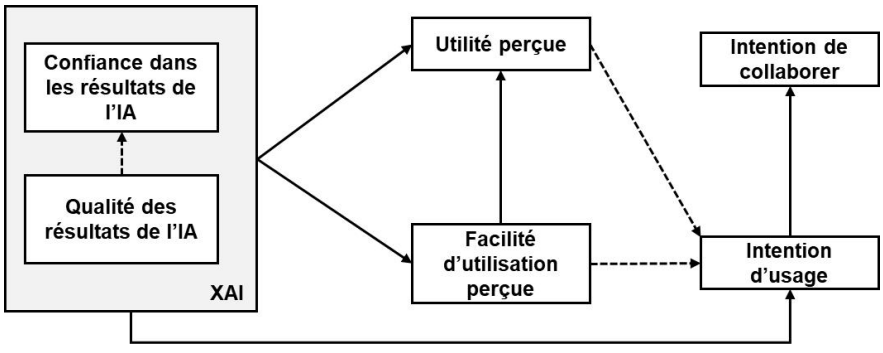


FIGURE 2.11 – AI - Technology Acceptance Model (AI-TAM) de Baroni et al. (2022)

2.5 Critique de l'acceptabilité

L'acceptabilité, comme d'autres concepts en IHM, n'est pas exempte de critiques. Si sa mesure est essentielle au déploiement et à l'appropriation de solutions technologiques en contexte professionnel, il est d'autant plus nécessaire d'en comprendre les limites. Nous nous intéresserons ici à offrir une vue d'ensemble de l'acceptabilité plus nuancée. Et ce, afin de mieux appréhender le concept pour une prise en compte plus efficiente.

Parmi les principales préoccupations autour du TAM, l'une d'entre elles concerne sa capacité prédictive de l'utilisation d'une technologie. D'après la littérature, le TAM semble mieux prédire l'intention d'usage que les autres modèles avec une variance expliquée qui se situe entre 30% et 60% selon le contexte dans lequel il est appliqué [14]. Même sa première extension, le TAM 2, n'explique qu'environ 60% de la variance de l'acceptabilité au maximum, et donc propose des résultats similaires mais avec beaucoup plus de dimensions à appliquer [2]. La deuxième extension, le TAM 3, explique entre 40% et 70% de l'intention d'usage. Par rapport aux précédentes versions, la marge de variance expliquée est réduite et le score maximal est d'autant plus élevé. Mais cette version reste difficile à administrer, sachant que plus nous avançons dans les versions du TAM, plus il y a de facteurs à évaluer, ce qui les rend plus difficiles à utiliser.

De plus, les fondements-même du TAM sont questionnés par plusieurs études qui ont montré que bien que le lien entre l'utilité perçue et l'intention d'usage est généralement confirmé, le lien entre la facilité d'utilisation perçue et l'intention a plus souvent tendance à être contradictoire dans la littérature. La facilité d'utilisation perçue semble souvent ne pas avoir d'effet significatif sur l'intention d'usage [118] [35] [13] [91] [156].

Au-delà des limitations actuelles du TAM, nous ouvrons également la porte à une multitude d'autres variables qui peuvent jouer un rôle crucial dans l'acceptabilité des technologies en milieu professionnel. C'est notamment le cas de l'anthropomorphisme. L'anthropomorphisme se caractérise par la projection des traits humains de l'utilisateur sur un artefact. Ce phénomène a été reconnu comme un facteur significatif dans l'acceptabilité des technologies [139] [149]. Ce trait influence la manière dont les utilisateurs interagissent avec la technologie et peut contribuer à une plus grande empathie envers l'artefact. Toute-

2. Le TAM 2 est donc plus coûteux à administrer car il demande plus de temps et de concentration aux répondants.

fois, l'anthropomorphisme n'est pas sans défis, car un artefact trop anthropomorphisé peut créer une sensation d'étrangeté. C'est la théorie de la Vallée de l'étrange. Plus un artefact présente des caractéristiques humaines, plus les utilisateurs ont tendance à éprouver une certaine empathie pour cet artefact, mais il ne doit pas être proche de l'humain au point de pouvoir provoquer une confusion. La relation entre anthropomorphisme et acceptabilité n'est donc pas linéaire [23] [149]. Dans leur étude sur l'utilisation de chatbot avec de l'IA dans le domaine du tourisme, Pillai et collaborateurs mettent en avant que l'anthropomorphisme joue un rôle dans l'acceptabilité de ces artefacts, au même titre que la confiance qu'il génère, donc le degré auquel l'utilisateur estime que l'artefact est fiable et crédible, ainsi que l'intelligence perçue, donc la compétence, la transmission de connaissances, la sensibilité et la réaction responsable de l'outil.

Comme observé précédemment, la confiance semble représenter un autre facteur clé pour comprendre l'acceptabilité des technologies. À mesure que les technologies deviennent de plus en plus complexes et s'intègrent dans nos vies, la confiance dans ces systèmes devient un enjeu majeur. La confiance accordée à un système intelligent est susceptible de passer par une multitude de déterminants, allant de la robustesse de l'outil et le respect des législations, jusqu'à l'aspect technique et l'exploitation des données [86] [106] [67] [130] [108]. Lee et al. (2018) se sont penchés sur l'acceptabilité des robots dans les services de restauration. Ils ont défini la confiance comme le degré de croyance qu'un individu a dans le fait que l'outil agit de manière sincère, bienveillante, compétente, fiable, respectueuse de l'éthique et de la société. Les auteurs ont découvert que la confiance a un effet significatif sur l'utilité perçue d'une technologie. Ils ont également constaté que la qualité des résultats obtenus à l'aide de la technologie et l'interactivité de celle-ci ont un effet sur la facilité d'utilisation perçue [106]. Les résultats de leur étude suggèrent que la confiance est un élément crucial de l'acceptabilité dans les environnements de travail. En outre, ils mettent en lumière l'importance de la conception et de la fonctionnalité d'une technologie, ainsi que de la nature de la tâche qu'elle est censée accomplir, dans la détermination du niveau de confiance des utilisateurs.

De son côté, Impact IA, qui travaille à la création d'une approche collective de l'IA en France, soutient qu'une IA acceptable est une IA digne de confiance [86]. Cette confiance se construit sur trois piliers : la robustesse de l'outil, le respect des législations, et le respect des principes éthiques. Ces piliers ont été détaillés dans la section 1.4, pour souligner

l'importance d'une approche globale qui comprend tous ces aspects pour établir la confiance dans une technologie IA et ainsi favoriser son acceptabilité.

D'autres études proposent une approche multi-factorielle de la confiance dans la solutions IA comme par exemple Oscar Hengxuan Chi et ses collaborateurs en 2021. Les chercheurs ont travaillé sur l'élaboration d'une échelle de mesure de la confiance dans les interactions avec des robots de service social intégrant de l'IA dans le cadre d'une prestation de service [80] (voir figure 2.12).

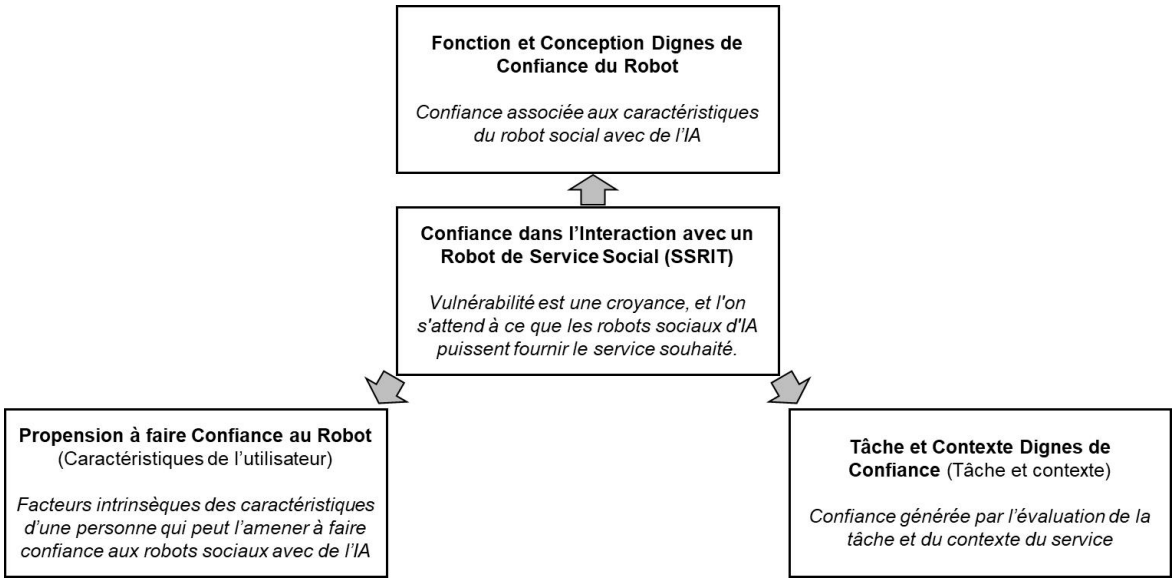


FIGURE 2.12 – Dimensions de la confiance dans les interactions avec des robots de service social intégrant de l'IA, notre traduction des travaux de Hengxuan Chi et al. (2021)

Dans ce contexte, les auteurs argumentent que la confiance repose sur trois dimensions : 1) la propension à faire confiance au robot avec IA, intrinsèque à l'utilisateur, 2) la fonction et la conception du robot, associées aux caractéristiques de l'IHM et 3) et la tâche désignée, renvoyant à l'évaluation des interactions. Cette classification propose de lier la confiance aux déterminants de l'acceptabilité du TAM. Elle la situe à un niveau supérieur puisque les auteurs imbriquent la performance perçue, correspondant à l'utilité perçue dans le TAM [43], dans la dimension Fonction et Conception du robot. Pour la facilité d'utilisation perçue du TAM [43], nous la retrouvons également dans cette dimension du modèle de Hengxuan Chi et al., mais en tant qu'effort attendu. Cette classification propose à nouveau de lier la confiance aux déterminants de l'intention d'usage du TAM.

2.6 Conclusion

L'acceptabilité reste aujourd'hui un domaine de recherche actif. Les chercheurs explorent comment différents facteurs, tels que les normes culturelles, les valeurs personnelles et les expériences passées, peuvent influencer l'acceptabilité des nouvelles technologies. Les modèles d'acceptation des technologies continuent d'être affinés et de nouveaux modèles sont développés. Ces modèles ont une double fonction : décrire non seulement les réactions des individus face à l'utilisation des outils, mais aussi élaborer et tester des hypothèses préalables concernant les attitudes des utilisateurs et acceptabilité de l'artefact.

L'évolution des recherches sur l'acceptabilité en milieu professionnel montre qu'il n'existe pas une seule et unique manière de mesurer ce concept. Ces méthodes se veulent évolutives, en fonction du contexte, des utilisateurs ou encore des spécificités du dispositif telles que les solutions IA. Dans le cas de l'IA, il est nécessaire de prendre en compte une multitude de facteurs, comme l'anthropomorphisme et la confiance, qui vont au-delà des éléments fondamentaux du TAM. Le défi pour les chercheurs et les praticiens est d'intégrer ces divers éléments dans une approche holistique de l'acceptabilité, qui reflète la complexité des environnements de travail et des technologies qui y sont déployées. Surtout lorsque ces dernières évoluent très rapidement.

Chapitre 3

Théories fondamentales de l'UX et apports dans le déploiement de solutions IA

Dans ce chapitre

Ce chapitre se concentre sur les définitions de l'UX et sa prise en compte dans la conception de solutions informatiques professionnelles. Une multitude de modèles théoriques montrent que la considération pour l'UX est question de maturité de l'entreprise à ce sujet. Et cette maturité se déploie au fur et à mesure que les stratégies de conception se détache d'une vision technocentrée. Face aux spécificités des technologies IA, intégrer une démarche de prise en compte de l'UX peut contribuer à faciliter l'adoption de la solution IA en veillant à construire une relation de confiance envers cette dernière.

3.1 Introduction

L'expérience utilisateur (UX) est devenue un concept récurrent dans le domaine des interactions humain-machine (IHM), visant à saisir la qualité de l'interaction avec un artefact dans son intégralité [133] [140]. Il existe de nombreuses définitions de l'UX, dont la plupart prennent en compte les émotions, les attitudes et les réponses d'un individu à l'interaction. Aujourd'hui, l'UX Design (ou l'approche de conception UX) est très appréciée dans le contexte professionnel. En prenant en compte l'utilisateur et l'écosystème dans la conception des artefacts, elle permet un changement significatif dans leur acceptation.

L'UX Design aborde un ensemble de problématiques, tant en termes de conception que d'usage prévu, qui, dans un environnement professionnel, favorise non seulement une bonne expérience utilisateur, mais aussi l'efficacité, la productivité et la satisfaction globale au travail [102]. L'UX Design est ainsi devenue un levier crucial pour la mise en œuvre réussie d'outils professionnels. Cependant, la création d'une UX optimale en contexte professionnel représente un défi particulier. Les professionnels interagissent avec leurs outils de travail différemment des consommateurs ordinaires, avec des attentes différentes en termes de fiabilité, d'efficacité et de précision. Très souvent, le choix des outils de travail est fait par les supérieurs, et non par les employés, qui sont pourtant les utilisateurs cibles. Avec l'introduction des solutions IA dans les situations de travail, cette question prend une dimension nouvelle et plus complexe. Les solutions IA, tout en offrant un énorme potentiel d'optimisation et d'automatisation, soulèvent également des questions pertinentes en matière d'UX.

Dans ce chapitre consacrée à l'exploration de la notion d'expérience utilisateur, nous aborderons tout d'abord les théories fondamentales de l'UX, et comment elle est définie dans le domaine des IHM. Ensuite, nous examinerons les modèles théoriques destinés à éclairer les entreprises sur l'évolution des stratégies d'intégration de l'UX. Puis nous discuterons des défis liés à l'application de méthodes valorisant l'UX dans le projets IA en entreprise.

3.2 Définition

Cette section vise à examiner l'évolution de la notion d'expérience utilisateur (UX) dans la littérature en IHM. Introduite par Norman en 1988, la notion d'UX est définie comme la qualité de l'expérience vécue par une personne qui interagit avec un artefact [132]. L'UX est alors considérée comme un concept holistique, englobant tous les aspects de l'expérience d'un individu, tout en tenant compte du processus de conception. Ce qui est central dans l'interprétation de ce concept est l'interaction entre un humain et un artefact [132] [10] [116] [41] [120] [74] [77] [115] [161] [72] [17]. La norme ISO 9241, qui couvre les normes relatives à l'ergonomie des interactions humain-système, recommande d'également prendre en compte le contexte dans lequel se déroule cette interaction. Cette recommandation se fait en écho à ce qui est présenté dans la définition de la notion

d'utilisabilité, qui est intrinsèquement lié à l'UX. L'utilisabilité y est définie comme le degré d'utilisation d'un artefact par une population cible dans un contexte donné pour une tâche spécifique [88]. Nous approfondirons la notion d'utilisabilité dans la section 3.5.

Le contexte joue un rôle crucial dans l'UX, car il peut motiver l'utilisation de l'artefact et rendre l'UX plus dynamique en s'adaptant aux variations contextuelles [116]. Elle évolue également dans le temps, car la perception de l'artefact, la compréhension de son fonctionnement, l'émotion qu'il suscite ou encore sa capacité à répondre aux besoins évoluent selon la perception de l'utilisateur en anticipation du premier usage (acceptabilité), des conditions dans lesquelles il s'en sert et des conséquences post-usage [10]. Ainsi, les expériences et attentes passées de l'utilisateur influencent son expérience présente, qui à son tour conduit à de nouvelles expériences et à des attentes en évolution. En 2006, Hassenzhal et Tractinsky réaffirment le concept en insistant sur la place de l'utilisateur dans la conception pour lui proposer une bonne expérience d'usage. L'utilisateur doit être au centre de la conception, car l'UX correspond à la conséquence de l'état interne de l'utilisateur, des caractéristiques du produit conçu et du contexte de l'interaction [74].

Cependant, l'UX est surtout un concept intrinsèquement subjectif. Alors que son objectif est de représenter la qualité de l'interaction avec un artefact, elle englobe également une dimension négligée dans la définition de l'utilisabilité : l'émotion suscitée par l'artefact. Étant donné que l'UX est une relation personnelle entre l'utilisateur et l'artefact, elle peut générer davantage qu'une simple satisfaction liée au degré d'utilisation pour réaliser un objectif. Cela englobe la satisfaction esthétique, les significations attribuées à l'artefact [77], ainsi qu'un sentiment de plaisir momentané ressenti lors de l'interaction avec l'artefact [72]. En 2009, Barcelina et Bastien ont adopté une approche plus généraliste, envisageant l'UX comme un cadre intégrateur des différentes composantes relatives à l'interaction utilisateur-produit [17]. Cette présentation vise à montrer une évolution dans la conception de l'UX, avec une compréhension de l'interaction qui intègre et met en évidence le rôle du contexte, des attentes, des perceptions et des émotions de l'utilisateur dans son expérience globale. Nous pouvons nous inspirer de la synthèse de l'UX d'après Berni et Borgianni (2021) pour mieux appréhender cette notion telle que proposée dans la littérature. Les chercheurs explique que l'UX repose sur trois éléments : l'utilisateur, le système et le contexte d'utilisation (voir figure 3.1)

Ces trois éléments constituent de la base de l'UX qui peut être identifiée et décrite

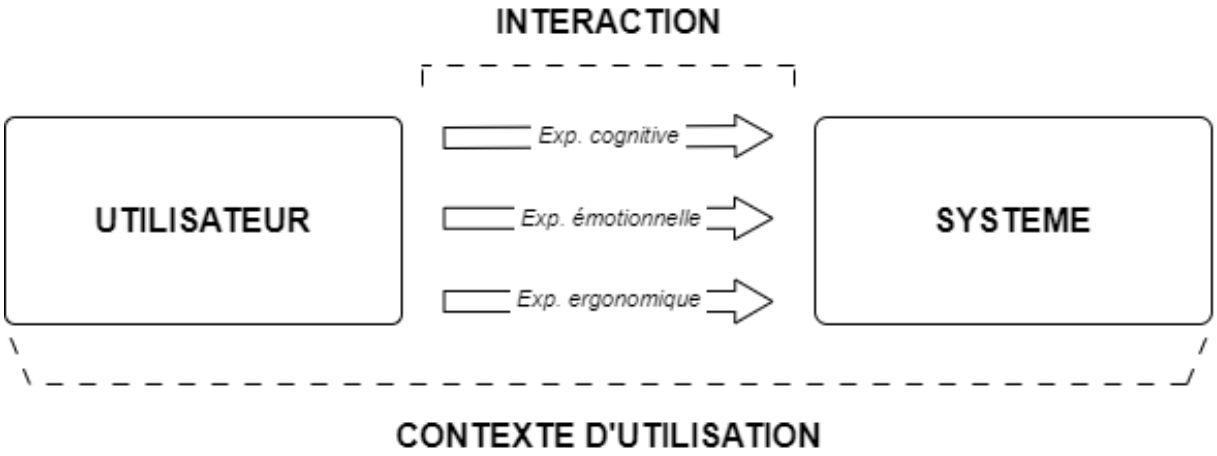


FIGURE 3.1 – Représentation de l'expérience utilisateur selon Berni et Borgianni (2021)

selon trois catégories d'expérience :

- L'expérience ergonomique, qui renvoie à des aspects qu'on retrouve notamment dans la littérature sur l'acceptabilité comme la facilité d'utilisation, l'utilité / efficacité. Nous parlons ici d'une expérience en terme d'atteinte du but, cette expérience semble plutôt reposer sur la fonctionnalité de l'artefact [10].
- L'expérience cognitive concerne plutôt la perception que l'utilisateur a de l'artefact. Ici, ce qui prévaut est ce que renvoie le système, notamment en termes d'esthétisme et, en particulier, la manière dont l'utilisateur perçoit l'apparence extérieure, l'esthétique d'un système [10] [74] [41].
- Et enfin l'expérience émotionnelle comprend toutes les composantes émotionnelles liées à l'interaction, tels que le plaisir ou le désagrément lié à l'interaction [131] [10].

3.3 L'UX en contexte professionnel : un vecteur d'acceptation

L'UX est un sujet de plus en plus présent dans le domaine des IHM. On remarque notamment que sa prise en compte permet de favoriser l'acceptation des produits conçus en répondant davantage aux besoins et contraintes des utilisateurs finaux. Mais toutes les entreprises n'ont pas la capacité et/ou les connaissances suffisantes pour correctement aborder cette notion [155]. La tendance pour beaucoup d'entreprises est de prioriser les besoins des clients sur ceux des utilisateurs finaux, en particulier dans les secteurs orientés

B2B (pour *Business to Business*), terme qui désigne que l'activité de l'entreprise est à destination d'autres entreprises ou du moins à un public professionnel. Dans ce cadre, les exigences commerciales ou techniques des clients, qui ont des intérêts financiers, peuvent éclipser l'expérience véritable des utilisateurs avec le produit conçu [2]. Nous retrouvons ainsi la prise en compte de l'UX souvent réduite à de simples considérations esthétiques ou fonctionnelles, ignorant la multiplicité de dimensions expérientielles que nous avons abordé dans la section 3.2.

Face à ces constats, nous examinerons dans cette section les modèles théoriques de maturité des entreprises dans leur prise en compte de l'UX. Ces modèles offrent un cadre visant à éclairer comment les entreprises évoluent dans leur approche de l'UX, de la simple prise de conscience à une intégration profonde dans leur culture et leurs pratiques. Il nous semble important, pour ces entreprises, de comprendre chacune de ces étapes afin d'identifier les actions à entreprendre pour concevoir des artefacts plus en adéquation avec les publics cibles. Cela doit se faire tout en répondant aux besoins et contraintes de tout le réseau de parties prenantes à ces projets. Dans un deuxième temps, nous porterons notre attention sur l'impact que la prise en compte de l'UX peut avoir sur les situations de travail. Nous nous pencherons plus spécifiquement sur les cas où les entreprises souhaitent déployer des solutions IA.

3.3.1 Les modèles de prise en compte de l'UX en contexte professionnel

Dans cette section, nous examinerons plusieurs cadres théoriques bien établis pour évaluer à quel point les entreprises intègrent l'UX dans la création de leurs produits. Nous verrons que dans ces cadres, les concepts d'UX et d'utilisabilité, définis précédemment, sont souvent entre-mêlés et parfois confondus. Pour rappel, l'utilisabilité se réfère au degré d'utilisation d'un artefact par des utilisateurs identifiés pour une tâche donnée dans un contexte spécifié [88]. En revanche, l'UX adopte une perspective plus dynamique et globale, incluant des aspects subjectifs tels que les émotions suscitées par l'interaction avec le produit ou encore la perception générale de l'utilisateur. Certains des modèles que nous allons présenter se concentrent principalement sur l'utilisabilité, mais ils restent pertinents pour l'évaluation de la maturité UX. Car l'utilisabilité en est une composante essentielle.

En 1998, Earthy travaille sur la maturité des organisations et étudie leur capacité à prendre en compte l'utilisabilité dans leur processus de conception centrée-utilisateur. Dans son approche, Earthy propose six stades de maturité, par lesquelles les entreprises vont toutes passer au fur et mesure qu'elles s'impliquent et intègrent l'UX à leurs pratiques [54].

Stade de maturité	**Description**
Stade X	**Inconnue**
Stade A	**Connue**
A1	Attribut de reconnaissance du problème
A2	Attribut de processus effectués
Stade B	**Considérée**
B1	Attribut de sensibilisation à la qualité d'utilisation
B2	Attribut d'orientation utilisateur
Stade C	**Implémentée**
C1	Attribut d'implication de l'utilisateur
C2	Attribut de technologie liée aux facteurs humains
C3	Attribut de compétences en facteurs humains
Stade D	**Maîtrisée**
D1	Attribut d'intégration
D2	Attribut d'amélioration
D3	Attribut d'itération
Stade E	**Institutionnalisée**
E1	Attribut de leadership centré sur l'humain
E2	Attribut de centrage sur l'humain dans l'organisation

TABLE 3.1 – Représentation des stades de maturité dans la prise en compte de l'utilisabilité par les organisations, selon Earthy (1998) [54]

Le premier stade, stade X – Inconnue, renvoie aux prémices, le fait qu'aucun indicateur de prise en compte de l'utilisabilité dans la démarche centrée-utilisateur n'est déployé. Á ce stade-là, l'entreprise n'a pas encore conscience de la nécessité d'être à l'écoute des problématiques des utilisateurs, et est plutôt auto-centrée. Évoluer depuis ce stade nécessite qu'un acteur du réseau de parties prenantes s'empare de ce sujet pour l'évangéliser. Le deuxième stade, stade A – Connue, adresse la reconnaissance des problèmes liés à la prise en compte de l'usage. Cet éveil de conscience s'accompagne d'une volonté d'améliorer les processus de conception, et de se focaliser davantage sur l'utilisateur. Porté par un faible nombre d'acteurs, l'UX à ce stade-là ne bénéficie souvent que de peu, voire pas, de budget. Ce qui est le prérequis pour passer au stade suivant. Le troisième stade, stade B –

Considérée, désigne l'étape où l'organisation commence à déployer des moyens permettant d'impliquer l'utilisateur dans la conception. De cette manière, l'UX intègre l'organisation et est considérée en termes de méthodologie dans les processus. Mais bien que de plus en plus présente, l'UX nécessite d'être portée par des profils spécialisés qui peuvent intervenir sur une diversité de projets de manière rigoureuse. Dans le quatrième stade, stade C – Implémentée, l'UX s'ancre durablement dans l'organisation avec une équipe dédiée qui peut faire participer les utilisateurs aux différentes phases de conception et valoriser ses compétences en termes de facteurs humains. L'utilisateur est alors au centre du processus pour accumuler un maximum d'informations sur ses besoins, contraintes, capacités, ressentis, etc. liés à l'interaction avec l'artefact en développement. Á force d'itérer dans ces processus de conception, l'entreprise gagnera en expérience pour passer au stade suivant. Pour le cinquième stade, stade D – Maîtrisée, l'UX est pleinement intégrée, l'utilisateur est au centre des processus de conception, la méthodologie UX est partagée avec les autres parties prenantes et permet d'être force de proposition. Des retours sont reçus et en cours d'intégration quant à ce qu'il faut améliorer et itérer. En prenant une place dominante dans la conception, l'UX se voue à devenir un pilier dans la culture de l'entreprise. Dans le sixième et dernier stade, stade E – Institutionnalisée, l'UX est devenue transversale et vecteur d'innovation. Sa culture atteint un niveau lui permettant de centrer toute approche possible de l'entreprise sur l'humain. Ses outils et méthodes sont partagés avec toutes les parties prenantes.

Avec une considération similaire, Plewes et Fraser font le constat que de nombreuses entreprises considèrent l'UX et les facteurs Humains comme des éléments non-centraux dans la proposition de valeur de leurs produits, et se focalisent parfois, uniquement sur des aspects esthétiques [64]. Partant de ce constat, les chercheurs proposent à leur tour un modèle représentant les stades de maturité UX d'une entreprise en se basant sur six indicateurs :

— Le moment de l'implication de l'UX dans le processus de conception et de développement.
— L'expertise et les ressources UX en interne et/ou la capacité à faire appel rapidement à l'expertise UX en cas de besoin.
— L'utilisation de techniques et de produits appropriés pour obtenir et comprendre les commentaires des utilisateurs.
— Le leadership et la culture de l'entreprise, c'est-à-dire dans quelle mesure les

dirigeants, et l'entreprise dans son ensemble, apprécient la valeur et la nécessité de l'UX Design d'un point de vue commercial.

— Le degré de connexion et d'intégration des processus UX avec les autres processus de l'entreprise qui permettent aux individus de travailler ensemble pour créer une bonne UX.

— La manière dans laquelle le *Design Thinking* - approche créative orientée vers la résolution de problèmes en mettant l'accent sur l'empathie envers les utilisateurs, la collaboration interdisciplinaire et l'itération - est appliqué dans la perspective la plus large possible pour favoriser une expérience client cohérente.

Bien qu'ils théorisent les six indicateurs clés présentés précédemment, seuls trois d'entre eux sont présentés dans leur modèle (voir figure 3.2).

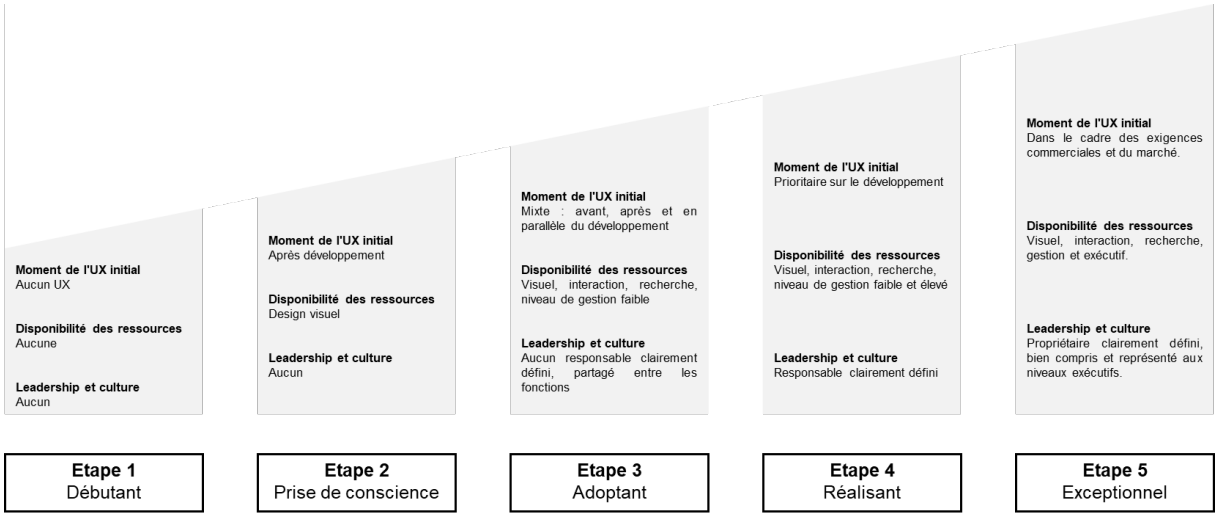


FIGURE 3.2 – Modèle de maturité UX des organisations, selon Fraser et Plewes (2015)

Ce modèle vise à aider les organisations à comprendre et à évaluer leur niveau de maturité dans la prise en compte de l'UX. Le modèle décrit cinq étapes clés de l'initiation à l'UX du stade initial jusqu'au stade dit exceptionnel. Chaque étape est accompagnée d'implications spécifiques, de signes clés indiquant le niveau de maturité de l'organisation et de facteurs de réussite critiques pour passer à l'étape suivante. L'étape 1, intitulée "Débutant", met en évidence le manque de conscience et d'expertise en matière d'UX Design au sein de l'organisation, où l'UX est perçue comme une couche superficielle ajoutée à la fin du processus de développement. L'étape 2, "Prise de conscience", décrit une organisation qui commence à considérer l'UX mais manque de structure et de compréhension claire de sa véritable nature. L'étape 3, "Adoption", montre les difficultés et les succès mitigés

de l'organisation dans l'adoption de pratiques UX plus sophistiquées. Les investissements sont faits, mais il existe un risque de stagnation ou de régression. L'étape 4, "Réalisation", met en évidence les organisations qui ont dépassé les débats sur l'importance de l'UX et se concentrent sur son amélioration concrète. Les objectifs UX sont clairement définis et intégrés dans l'ensemble de l'organisation. Et enfin, l'étape 5, "Exceptionnel", décrit les organisations qui ont pleinement réalisé les objectifs précédents et ont intégré l'UX Design dans tous les aspects de prestations tournées vers l'expérience client. Elles sont reconnues pour leur excellence en matière d'UX et utilisent cela comme un avantage concurrentiel. Ce modèle souligne l'importance d'une prise de conscience, d'une éducation et d'une formation adéquates pour passer d'une étape à l'autre. Il met également en évidence l'importance d'une direction stratégique solide et de la coordination entre les différentes fonctions de l'organisation pour atteindre un niveau élevé de maturité UX.

En 2006, Nielsen propose son modèle de maturité UX (voir tableau 3.2) qui vise notamment à permettre aux entreprises de se situer dans la mise à disposition de ressources dédiées à l'utilisabilité et l'intégration d'une conception centrée-utilisateur [128].

Ce modèle de maturité UX, proposé par Nielsen, offre un cadre d'évaluation de l'UX principalement sous le prisme de l'utilisabilité. Il décrit huit étapes de maturité UX, allant de l'hostilité envers l'expérience utilisateur à l'entreprise centrée sur l'utilisateur. Chaque étape se caractérise par le degré d'implication de la direction, la qualité et la fréquence des processus UX, la présence et le rôle d'une équipe UX, et les résultats obtenus en termes de satisfaction et de fidélisation des utilisateurs. Le modèle de maturité UX de Nielsen cherche vise à aider les organisations à identifier leurs forces et leurs faiblesses en matière d'UX, et à définir des actions pour améliorer leur maturité et leur performance. Il comporte six niveaux :

1. Hostilité : L'entreprise ne se soucie pas de l'UX et peut même y être opposée. Elle considère que les utilisateurs doivent s'adapter à ses produits ou services, et non l'inverse.
2. Centrée-développeurs : L'entreprise confie la conception de l'UX aux développeurs, sans faire appel à des experts et/ou à des utilisateurs. Elle se base sur des hypothèses ou des préférences personnelles, sans tenir compte des besoins réels des utilisateurs.
3. Bricolage : L'organisation reconnaît l'importance de l'UX, mais ne lui accorde pas de ressources suffisantes. Elle compte sur des initiatives souvent isolées de quelques employés.

Étape de maturité UX	**Caractéristiques**	**Temps nécessaire pour accéder à l'étape suivante**
Hostilité	Les développeurs ne veulent simplement pas entendre parler des utilisateurs ou de leurs besoins	Jusqu'à des décennies
Centrée-Développeurs	L'équipe de conception répond à sa propre intuition	2 – 3 ans
Bricolage	Recherche utilisateur en mode guérilla ou experts de l'utilisabilité externe	2 – 3 ans
Budget dédié	L'utilisabilité est prise en compte et pensée dans le projet	2 – 3 ans
Géré	Quelqu'un se charge de penser à l'utilisabilité dans toute l'organisation	6 – 7 ans
Processus systématique	Suivi de la qualité de l'expérience utilisateur	6 – 7 ans
Conception centrée-utilisateurs intégrée	Utilisation de données d'utilisabilité pour déterminer ce que l'entreprise devrait construire	20 ans
Entreprise axée sur les utilisateurs	L'utilisabilité influence la stratégie d'entreprise et les activités au-delà de la conception d'interface	40 ans à partir du début

TABLE 3.2 – Modèle de maturité UX des organisations, selon Nielsen (2006)

4. Dédiée : L'entreprise crée une équipe dédiée à l'UX, qui dispose d'un budget et d'une autorité propres. Cette équipe implique les utilisateurs dans la conception, et utilise des méthodes UX.
5. Institutionnalisée : L'entreprise intègre l'UX dans sa culture et sa stratégie. Elle forme tous ses employés aux principes de l'UX, et met en place des processus et des standards pour assurer la qualité et la cohérence de l'UX.
6. Visionnaire : L'entreprise innove en matière d'UX, et se positionne comme un leader dans son domaine. Elle anticipe les besoins et les attentes des utilisateurs, et crée des produits ou services qui leur apportent une valeur ajoutée.

En 2010, Renato Feijo s'intéresse aux stratégies UX mises en place par les entreprises. Il les définit comme un ensemble d'actions coordonnées pour obtenir une bonne expérience utilisateur à l'échelle organisationnelle. Feijo différencie la stratégie UX de la conception UX, expliquant que la stratégie ne doit pas se confondre avec les objectifs de conception des

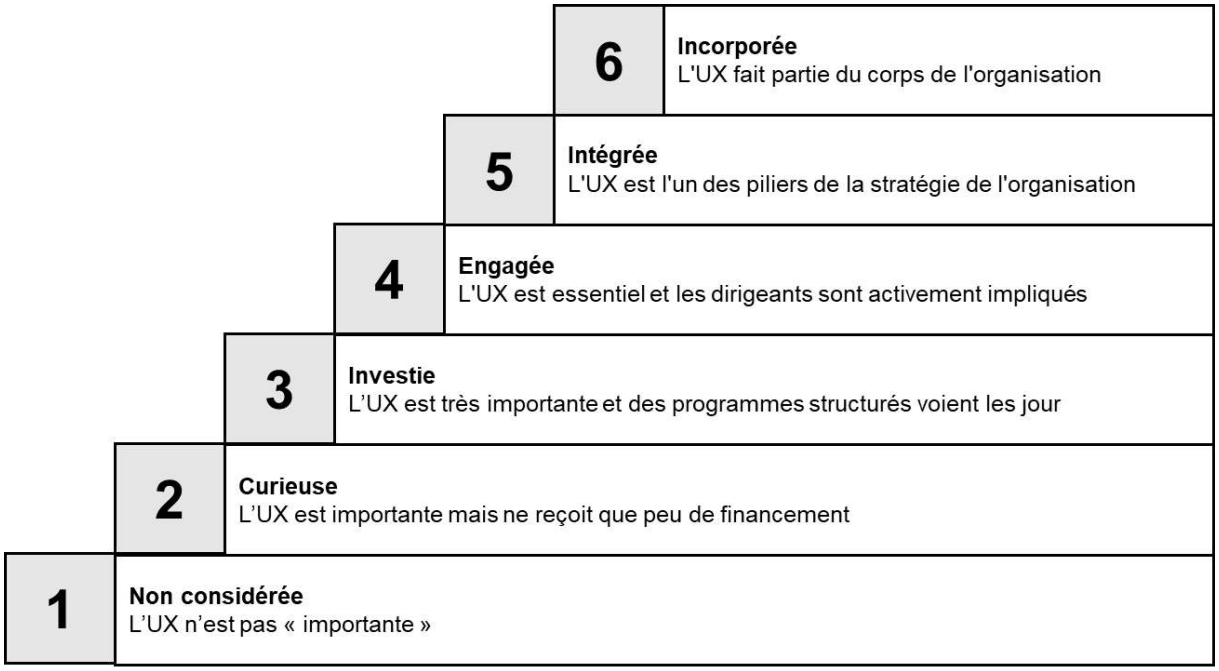


FIGURE 3.3 – Niveaux de maturité en stratégie UX, selon Feijo (2010)

projets. Pour Feijo, la capacité d'offrir une suffisamment bonne UX passe par la capacité de l'entreprise à coordonner les efforts des différentes parties prenantes (marketing, ingénierie, développement produits, ventes, etc.). Mais cette coordination passe par la stratégie UX, qui doit être abordée à des niveaux supérieurs, organisationnels (voir figure 3.3). Une stratégie bien définie propose un plan des actions à mener pour élargir sa culture globale et ses objectifs stratégiques, généralement sur une période de 3 à 5 ans [58].

Dans son modèle, Feijo propose 6 niveaux de maturité UX, desquelles découlent les choix stratégiques mis en place :

— Non considérée - L'expérience utilisateur (UX) n'est pas jugée pertinente, l'accent est mis sur ce que l'organisation estime important.
— Curieuse - L'importance de l'UX est reconnue ; le sujet commence à susciter des discussions.
— Investie - L'UX est considérée comme essentielle, des activités liées à l'approche de conception UX commencent à émerger.
— Engagée - L'UX est perçue comme cruciale, les acteurs s'impliquent activement.
— Intégrée - L'UX est considérée comme une des stratégies fondamentales de l'organisation.
— Incorporée - L'UX n'est plus un sujet distinct, elle est intrinsèque à la structure

organisationnelle.

Dans cette sous-section, nous avons exploré la notion de prise en compte de l'UX dans le contexte des entreprises. Nous pouvons constater que l'UX, malgré son importance croissante dans le paysage numérique, n'est pas toujours correctement abordée par toutes les entreprises, en raison de contraintes diverses telles que le manque de ressources, de compétences ou de sensibilisation [155]. Nous avons donc examiné différents modèles de maturité en termes de considération de l'UX [54] [128] [58] [64]. Ces modèles décrivent des étapes graduelles d'adoption et d'intégration de l'UX, allant de l'absence totale de prise en compte de l'UX à son intégration pleine et entière dans la culture et les pratiques de l'entreprise. Chacun de ces modèles met en évidence l'importance d'une sensibilisation, d'une formation et d'un leadership adéquats pour favoriser le passage d'une étape à l'autre. L'objectif de ces modèles reste de fournir des lignes directrices, un cadre précieux pour évaluer son niveau actuel de maturité UX, identifier les domaines d'amélioration potentiels et planifier les actions nécessaires pour s'améliorer. Ces modèles soulignent tous l'importance d'une prise de conscience, d'un engagement et d'une coordination à tous les niveaux de l'organisation, mais aussi de la nécessité d'un leadership fort et d'une culture organisationnelle propice à l'intégration de l'UX. Il est donc crucial pour les entreprises de comprendre et de suivre ces modèles pour concevoir des artefacts numériques qui répondent de manière efficace et efficiente aux besoins et attentes de leurs utilisateurs, tout en tenant compte des contraintes de leurs réseaux de parties prenantes.

3.3.2 Impact de l'UX sur le travail

Une majorité des investissements technologiques en milieu professionnel ont pour objectif d'accroître la productivité, d'améliorer la qualité de travail de l'humain ou encore de répondre aux nouveaux besoins émergeant de la société. Pourtant de nombreux cas ont montré le déploiement d'outils inadaptés à l'humain. Bien que les raisons puissent être multiples, nous portons notre attention sur le manque de considération envers l'UX, qui peut donner lieu, par exemple, à des interfaces inadaptées entraînant une surcharge cognitive [1] des utilisateurs cibles, et donc de l'indésirabilité. Comme nous avons pu le voir, la littérature en IHM met pourtant en avant que dès le début du processus de conception, il est nécessaire de prendre en compte les utilisateurs sous leurs différentes caractéristiques

1. État mental d'un individu engagé dans la réalisation d'une tâche extrêmement exigeante pour lui. Pour cette tâche, l'individu ne dispose pas des ressources cognitives suffisantes à une réalisation aisée de cette tâche.

(facteurs humains, besoins, perception et émotions ressenties en lien avec l'interaction) pour se prémunir de problèmes de viabilité de la solution proposée.

En 2013, Chaudet et ses collaborateurs ont démontré, au travers d'une étude de cas, l'intérêt de considérer les utilisateurs lors de la conception d'un dispositif dans le domaine médical. Pour cela, les chercheurs ont étudié la qualité d'usage d'un outil d'accompagnement du personnel soignant dès son implémentation, au sein d'un service d'urgence à Londres. La tâche prescrite de l'outil est de favoriser le confort du personnel soignant dans la gestion organisationnelle du service. Au bout de 4 ans, la moitié des fonctionnalités de l'outil ont été abandonnées. Les procédures initiales, ainsi que les capacités et perceptions des équipes soignantes n'étaient pas suffisamment prises en compte (mise en place de postes fixes malgré les nombreux déplacements, interfaces peu intuitives, problèmes techniques, etc.). L'usage de l'outil a donc été partiellement détourné par les utilisateurs cibles, ils l'utilisaient différemment de ce qui était prévu initialement. A l'issue de leurs travaux, les chercheurs avertissent qu'une considération insuffisante des facteurs humains lors de la conception de l'outil accentue l'apparition des freins à l'usage de certaines fonctionnalités prescrites [36]. Ils préconisent également que les utilisateurs aient des connaissances sur l'état du système et qu'un feedback cohérent leur soit apporté, faisant écho au fait que "plus un système est avancé, plus la contribution de l'opérateur humain [et sa compréhension de ce système sont cruciaux]" [16].

3.4 Quelle place pour l'UX dans la conception, le déploiement et l'utilisation de solutions IA en contexte professionnel ?

Dans le contexte de la prise en compte de l'UX pour déployer des solutions IA en entreprise, plusieurs points clés émergent de nos analyses théoriques et pratiques. Il nous apparaît, via notre revue de la littérature que l'intégration de l'UX dans les processus de conception est fondamentale pour garantir l'acceptation, l'efficacité et la satisfaction des utilisateurs finaux. La préoccupation principale pour les entreprises est de mettre à disposition une solution IA qui est digne de confiance. La solution IA doit permettre d'identifier clairement comment elle fournit ses résultats ou encore d'où proviennent ses données d'entraînement, et donc que la décision finale est prise de manière responsable,

qu'elle provienne de l'opérateur humain et/ou de la solution IA elle-même. Pour s'assurer de mettre à disposition une solution responsable, des lignes directrices aidant à guider la conception, le développement, le déploiement et l'utilisation de l'IA sont de plus en plus mises en avant, en particulier par des grandes entreprises technologiques [110]. Pour cet ensemble de principes, nous parlons alors d'IA Responsable (RAI, pour *Responsible Artificial intelligence*), qui représente des moyens de renforcer la confiance dans les solutions d'IA. La RAI met en avant que les systèmes d'IA ont un impact plus large que simplement permettre d'accomplir une tâche. Elle considère également un impact sur les valeurs sociétales et explique qu'il est donc nécessaire d'aligner ces technologies sur les valeurs des parties prenantes, les normes juridiques et les principes éthiques. Ces considérations mèneraient à une meilleure adoption et utilisation des solutions IA dans l'environnement professionnel.

Face à cette volonté de mobiliser davantage de moyens pour répondre aux recommandations de RAI et respecter les réglementations en vigueur pour favoriser l'IA de confiance, des acteurs industriels mais aussi académiques s'intéressent au développement de pratiques UX dans la conception et le déploiement de solutions IA. Par exemple, Wang et ses collaborateurs ont exploré dans leur étude comment les praticiens UX intègrent et examinent les considérations de la RAI durant les phases initiales de la conception de solutions IA au sein d'une grande entreprise technologique [177]. Leur travail révèle un besoin constant d'adaptation des rôles et pratiques pour tenir compte des enjeux éthiques, sociaux et réglementaires liés à l'IA. Les chercheurs ont interrogé un certain nombre de praticiens UX mêlés à des projets IA pour faire ressortir que la prise en compte de l'UX est essentielle dans ce type de projet pour répondre à trois pratiques émergentes :

1. La sensibilisation et l'éducation aux problématiques de responsabilité des solutions IA. Ce qui vise à instaurer un état d'esprit partagé où les considérations éthiques et sociales sont intégrées dans toutes les décisions de conception et de développement.
2. Le prototypage responsable qui aborde les défis d'interaction entre les opérateurs humains et les solutions IA. Cela vise à préparer les opérateurs humains à la nature parfois imprévisible de ces systèmes, pour aider à mieux comprendre leurs capacités et limites. Le prototypage responsable permet donc de valider les idées de conception mais aussi d'attribuer une représentation de l'outil plus ajusté au réel en détectant et atténuant d'éventuelles dissonances entre ce qui est attendu de l'outil et ce qu'il fait vraiment.

3. Et enfin, l'évaluation responsable qui souligne l'importance d'impliquer les utilisateurs finaux dans l'évaluation des solutions IA pour s'assurer qu'elles sont fonctionnelles et qu'elles répondent aux principes éthiques et sociaux. Cette pratique met en avant la nécessité d'engagements à long terme avec les utilisateurs pour évaluer de manière adéquate l'impact et les implications des systèmes IA.

Sur la base de cette lecture, nous estimons que valoriser l'UX dans la conception et le déploiement de solutions IA en entreprise a pour objectif de s'aligner avec les besoins réels, les attentes et les contextes d'utilisation des employés. En mettant l'accent sur des démarches de prises en compte de l'UX, les entreprises peuvent améliorer l'efficacité et la productivité de leurs services, en s'assurant que les solutions IA sont non seulement efficaces (permettant d'atteindre le but donné) mais également intuitives, pertinentes et surtout voulues par les utilisateurs finaux. En se centrant sur l'utilisateur, cette approche vise à surmonter les barrières potentielles à l'adoption de l'IA, telles que la résistance au changement, la méfiance envers les technologies automatisées et les préoccupations liées aux représentations négatives (remplacement/perte d'emplois, déshumanisation du travail, soumission à la solution IA) [12] [182].

L'impact de l'UX sur l'adoption de l'IA en entreprise semble donc significatif. Une UX bien conçue peut faciliter la transition vers des systèmes plus automatisés en contribuant à réduire la courbe d'apprentissage de l'utilisation de la solution IA et en augmentant la confiance des utilisateurs envers ce même outil. [12]. Cela est particulièrement pertinent dans des contextes où les décisions prises par l'IA peuvent avoir des conséquences importantes, nécessitant une compréhension claire des processus et des résultats par les utilisateurs finaux. De plus, une UX efficace peut contribuer à démysthifier l'IA, en rendant ses processus plus transparents et compréhensibles, ce qui est crucial pour bâtir la confiance et l'acceptabilité parmi les employés [86] [12] [174] .

En conclusion, l'intégration des méthodes UX dans la conception, le déploiement et l'utilisation de solutions IA en contexte professionnel est essentielle pour maximiser leur potentiel et assurer leur succès. Une attention particulière à l'UX peut transformer la manière dont les solutions IA sont perçues et utilisées dans le milieu professionnel, conduisant à une plus grande efficacité, satisfaction et adoption par les utilisateurs finaux. Cependant, cette intégration requiert une approche multidisciplinaire pour relever les défis techniques et éthiques associés.

3.5 Critique et limites de l'UX

Dans le domaine des IHM, la prise en compte de l'UX est de plus en plus présentée comme fondamentale. Ceci est notamment dû au nouveau regard apporté qui intègre les composantes instrumentales [2] et non-instrumentales [3] [115] [70]. Mais l'UX fait également l'objet de critiques dans la littérature. Premièrement, l'UX est parfois perçue comme une simple tendance répondant à des intérêts principalement marketing, sans apporter d'innovation suffisante pour révolutionner les paradigmes liés à l'utilisabilité. Deuxièmement, l'approche par l'UX Design est critiquée pour son manque de validité scientifique avec une démarche empirique jugée peu rigoureuse [18] [70].

Cependant, il est important de préciser que ces critiques ne reflètent pas nécessairement notre positionnement. Selon nous, les problèmes d'intérêts marketing et de manque de validité dans la démarche sont souvent le résultat de l'appropriation des méthodes UX par des profils non spécialisés en UX. Ces acteurs, sans formation adéquate, peuvent détourner les principes de la démarche à des fins superficielles ou mal alignées avec les objectifs originaux de cette discipline. Cette appropriation peut conduire à une application simpliste et incorrecte des méthodes, nuisant à la perception globale de leur validité et de leur valeur ajoutée. Par conséquent, les critiques adressées à l'UX peuvent en réalité être des critiques envers une mauvaise mise en œuvre des pratiques UX, plutôt qu'envers les principes fondamentaux des méthodes de prise en compte de l'UX elles-même.

De plus, la notion d'UX est parfois imbriquée ou confondue avec celle d'utilisabilité, comme nous l'indique la norme ISO 9241-210 [88]. Ce qui amène certaines équipes de conception à prendre parti pour l'un ou l'autre des termes. Mais l'utilisabilité, introduite par la première fois en 1984 par Eason, examine les réactions des utilisateurs face à un système informatique. Il en ressortira un premier modèle d'analyse ergonomique, qui prend en compte la facilité d'apprentissage du système, sa facilité d'utilisation et le degré auquel la tâche définie est accomplie. L'utilisabilité sert davantage à évaluer le rapport coût-bénéfice de l'interaction pour l'utilisateur [55]. En 1998, l'utilisabilité a été officiellement définie par la norme ISO 9241-11 (1998) comme étant le "degré selon lequel un produit peut être utilisé par des utilisateurs identifiés pour atteindre des objectifs définis, avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifié" [88]. Nous soulignons que

2. Que l'on retrouve notamment dans l'utilisabilité.
3. Telles que les émotions.

l'UX va au-delà de la simple notion d'utilisabilité. Tandis que l'utilisabilité se concentre principalement sur l'efficacité, l'efficience et la satisfaction dans un contexte spécifique d'utilisation.

L'UX design englobe une gamme plus large de facteurs, y compris les émotions, les valeurs et les besoins des utilisateurs, qui peuvent influencer leur expérience globale avec un produit ou un service. Cette approche, plus holistique [102], permet de concevoir des solutions qui résonnent plus profondément avec les utilisateurs, en créant des expériences plus riches et plus engageantes [74]. En intégrant des principes issus de la psychologie, de l'ergonomie et autres sciences qui s'intéressent à la compréhension des utilisateurs, l'UX Design favorise une compréhension plus nuancée des interactions humaines avec la technologie [73]. Cette interdisciplinarité permet d'adresser des problématiques complexes et de proposer des innovations significatives dans la conception de produits et services numériques [181].

Malgré cela, il ne faut pas négliger que les notions d'utilisabilité et d'UX sont interconnectées. L'UX encapsule les dimensions de l'utilisabilité, présentées comme instrumentales, et y ajoute des dimensions non-instrumentales, relatives aux sentiments et attitudes d'un utilisateur découlant de l'esthétique, la symbolique et la motivation dégagées par l'interaction avec l'artefact de façon longitudinale [18] [61] [70]. L'UX propose une perspective bien plus large que celle de l'utilisabilité, mais ce qui la rend également d'autant plus complexe à cerner et à étudier.

3.6 Conclusion

La notion d'expérience utilisateur est au fur et à mesure du temps devenue une notion holistique des réactions de l'utilisateur liées à un artefact. Elle est autant liée à l'utilisabilité qu'à la perception de l'artefact, la compréhension de ce dernier, ou encore les émotions suscitées. Elle est temporelle et contextuelle. Cette notion a réussi à devenir essentiel dans la conception d'artefact, en particulier dans le domaine du numérique. Mais pour autant, elle semble encore avoir tendance à être négligé, ce qui a conduit à la création d'échelle de mesure de maturité en UX des entreprises pour identifier leur degré de prise en compte du concept.

Son implication dans la conception de solutions IA semble de plus en plus nécessaire face à à la complexité dans l'appropriation, compréhension et collaboration avec ces dernières. Configurer les solutions IA en prenant en compte l'UX faciliterait l'interaction avec l'humain pour atteindre la tâche, dans des conditions favorables à l'utilisateur.

Conclusion de la Partie I

Cette première partie se focalise sur les théories autour de l'acceptabilité des solutions IA en contexte professionnel. Nous y considérons l'acceptabilité comme un concept essentiel au bon déploiement de ce type d'outils dans un environnement de travail. En effet, l'acceptabilité est une attitude à l'égard de l'usage d'un artefact par anticipation, c'est-à-dire que les collaborateurs vont exprimer un certain degré d'intention d'usage à partir des représentations et perceptions qu'ils ont des solutions IA. Un degré admissible d'acceptabilité et d'acceptation contribue à l'adoption des solutions IA et donc à leur inscription dans les usages des collaborateurs. Notre exploration de la littérature montre que les perceptions des solutions IA sont très disparates et que les intérêts à leur déploiement sont très changeants selon que l'on soit décideurs, concepteurs ou utilisateurs. Mais ce qui reste commun à ces différentes parties prenantes est le besoin de comprendre le fonctionnement du modèle IA dans la solution pour pouvoir cadrer correctement son usage, savoir à qui revient la responsabilité des décisions prises par ou avec l'outil et comment le positionner dans l'environnement de travail par rapport au collaborateur. Mieux comprendre les solutions semble ainsi un élément essentiel pour accroître la confiance que leur accordent les collaborateurs, parfois de manière similaire à des interactions Humain-Humain. Et lorsque l'on a confiance en l'outil avec lequel on collabore, alors on est d'autant plus susceptible de le trouver acceptable.

Notre état de l'art met également en avant que pour rendre les solutions IA acceptables, il est nécessaire que les entreprises s'intéressent aux perceptions, besoins et contraintes de chacun des acteurs de ces projets. Pour cela, il faut employer une méthodologie de prise en compte des usages, mais leur mise en place nécessite une certaine maturité de la part des entreprises à ce sujet. Nous explorons donc les méthodes utilisées par les entreprises pour mettre l'utilisateur au centre des processus de conception et d'évaluation, avec pour principal objectif de favoriser l'expérience utilisateur (UX). L'UX désigne les ressentis liés à l'usage d'un outil de manière holistique, donc autant sur les aspects d'utilisabilité de l'outil que sur les aspects émotionnels et contextuels. Pour une entreprise, gagner en maturité sur les démarches UX est essentiel dans la conception et l'évaluation des solutions

IA. Cela est susceptible de contribuer à l'identification des facteurs d'acceptabilité de ces outils comme la confiance. D'après nos recherches, rendre ces outils plus acceptables contribue l'adoption de ces solutions IA en contexte professionnel et de surcroît, contribue à la collaboration entre les utilisateurs et les systèmes IA.

La problématique centrale de cette thèse réside donc dans l'identification des facteurs clés influençant l'acceptabilité des solutions IA en contexte professionnel. Elle cherche à répondre à des questions cruciales telles que la manière dont les perceptions et attentes des utilisateurs finaux affectent leur intention d'usage de ces nouveaux outils professionnels. Mais elle s'intéresse également aux mécanismes qui peuvent être mis en place pour garantir une adoption consentie et durable de ces technologies. En abordant ces aspects, la thèse vise à offrir un cadre méthodologique pour intégrer efficacement les solutions IA dans les environnements de travail, tout en tenant compte des besoins et des attentes des utilisateurs.

La suite de ce manuscrit sera consacrée à la présentation des différents travaux que nous avons réalisés sur des projets de conception et déploiement de solutions IA. Ces travaux s'inscrivent dans une volonté d'explorer et d'identifier 1) les facteurs d'influence potentiels de l'acceptabilité des solutions IA et 2) des cadres efficients de collaboration Humain-IA permettant une meilleure intégration dans les environnements professionnels.

Deuxième partie

Mesure de l'acceptabilité des outils d'aide à la prise de décision

Chapitre 4

Mobilisation des méthodes UX pour étudier la place des utilisateurs dans le processus de conception des projets SIGMA

Dans ce chapitre

Nous présentons ici les différents projets IA mis à disposition par SIGMA Informatique et ses partenaires. Après avoir présenté les études ethnographiques réalisées sur ces terrains, nous nous intéresserons au degré de maturité dans la prise en compte de l'UX pour la conception de ces projets. Nos recherches montrent une volonté des parties prenantes de s'orienter vers de la conception centrée utilisateur, mais avec prudence. Les équipes de conception étant le plus souvent centrées sur l'outil, parfois au détriment du projet.

4.1 Introduction

Les méthodes d'enquête s'avèrent cruciales dans la méthodologie UX, car elles offrent un aperçu des attitudes, des comportements et des besoins des utilisateurs. Par exemple, les méthodes d'entretien, bien que plus en proie à la subjectivité, fournissent des données qualitatives qui peuvent faciliter la compréhension des utilisateurs et autres parties prenantes (leurs préférences, leurs craintes, leurs motivations, leurs habitudes, etc.), pour ensuite impacter la conception des produits aux services qui leur sont destinés. Nous avons donc privilégié ce type de méthodes dans la récolte de données relatives à l'étude de la

place des utilisateurs dans la conception, le déploiement et l'utilisation des solutions IA au sein de SIGMA Informatique.

Nous présentons, ici, les résultats des entretiens menés auprès d'un échantillon de professionnels au sein de SIGMA Informatique, qui constitue le réseau de parties prenantes à la conception et/ou au déploiement de solutions IA décrites dans la section 4.3. L'objectif principal de ces entretiens était de comprendre comment les différents acteurs liés aux projets IA en cours perçoivent ces solutions technologiques, ainsi que leurs applications potentielles. Les participants aux entretiens provenaient de divers horizons professionnels, incluant des rôles techniques (tels que développeur-concepteur, architecte cybersécurité, etc.), ainsi que des rôles plus orientés métier (comme responsable de service, managers). Les développeurs-concepteurs interrogés étaient des alternants en IA, ce qui a permis d'identifier la vision des acteurs de l'IA qui sont au début de leur carrière dans ce domaine.

4.2 Comment les parties prenantes des projets IA perçoivent-elles ces solutions ?

Comme amorcé précédemment, nous explorons dans cette section les perceptions des parties prenantes à différents projets IA que nous avons suivis au sein de SIGMA Informatique, ainsi que chez certains de ses partenaires. Pour cela, nous avons choisi de procéder à des entretiens semi-directifs avec les collaborateurs participant, ou ayant participé, à ces projets. Pour sélectionner nos répondants, nous sommes d'abord allés voir les équipes de conception des projets IA, puis les utilisateurs finaux, les décideurs/responsables et enfin toutes personnes susceptibles d'être associées à ces projets et/ou recommandées par les autres répondants (ex. architecte cybersécurité qui s'intéressent au sujet dans les projets, responsables opérationnel du service où sont réalisés ces projets). Comme nous pouvons le voir dans la figure 4.1, l'entretien est composé de trois, voire quatre dimensions.

Tout d'abord, nous interrogeons le collaborateur sur son poste de travail, plus précisément nous cherchons à contextualiser son cadre de travail. Cette partie de l'entretien vise à identifier le type de tâches qu'il effectue, la durée depuis laquelle il occupe le poste ainsi que ses habitudes de travail, notamment s'il travaille principalement en équipe ou individuellement, et s'il opère sur site, chez le client, ou en télétravail, etc. Ces informations permettent de comprendre l'environnement professionnel direct du collaborateur et d'évaluer si cela peut influencer sa perception des solutions IA et sa manière d'interagir

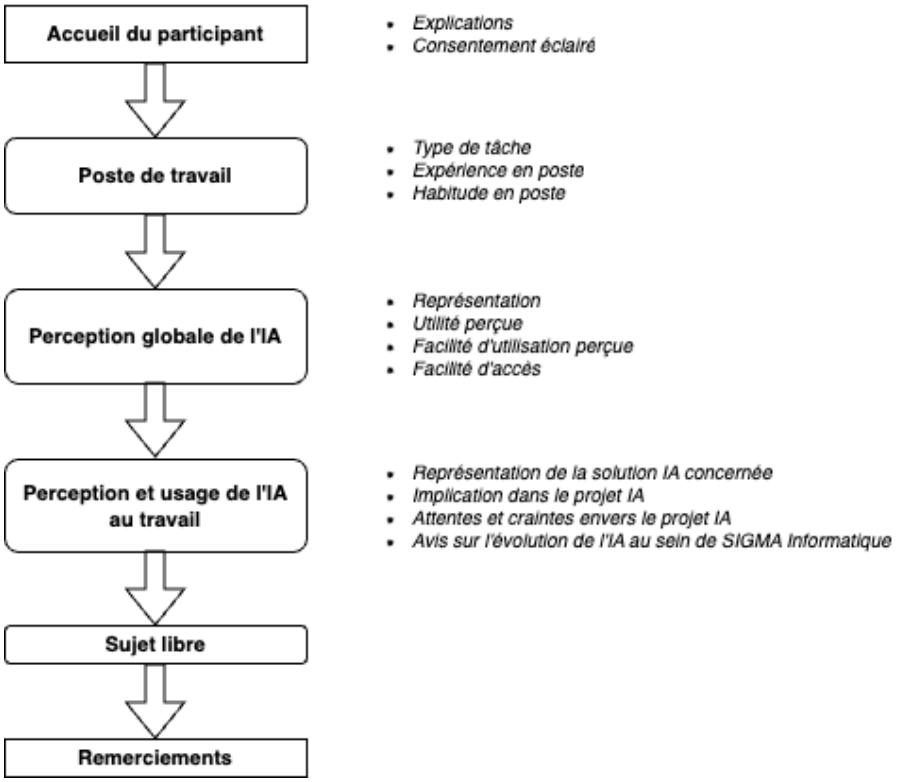


FIGURE 4.1 – Déroulé des entretiens semi-directifs auprès des parties prenantes aux projets IA au sein de SIGMA Informatique

avec. Ensuite, nous explorons sa perception globale de l'IA. Nous y recueillons notamment les représentations que les collaborateurs se font de l'IA, mais également à quel point il trouve facile d'accès et d'utilisation les technologies qui en découlent. Nous l'interrogeons également sur l'utilité perçue de l'IA, tant sur le plan personnel que sur le plan sociétal, en considérant les impacts possibles sur l'emploi et la société dans son ensemble. Et enfin, nous nous concentrons spécifiquement sur la solution IA concernée par le projet en cours. Nous interrogeons le collaborateur sur sa perception et l'usage de la solution IA envisagée sur le(s) poste(s) de travail concerné(s). Il est important de déterminer quelle solution IA est déjà connue et utilisée, ainsi que le degré d'implication du collaborateur interrogé dans la conception, et de savoir si la solution IA a été introduite pour remplacer, assister ou encore collaborer avec les utilisateurs finaux sur les tâches cibles. Nous examinons également les attentes et les craintes du collaborateur vis-à-vis de l'outil conçu, ses premières impressions lors de l'introduction de l'outil dans son environnement de travail, ainsi que sa vision de l'impact futur de l'IA au sein de SIGMA Informatique ou de ses partenaires.

Ces discussions approfondissent notre compréhension de l'intégration de l'IA dans les activités quotidiennes et ses répercussions potentielles sur les individus et les processus organisationnels. Nous finalisons l'entretien en laissant la possibilité au collaborateur interrogé de s'exprimer librement sur le sujet.

Ces entretiens auprès d'une vingtaine de collaborateurs nous montrent que les perceptions des solutions IA sont très variables parmi les interviewés. Nous y trouvons un éventail de perceptions plutôt positives avec des solutions IA décrites comme des outils d'optimisation qui facilitent la prise de décision, éliminent les tâches "ingrates" et offrent des gains de productivité. Par exemple, un répondant a dit être "convaincu des gains de productivité en automatisant certaines tâches [et qu'il reconnaît] l'utilité accrue de l'IA sur les tâches récurrentes". Mais nous retrouvons également des perceptions bien plus nuancées avec des répondants qui expriment plutôt des réserves face au "coût élevé associé à l'IA et la tendance à vouloir intégrer de l'IA là où ce n'est pas forcément nécessaire" [1].

Avec le rapide développement des solutions IA ces dernières années, certains perçoivent son utilisation plus comme une tendance que comme une nécessité. En croisant les profils métiers avec les retours obtenus, nous nous apercevons que les principaux décideurs (responsable de service, d'équipe) envisagent les solutions IA comme un bon axe de développement économique qui nécessite un haut taux de performance pour se rendre acceptable. Comme exploré précédemment dans le chapitre 1 et le chapitre 2, la seule performance des solutions IA ne suffit pas à leur acceptabilité. Des profils plus techniques tels que les concepteurs-développeurs et un architecte cybersécurité reconnaissent "les gains de productivité et l'utilité de l'IA pour des tâches spécifiques", mais à condition que de nombreux efforts soient faits dans la conception. Ce qui en l'état actuel semble les décourager, en l'absence de stratégies et directives précises. Ces propos rejoignent la perception d'un profil plus axé sur la prise en compte des utilisateurs cibles (UX Leader), qui mentionne ses préoccupations concernant les ressources à la bonne réalisation des projets IA (ex. expert en Data/IA, temps nécessaires, coût financier du projet) et la "tendance à intégrer []des solutions IA] là où ce n'est pas nécessaire".

Plusieurs axes émergents ont été identifiés lors de ces entretiens. Parmi eux, la volonté

1. Ces verbatims sont extraits d'entretien concernant l'utilisation de l'IA en général en entreprise et non pas d'une solution spécifique.

des collaborateurs interrogés de bénéficier d'une cartographie des compétences au sein de SIGMA Informatique pour la mise en œuvre de projets IA. En effet, la discipline qu'est l'IA représente un domaine qui peut nécessiter une pluralité d'acteurs d'orientations différentes, le besoin d'accompagnement pour renforcer les compétences en IA, et une réflexion profonde sur l'acceptabilité des solutions IA. Parmi les autres sujets saillants, il y a l'importance accordée à la formation et à l'accompagnement pour maîtriser les technologies IA. Mais aussi les besoins de clarification de l'objectif sous-jacent au déploiement de la solution, parfois ressenti comme un outil d'aide et d'autres fois plutôt comme remplaçant potentiel de l'expertise humaine [59] [66] [65].

Finalement les entretiens révèlent une perception nuancée des solutions IA dans ce contexte. Si certains n'hésitent pas à exprimer leur enthousiasme face à son potentiel sur les situations de travail, d'autres expriment leurs préoccupations concernant son coût, sa mise en œuvre et son acceptabilité. Enfin, ces entretiens soulignent l'importance de l'accompagnement, de la formation et de la réflexion stratégique pour réussir l'adoption de ce type de solution.

4.3 Étude ethnographique de la prise en compte de l'UX : de l'approche technocentrée à l'approche centrée utilisateur

Comme énoncé précédemment dans la section 4.1, nous avons étudié certains terrains en contexte professionnel lié à la conception et aux déploiements de solutions IA sur des situations de travail. Avec la disponibilité des données pour nourrir les modèles et l'ingéniosité des algorithmes, les solutions IA permettent un réel confort d'assistance pour répondre à des problèmes initialement coûteux pour l'humain. Ces mutations préoccupent cependant de plus en plus les organisations sur le plan éthique. Elles s'interrogent sur la prise en compte de l'humain dans la réalisation des tâches et de l'usage qui peut être fait de ces outils [86]. Pourtant il semble qu'actuellement peu d'études s'intéressent à la manière dont les entreprises prennent en compte les travailleurs [2]. Dans un contexte générique, il est commun de faire pleinement profiter l'utilisateur d'un outil en faisant le nécessaire pour qu'il ait une expérience agréable. Mais qu'en est-il des opérateurs ? Eux qui ont rarement le choix de leurs outils professionnels, mais qui subissent plutôt leurs

implantations suite à la décision de leur hiérarchie [59] [65] [142].

Au travers de cette recherche, notre objectif est de présenter les pratiques actuelles mobilisées par le SIGMA Informatique et ses partenaires dans le domaine de l'IA et les conséquences sur les collaborateurs et sur la réalisation de la tâche. Ces constats émanent d'un travail exploratoire et d'une étude ethnographique sur des projets mis à disposition par SIGMA informatique et ses partenaires. Dans cette section, nous étudierons donc 1) les motivations de ces entreprises à implanter ce type de solution, 2) la place donnée aux travailleurs qui vont les utiliser, 3) les moyens déployés pour favoriser leur acceptabilité et 4) si les méthodes UX sont en adéquation avec ce type de projet pour mieux comprendre les besoins et perceptions des collaborateurs. Face à ces interrogations, nous avons exploré quatre terrains d'étude présentant les collaborateurs comme augmentés par l'IA : une aide à l'infogérance, un chatbot d'assistance pour le support informatique, une aide à la commande dans la grande distribution et un outil d'aide au traitement de rapports diagnostiques.

4.3.1 Optimisation du Capacity Planning

Contexte du projet Ce premier terrain concerne une équipe du service infogérance de SIGMA Informatique. Son rôle est notamment la gestion d'une partie de l'activité des clients à leur place (ex. hébergement de données, cybersécurité, gestion des infrastructures informatiques, etc.). Dans le cadre de ce projet, le focus est fait sur l'hébergement de données de clients. Dans un premier temps, un audit a été réalisé par un prestataire spécialisé dans l'exploitation de la donnée et l'automatisation par l'IA. Cette analyse a mis en lumière un intérêt économique à optimiser la tâche de gestion de l'hébergement de données en accompagnant le service responsable de celle-ci dans la prédiction des besoins à venir de baies de stockage. Cette tâche est attribuée à une seule personne : le gestionnaire de plateforme, qui estime ponctuellement (une à deux fois par an) l'évolution des besoins de la clientèle en termes de stockage afin de déterminer la quantité de baies de stockage à commander pour les centres de données (*datacenters*). On parle de *Capacity Planning*[2]. Cette tâche a de grandes répercussions sur l'activité de l'entreprise. Si la commande de baies de stockage est insuffisante, il faudra en recommander pour répondre aux besoins de stockage de la clientèle, mais sûrement à un tarif plus élevé (un achat global permet de

2. Ce terme désigne le fait de déterminer/prédire la capacité de stockage nécessaire répondant à l'évolution de la demande

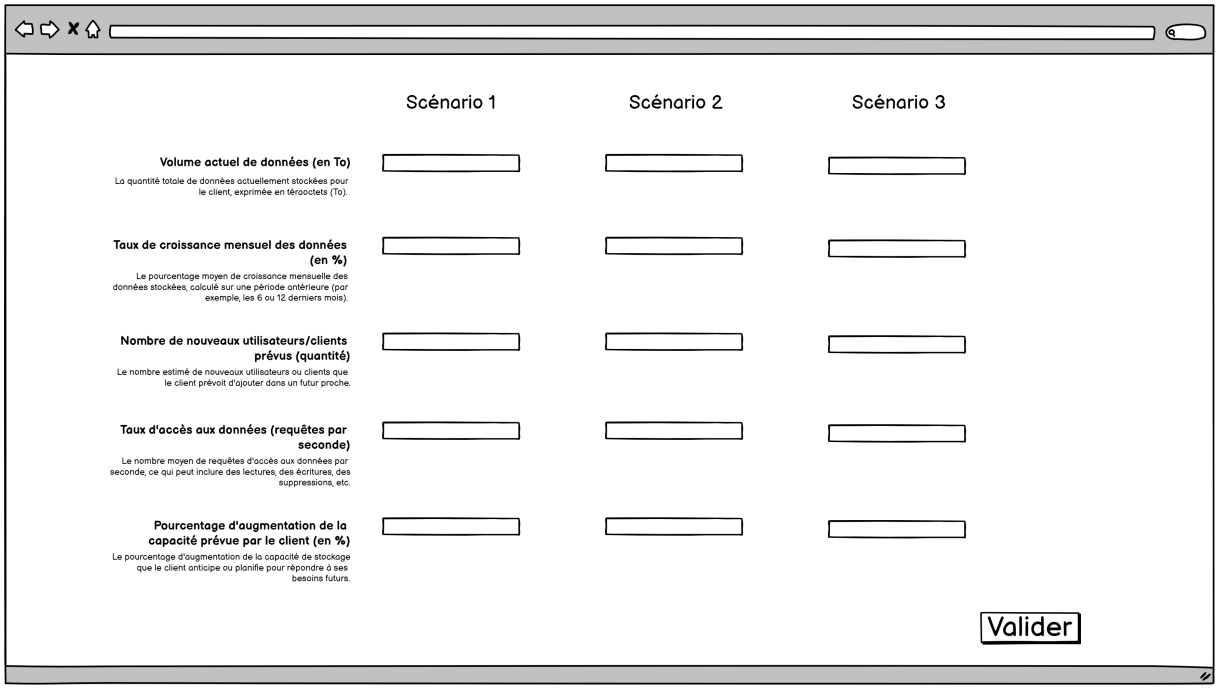


FIGURE 4.2 – Maquettage du dashboard de sélection des scénarios d'activité

bénéficier de tarifs préférentiels). Au contraire, si la commande de baies est trop importante par rapport aux besoins réels, il y aura des baies de stockage non allouées et donc une perte financière.

Objectifs du projet IA Trois étudiants alternants d'une école de formation à la conception de modèles IA ont été recrutés pour travailler sur ce projet. Ils ont eu pour objectif de concevoir un outil d'aide à la décision. L'outil développé repose sur deux fonctionnalités : la centralisation des données des clients sur une même interface et la prédiction des besoins globaux de stockage sous forme d'un modèle auto-régressif [3] en renseignant trois scénarios d'activités potentiels pour moduler les prédictions de l'outil (voir figure 4.2). La rentabilité attendue de ce projet est que le modèle soit suffisamment précis pour être au plus proche de la réalité afin de faciliter la prédiction du gestionnaire et de lui faire gagner du temps. Le gestionnaire de plateformes a été sollicité à la fin du processus de développement pour tester l'outil, il a également déclaré à l'équipe de conception qu'il en était satisfait.

3. Modèle statistique de régression pour séries temporelles. Les données d'une série y sont uniquement expliquées par ses valeurs passées plutôt que par d'autres variables.

Notre intervention Dans le contexte de notre recherche, nous n'avons pas été en mesure d'observer directement la conception et le déploiement de l'outil, notre intervention ayant eu lieu a posteriori de son déploiement. Nous avons donc commencé par interroger l'équipe de conception (les alternants et le responsable du projet) à visée exploratoire. Un entretien non directif a été réalisé pour obtenir des détails sur le processus de conception. Celui-ci a aussi permis de préciser comment cette équipe a perçu les enjeux du projet et la manière dont l'utilisateur a été intégré.

Pour approfondir ce point, nous avons également procédé à un entretien avec cet utilisateur unique. Il présente son implication dans le projet comme tardive (phase de tests finaux), dès lors qu'il a pris l'initiative de comparer ses propres prédictions basées sur son expérience à celles de l'outil. Il se dit effectivement satisfait d'un outil qui facilite sa tâche en centralisant les données clients et dont les résultats sont suffisamment similaires à ses prévisions, mais obtenus bien plus rapidement. Face à une tâche qu'il ressentait comme coûteuse initialement — au moins deux jours de travail à plein temps —, il perçoit divers avantages à l'outil déployé : le gain de temps qui le soulage face à une tâche qu'il décrit comme parfois trop stressante, une meilleure visibilité des volumétries de données qui sont centralisées sur une même interface et des prévisions liées aux scénarios d'activités qu'il envisage. Il décrit l'outil comme un "vrai soulagement" et admet lui accorder sa confiance puisque celui-ci "a fait ses preuves en termes de performances". Si l'outil est utilisé pour des projections de manière ponctuelle, le gestionnaire de plateformes explique également que les prévisions de celui-ci ne remplacent pas ses propres prédictions, et qu'il s'en sert plutôt pour justifier ses commandes. Il explique également avoir conscience de certaines limites, telles que la linéarité des prédictions, ne lui permettant pas d'observer les potentielles variations (ex. une baisse des volumétries qui pourrait signifier une baisse d'activités de la clientèle) (voir figure 4.3). Une autre limite identifiée est que l'usage ponctuel de l'outil complique sa tâche, car il rencontre parfois des difficultés à se remémorer les informations à renseigner dans les champs de l'interface. De plus, les informations renseignées dans l'outil sont affichées en nombre de fichiers (ex. deux fichiers de type Application), ce qui complique sa compréhension des résultats de l'outil, ne connaissant pas la volumétrie réelle des données.

Analyse de la démarche UX De manière similaire à une approche UX, les concepteurs ont souhaité prendre en compte l'usage de l'outil en s'intéressant aux besoins initiaux du

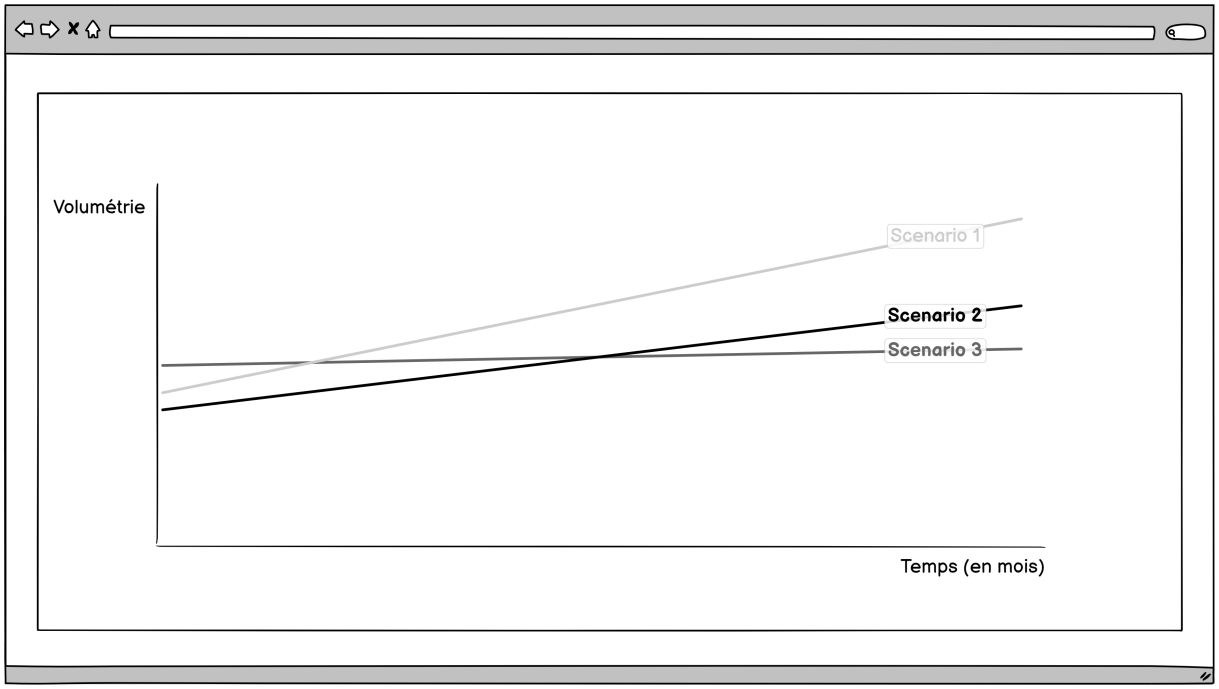


FIGURE 4.3 – Maquettage du dashboard de visualisation des scénarios d'activité

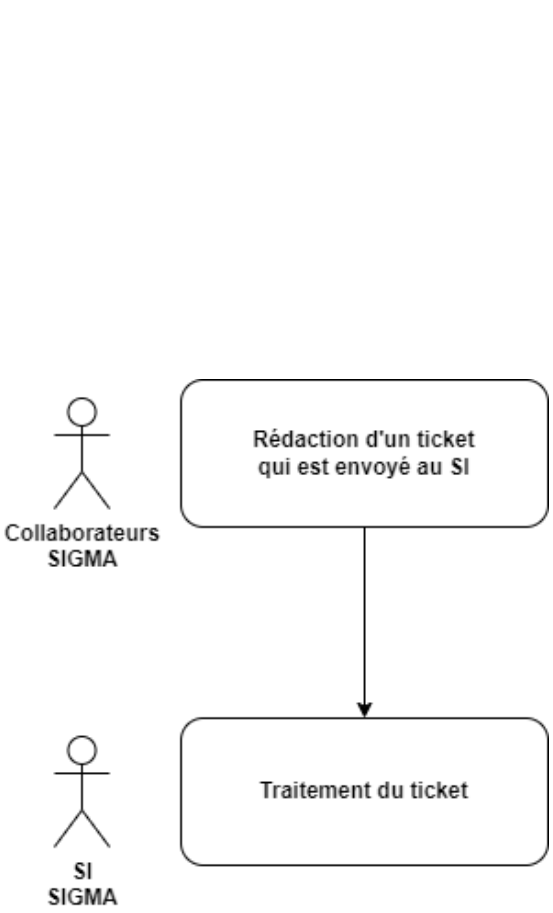
gestionnaire, d'autant que ce dernier a pu tester la solution. Cependant, son inclusion dans la conception s'est faite très tardivement et l'absence d'itérations, après que le gestionnaire a testé la solution, n'a pas permis d'aller au bout de la démarche et de prendre en compte les besoins émergents. En se basant sur l'approche de l'acceptabilité de Nielsen (voir figure 2.1), celle de Capacity Planning est partielle. Le gestionnaire de plateformes ressent de la satisfaction dans la confirmation de ses prédictions, mais cette dernière reste limitée par les freins qu'il identifie. La mémorabilité (ou facilité à apprendre) est faible puisqu'il rencontre des difficultés à se remémorer quelles informations sont à renseigner dans l'interface. C'est un point qui montre une utilisabilité déjà restreinte par un problème lié à l'ergonomie de l'interface. Et si l'utilité théorique de cet outil était de faire des prédictions pour aider l'utilisateur, en pratique c'est plutôt la centralisation de données qui prime. En conclusion, l'hypothèse la plus probable est que résoudre ces problèmes d'ergonomie réorienterait l'intérêt de l'utilisateur sur les fonctionnalités primaires de l'outil (prédictions).

4.3.2 Support augmenté : assistance du service de support informatique

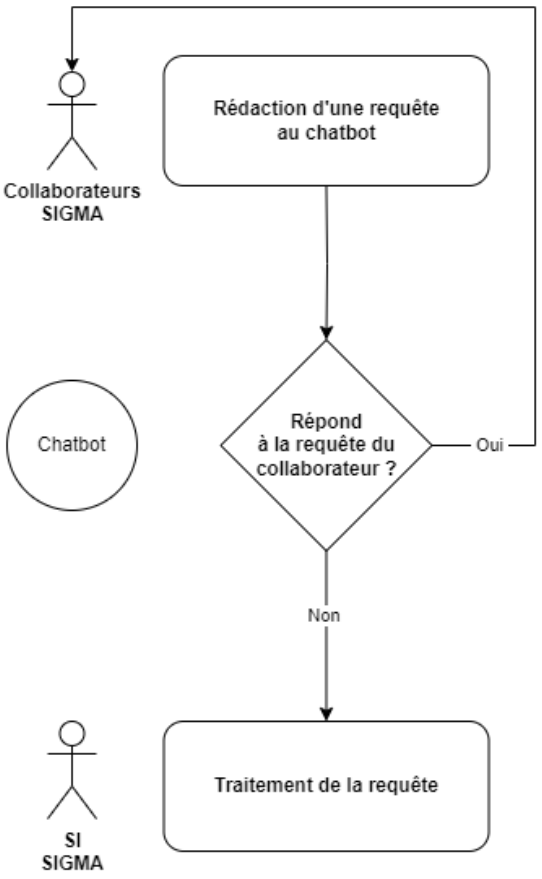
Contexte d'intervention Proposé par le même audit que pour le premier outil, ce deuxième terrain concerne le déploiement d'un agent conversationnel (*chatbot*) pour accompagner le service de support informatique (SI) de SIGMA Informatique qui s'occupe de la gestion des requêtes et des déclarations d'incidents liés aux matériels physiques et logiciels des postes de travail. Ces requêtes et déclarations prennent la forme de tickets, rédigés par les collaborateurs SIGMA eux-mêmes. Face à des enjeux d'amélioration de la satisfaction des collaborateurs, de la qualité de service et d'optimisation du support, une proposition de semi-automatisation robotisée des processus (RPA) a été faite.

Objectifs du projet IA Ce chatbot prend place entre les collaborateurs et le SI afin d'effectuer un pré-traitement des tickets (voir figure 4.4). La stratégie de conception mise en place par l'équipe de cadrage est de faire monter en compétences trois étudiants alternants, recrutés dans la même école que ceux du premier projet, et de développer un socle technique réutilisable pour d'autres cas d'usages. La plus-value de ce projet réside également dans ses objectifs de gain de temps avec l'optimisation de la recherche d'informations par les collaborateurs, estimée à environ dix minutes en moyenne. Mais il s'agit aussi de les faire gagner en autonomie dans la gestion de certaines requêtes et d'alléger la charge de travail du SI. Le chatbot a été conçu sur la base d'outils à disposition sur une plateforme applicative (service numérique hébergeant diverses applications, données et services). L'équipe de conception est donc partie d'une solution déjà existante et mise à disposition. Le chatbot est initialement accompagné de nombreux ensembles additionnels contenant près d'un millier d'énoncés, des requêtes formalisées que l'outil est en mesure de traiter. Mais ces énoncés sont génériques. Pour que l'outil réponde à la tâche prescrite, il faut renseigner manuellement un maximum d'énoncés (utterances) que l'utilisateur pourrait adresser à l'outil, en considérant que plusieurs énoncés peuvent désigner la même requête (voir figure 4.5).

Notre intervention Notre approche est similaire à celle du premier projet : nous nous sommes focalisés sur la récolte de données subjectives et qualitatives auprès des parties prenantes, l'objectif étant de mieux cerner la considération des utilisateurs dans le processus de conception, les critères d'interaction choisis et leur acceptabilité par les parties prenantes. Le déploiement de cet outil entraîne également d'importants changements dans



(a) Processus traditionnel de gestion des tickets des collaborateurs SIGMA



(b) Optimisation du système de ticketing avec l'intégration d'un chatbot pour prétraiter les tickets

FIGURE 4.4 – Évolution du processus de gestion des incidents, où un chatbot réalise le prétraitement des tickets

l'interactivité entre le support et les collaborateurs. Cette réflexion nous a conduit à interroger le support pour mieux cerner leur perception de la solution censée être déployée. Analyser le processus de conception est également un moyen de se rendre compte de la manière dont l'équipe de conception a priorisé l'implantation des énoncés. Lorsque l'on interroge le responsable du support, il ne semble pas croire en l'utilité de l'assistant virtuel, mais plutôt craindre que cela entraîne du travail supplémentaire pour son équipe, dû à un outil insuffisamment performant. Pour les membres de l'équipe de conception, le niveau de performance que doit atteindre l'assistant virtuel pour être considéré comme acceptable et industrialisable n'est pas suffisamment clair. Ils déclarent avoir un sentiment de "manque de moyens techniques" qui les conduit à n'avoir que peu de visibilité sur comment peut

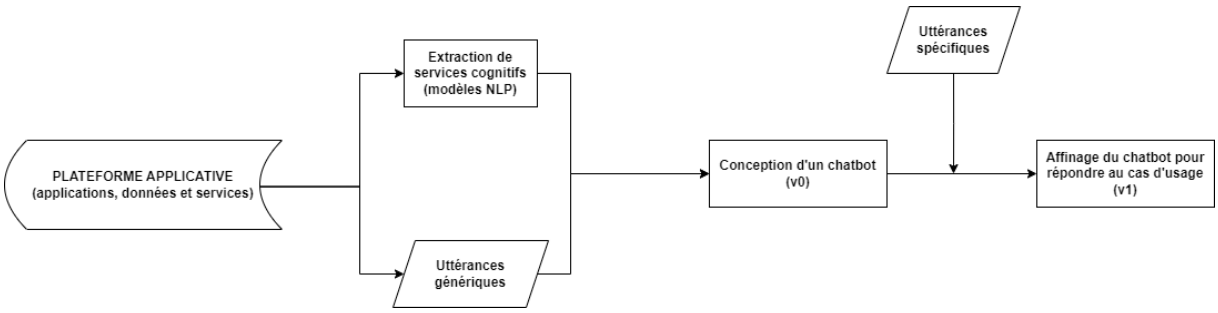


FIGURE 4.5 – Processus de conception du chatbot d'assistance du SI

s'implanter l'outil. En complément de cette préoccupation, le nombre conséquent d'énoncés envisagé "représente une importante charge de travail" pour un outil dont ils doutent qu'il soit utilisé. Un décalage s'est installé entre les attentes de l'équipe de conception et l'analyse d'activité du prestataire. Face à un trop grand effort de conception identifié, le projet s'est vu récusé, n'a donc finalement pas abouti, conduisant notre intervention à être faite de manière rétrospective.

Analyse de la démarche UX Dans ce cas de figure, le RPA est téléguidé par une analyse de besoins pour orienter les capacités du chatbot, mais pour autant, aucune stratégie de prise en compte de ces besoins n'est proposée par la chefferie de projet. Les utilisateurs cibles n'ont d'ailleurs pas été sollicités durant toute la conception. Le service support envisage finalement que l'outil aurait une faible efficience et des difficultés à prévenir les erreurs, si nous nous référons à nouveau à l'approche de l'acceptabilité de Nielsen (voir figure 2.1). De plus, la difficulté à prendre en compte les besoins et attentes des utilisateurs a été exprimée lors d'un audit externe, mais non exploitée par l'équipe de conception. Ce qui a fortement impacté les dimensions subjectives de favorabilité envers l'assistant virtuel, et finalement nui à sa conception, malgré une certaine maturité des algorithmes de traitement du langage. Les croyances, attentes et attitudes envers l'outil étant relativement défavorables de la part des concepteurs et du service support, la sphère sociale de l'acceptabilité en fut fortement impactée.

4.3.3 Aide à la prédiction de commandes de marchandise

Contexte d'intervention Ce troisième terrain implique l'implantation d'un outil d'aide aux commandes promotionnelles au sein d'une entreprise de la grande distribution. Face à des représentants de points de vente (supermarchés, hypermarchés, commerces de

proximité) qui expriment, lors d'ateliers participatifs, rencontrer des difficultés à prévoir les ventes sur des périodes données, le SI de cette entreprise a proposé d'intégrer une fonctionnalité d'aide à la prédiction de ventes dans l'outil de commande déjà en place.

Objectifs du projet IA Un prestataire spécialisé en élaboration de modèles IA a accompagné le SI pour concevoir le modèle prédictif et lui permettre de devenir autonome dans sa maintenance et la gestion des variables sur lesquelles repose le modèle. Dix-huit mois de données de ventes ont été récoltés pour créer le modèle actuel à destination des points de vente. Dans l'interface utilisateur de l'outil, les prédictions de ventes sont accompagnées d'indicateurs numériques [4] (voir figure 4.6). Ces indicateurs ont pour vocation de donner plus de crédit aux prédictions du modèle. L'idée est d'être plus transparent sur les prédictions du modèle en donnant à l'utilisateur une partie des éléments sur lesquels il s'appuie pour donner son résultat. L'objectif de cette mise à jour "intelligente" est de permettre aux points de vente de gagner du temps sur les tâches de commande et d'opérer avec le moins d'écart possible entre la commande et les ventes réelles.

Notre intervention Notre intervention a démarré à la suite des interrogations de la direction du SI concernant l'évaluation de l'acceptabilité de leur outil. La direction souhaitait également bénéficier d'une méthodologie permettant d'assurer la réussite de l'implantation de l'IA. Une attention particulière est portée par l'entreprise sur les éléments de communication mis en place pour promouvoir le déploiement de la fonctionnalité, mais aussi les potentiels freins à l'usage. Comme le SI ne bénéficie pas d'outil objectif pour mesurer l'usage effectif de l'outil et le gain de temps, nous gardons une perception floue du nombre de points de vente utilisant effectivement l'outil, de même que des gains en temps et en performance ainsi que de l'UX associée.

Nous avons commencé par retracer tous les efforts liés à la prise en compte de l'expérience utilisateur dans ce projet. Nous souhaitions vérifier que la conception répondait de

4. Ces indicateurs représentent respectivement les quantités vendues :
   1. lors de la dernière période promotionnelle,
   2. durant le même mois de l'année précédente,
   3. durant le même mois de l'année précédente par des points de vente de superficie similaire,
   4. durant le même mois de l'année précédente par des points de vente situés dans la même zone géographique.

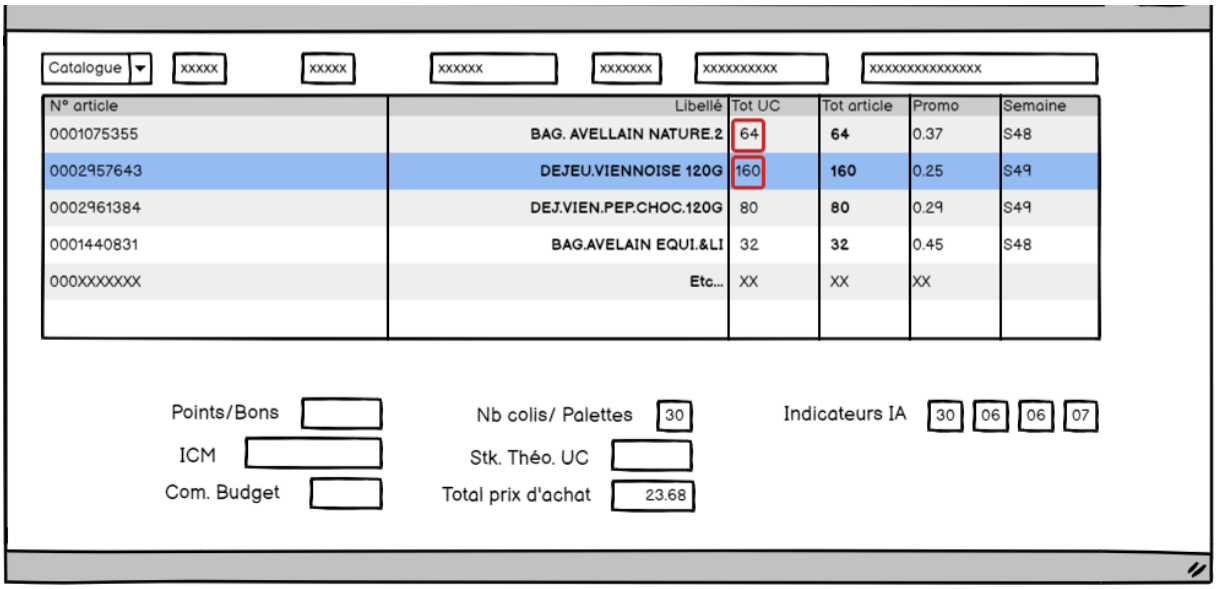


FIGURE 4.6 – Maquettage du dashboard avec les indicateurs IA (les recommandations IA sont pré-saisies dans le dashboard. Les valeurs entourées en rouge dans la colonne "Tot UC" sont les valeurs pour lesquels l'utilisateur a saisi une quantité différente de la prédiction IA. La ligne bleue est la ligne de commande sur laquelle l'utilisateur a cliqué. Les indicateurs IA en bas à droite sont affichés quel que soit le paramétrage, ils sont dépendants de la ligne de commande sur laquelle l'utilisateur clique et ils ne sont pas étiquetés, ils n'ont pas d'en-tête dans le dashboard)

manière adéquate aux besoins identifiés des différents points de vente. Afin d'approfondir comment les utilisateurs cibles ont été pris en compte dans la conception, nous avons donc proposé des méthodes de sondage pour estimer les perceptions des utilisateurs de la fonctionnalité : des entretiens et un questionnaire. La première intervention est un entretien individuel avec vingt employés missionnés aux commandes promotionnelles. L'objectif de ces entretiens est d'obtenir des données qualitatives et individuelles afin d'identifier la perception de l'implantation de solution IA selon trois dimensions : l'IA en général, les éléments de communication mis à disposition pour favoriser l'acceptabilité de la fonctionnalité, et l'IA dans leur point de vente pour identifier si elle est en adéquation avec la situation spécifique de ces derniers. La deuxième intervention proposée consiste en un sondage d'un échantillon plus large par la diffusion de questionnaire (200 répondants envisagés). L'objectif est de solliciter un échantillon bien plus vaste qu'avec les entretiens pour valoriser la représentativité et mettre en avant une approche plus quantitative. Ce questionnaire vise à mesurer quatre dimensions distinctes :

— Les informations générales pour identifier les proportions de points de vente ayant répondu à l'enquête en fonction du type de point de vente, de leur activité et de leur localisation géographique. Ces données servent également à identifier les profils des gestionnaires de stock qui sont amenés à utiliser l'outil de commande pour effectuer les commandes promotionnelles

— L'utilisabilité perçue de l'outil, qui est une notion fortement liée à l'acceptabilité selon la littérature, comme nous avons pu le voir dans les chapitres 2 et 3 [129]

— L'acceptabilité, qui est la notion centrale de l'intervention et que l'on définit comme étant les facteurs conduisant à l'intention d'usage de l'artefact [129] [165]

— La confiance, qui est mise en avant, dans la littérature, comme facteur fondamental de l'adoption d'une solution [67]. Dans cette étude, nous faisons donc de la confiance le second élément central de notre enquête, au vu de la place que ce concept occupe dans l'acceptabilité des solutions IA, comme nous avons pu en discuter dans les chapitres 1 et 2. Bien que les "indicateurs IA" visent à donner plus de transparence sur les décisions du modèle pour favoriser la confiance accordée, les utilisateurs sont cependant trop éloignés du processus de décision dans le projet et ces indicateurs sont donc susceptibles de ne pas être la donnée la plus pertinente pour favoriser leur confiance. Pour la bonne raison qu'ils n'ont pas pu donner leur avis sur ces informations.

Ces interventions n'ont cependant pas reçu de retour favorable et n'ont donc pas été

réalisées, la principale raison semble liée à une politique interne. Au sein de cette coopérative de magasins, les retours d'expérience sur l'usage des outils se font principalement de manière ascendante : s'ils le souhaitent, les points de vente font remonter leurs problèmes et remarques liés à l'outil par des rapports d'incidence, des appels au SI ou lors de groupes de travail. Face à cette hiérarchie récalcitrante à donner davantage de poids aux utilisateurs dans la conception de l'outil, une des hypothèses possibles est que le contexte économique les incite à maximiser leur performance en réduisant la marge de manœuvre octroyée au personnel. Mais nous supposons également qu'elle privilégie des cycles courts de production pour les mêmes raisons. Rapportée à notre étude, il faut entendre que l'IA pourrait être un moyen de réduire l'agentivité du personnel tout en maintenant ou en accroissant la productivité.

4.3.4 Centralisation de rapports diagnostiques

Présentation de la solution Ce quatrième et dernier terrain explore un outil numérique de saisie de diagnostics immobiliers. Cela comprend le diagnostic de performance énergétique, l'état d'amiante, le diagnostic bruit, etc. En tout, la solution couvre plus d'une dizaine de diagnostics relatifs aux obligations réglementaires de bonne tenue de logement. Les utilisateurs ciblés par cet outil sont les diagnostiqueurs de bailleurs sociaux [5].

Concrètement, les bailleurs sociaux ont accès, via la solution, à tous les diagnostics réalisés et à réaliser, le plan des logements concernés et le calendrier des travaux réalisés sur leurs terrains immobiliers. Toutes ces informations sont centralisées sur une même interface. Les utilisateurs ont la possibilité d'y personnaliser leurs tableaux de bord et les données sont également partageables selon leur volonté. Les utilisateurs ont également la possibilité de connecter la solution à leurs autres outils métier, et de générer et consulter des fiches récapitulatives et de synthèse de l'état des logements.

A partir du moment où ils souscrivent à la solution, les bailleurs sociaux renseignent les diagnostics directement sur l'interface. Mais pour les besoins de centralisation de l'information, ils doivent renseigner les précédents diagnostics dans l'outil. Il n'existe pas de norme universelle de rapport diagnostique dans l'immobilier en France, ce qui fait que les diagnostiqueurs peuvent proposer des rapports de format très différent, mais qui

5. Tels que les Offices Publics de l'Habitat (OPH), les Sociétés d'Economie Mixte (SEM) ou encore les Entreprises Sociales de l'Habitat (ESH).

renseignent le même type d'informations. Face à ça, une entreprise prestataire (que nous appellerons prestataire en charge de la transcription) effectue la tâche de transcription des rapports antérieurs (au format PDF) vers la solution. Cette tâche est relativement fastidieuse puisqu'elle nécessite d'explorer et référencer l'intégralité des anciens rapports diagnostiques de chaque bailleur-client dont la taille peut être assez conséquente (jusqu'à 200 pages pour un seul rapport), et que le format peut énormément changer d'un rapport à un autre (voir figure 4.7). Ce travail renvoie à un coût de prestation assez élevé. En conséquence de quoi, le responsable de l'équipe en charge de la gestion de cette solution chez SIGMA Informatique a envisagé d'automatiser une partie, voire l'entièreté, de cette tâche de transcription grâce à l'IA.

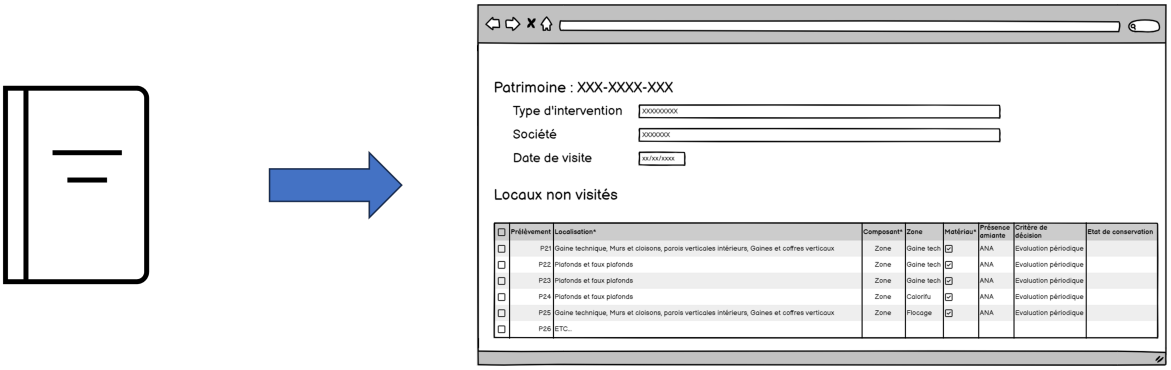


FIGURE 4.7 – Maquettage du dashboard de l'outil de centralisation des diagnostics immobiliers

Ce qui a été proposé par un autre prestataire, spécialisé dans le déploiement de modèles IA sur-mesure, est de concevoir une solution IA capable d'extraire l'information pertinente du rapport diagnostique et de la placer dans la solution de centralisation. Nous avons pu accompagner cette équipe sur la conception de l'outil avec le prestataire en question et étudier l'acceptabilité de l'IA par le prestataire en charge de la transcription dont les salariés sont les premiers impactés.

Objectifs du projet IA Étant donné que le référencement des informations des rapports diagnostiques représente un vrai coût, intégrer une solution IA qui extrait automatiquement l'information pertinente pour la transcrire dans la solution cible permettrait de faciliter cette tâche et ainsi de réduire ce coût. Plus précisément, l'objectif est que le modèle explore le contenu du rapport PDF, en faisant une analyse sémantique et graphique, puis prédise quelle information est pertinente pour les champs à remplir dans la solution cible.

Sur le moyen terme, la volonté est de transformer l'activité du prestataire qui fait le référencement. Les utilisateurs n'auraient presque plus à enregistrer les informations, mais feraient principalement de la vérification des informations transcrites par le modèle IA.

Notre intervention Nous sommes intervenus dans ce projet dès le début de la conception, l'objectif était d'accompagner l'équipe en charge de la solution cible qui s'interrogeait sur l'acceptabilité de solutions IA pour les tâches d'extraction de données et les stratégies de déploiement à mettre en place pour faciliter leur adoption. Dans un premier temps, nous avons rencontré le responsable de la société prestataire qui fait le travail de transcription pour identifier l'écosystème dans lequel sont transcrites les informations. Notre première intervention se présentait comme une analyse heuristique, c'est-à-dire que nous découvrions comment était réalisée la tâche avec un esprit critique et analytique pour évaluer l'utilisabilité actuelle de la solution-cible (avant intégration du modèle IA) et à quel degré le modèle IA pourrait contribuer à améliorer la situation de travail et la productivité des employés. Basés sur l'évaluation heuristique, nous souhaitions identifier les problèmes potentiels d'utilisabilité [129]. Nous nous basons principalement ici sur les heuristiques de Bastien et Scapin [152] qui comporte 8 critères ergonomiques :

- Le guidage, les moyens mis en oeuvre pour assister l'utilisateur dans l'utilisation de l'outil
- La charge de travail, les moyens mis en oeuvre pour réduire la charge perceptive et mnésique de l'utilisateur en réduisant le nombre d'informations que l'utilisateur doit prendre en compte
- Le contrôle explicite, le degré de maîtrise que l'utilisateur conserve sur les actions réalisées avec ou par l'outil
- L'adaptabilité, la capacité du système à réagir selon le contexte aux besoins et préférences de l'utilisateur
- Le traitement des erreurs, les moyens mis en place pour protéger l'utilisateur d'erreurs potentielles et lui permettre de les corriger.
- L'homogénéité, la cohérence global de l'interface de l'outil qui respecte une certaine logique d'utilisation
- La signifiance des codes et dénominations, l'adéquation entre l'objet ou l'information affichée ou entrée, et son référent
- La compatibilité, la capacité de l'outil à s'intégrer dans l'activité réelle en étant en favorisant l'accord entre les caractéristiques de l'utilisateur et les tâches définies

Ces informations ont été récoltées partiellement et leur fiabilité reste faible, car nous n'avons pu passer que très peu de temps à explorer l'interface et à voir les utilisateurs interagir avec. Le responsable de la structure a principalement transmis les informations relatives à l'utilisabilité, laissant peu de place aux employés pour exprimer leur ressentis et besoins à l'égard de l'outil avec lequel ils travaillent quotidiennement. Nous avons recensé les informations récoltées dans le tableau 7, situé en annexe.

Concernant l'intégration du modèle IA pour automatiser la transcription, nous exprimons un certain scepticisme. Comme développé dans la Partie I de ce manuscrit, il nous semble essentiel d'aller au contact des utilisateurs cibles pour étudier leur représentation de l'IA afin de choisir une stratégie adéquate d'accompagnement dans le cas de figure où on choisit de déployer ce type de technologie sur leur poste de travail. Dans ce projet, la stratégie des décideurs (chef de projet et financiers) est de s'assurer d'un certain niveau de performance de la part du modèle, avant de solliciter les utilisateurs pour comprendre leur perception de l'automatisation d'une partie de leurs tâches.

4.4 Évaluation de la maturité UX de SIGMA Informatique et de ses partenaires dans les projets IA

Dans cette section, nous explorerons le niveau de maturité du SIGMA Informatique et de ses partenaires dans sa capacité à prendre en compte l'expérience utilisateur dans les projets IA présentés précédemment. Ce travail d'investigation s'intéresse tant à la conception, qu'au déploiement et à l'utilisation finale des solutions IA. L'objectif est d'estimer comment ces entreprises abordent les aspects cruciaux, explorés dans le chapitre 3 sur la maturité UX, pour garantir une adoption réussie des solutions IA.

Bien que critiquée pour sa considération partielle des employés, l'intégration des utilisateurs dans la création de solutions IA montre une certaine adoption des principes UX par SIGMA Informatique et ses partenaires. Les principes UX, caractérisés par leur transdisciplinarité, polyvalence et capacité d'évolution, visent à une compréhension globale de l'interaction entre le dispositif, les utilisateurs et le contexte, en favorisant une approche de conception collaborative et itérative [102], qui ici n'a pas su se concrétiser. Les initiatives

lancées en ce sens ont cependant joué un rôle clé dans le développement d'innovations technologiques, visant à enrichir l'UX par une approche engageante, intuitive et accessible [121]. Cette méthode, n'étant que partiellement appliquée dans nos terrains, voit ainsi son efficacité limitée et mène parfois à des échecs. Malgré l'émergence de la pensée UX, ces entreprises montrent finalement une capacité et une volonté encore limitées à adopter une conception centrée sur l'utilisateur, ce qui nécessiterait des changements organisationnels profonds pour valoriser cette approche [138].

L'assistance à la gestion des commandes illustre bien l'approche participative, où les utilisateurs, experts de leurs besoins, contribuent activement au développement des solutions IA. Cependant, l'absence de test utilisateur approfondi et de compétences UX au sein des équipes de conception a conduit à des problèmes d'ergonomie, notamment dans l'agencement et la sélection des éléments dans l'interface, augmentant la charge cognitive des utilisateurs. Pour le chatbot d'assistance du SI, une analyse initiale a identifié des besoins connexes à la tâche prescrite, à savoir alléger la charge de travail du SI. La focalisation sur l'historique des demandes sans une compréhension approfondie des besoins réels des utilisateurs a limité la bonne réalisation du projet, soulignant l'importance d'une démarche UX plus engagée pour définir les priorités de développement.

Pour rappel, notre référence sur la maturité UX se base sur les travaux de Earthy et son modèle de maturité en utilisabilité [54], ceux de Nielsen avec son modèle de maturité UX des organisations, ceux de Fraser et Plewes et leur modèle de maturité UX [64], et enfin les travaux de Feijo et ces niveaux de maturité en stratégie UX [58]. Á partir de ces modèles, nous avons constitué un tableau récapitulatif d'où situe les projets IA en terme de maturité UX (voir tableau 4.1).

4.5 Conclusion

Nos travaux révèlent plusieurs points clés dans l'intégration des méthodes UX au sein du SIGMA Informatique et ses partenaires dans leurs projets IA. Bien qu'une certaine adoption des principes UX soit observée, celle-ci est critiquée pour sa considération partielle des intérêts des parties prenantes et des utilisateurs finaux en particulier, mettant en lumière la nécessité d'une approche plus centrée sur l'humain. Les initiatives prises pour intégrer l'UX dans les solutions IA, bien que cruciales pour le développement d'innovations

Modèle de maturité en utilisabilité (Earthy, 1998)	**Stade A**	**Stade B**	**Stade C**	**Stade D**	**Stade E**
	ASI	CP	CM + RD		
Modèle de maturité UX des organisations (Nielsen, 2006)	**Étape 1**	**Étape 2**	**Étape 3**	**Étape 4**	**Étapes 5 à 8**
	ASI	CP + CM + RD			
Modèle de maturité en stratégie UX (Feijo, 2010)	**Étape 1**	**Étape 2**	**Étape 3**	**Étape 4**	**Étapes 5 et 6**
	ASI	CP + CM + RD			
Modèle de maturité UX (Fraser et Plewes, 2015)	**Étape 1**	**Étape 2**	**Étape 3**	**Étape 4**	**Étape 5**
	ASI	CP + CM + RD			

TABLE 4.1 – Positionnement des projets IA sur les échelles de maturité (ASI : Assistance au SI, CP : Capacity Planning, CM : Aide à la prédiction de commande de marchandise, RD : Centralisation de rapports diagnostiques)

technologiques engageantes, intuitives et accessibles, sont souvent insuffisantes. Ce qui met également en avant que les projets sont encore très technocentrés. Pourtant nous réaffirmons qu'un socle technique performant ne suffit pas à assurer son acceptabilité.

Les projets étudiés montrent une volonté de déployer une démarche UX dans la mise en place de solutions IA, mais celle-ci reste souvent limitée par des problématiques organisationnelles et une compréhension incomplète des besoins réels des utilisateurs. Cette situation conduit à des problématiques de stratégie de conception, d'ergonomie et d'acceptabilité des solutions, limitant leur efficacité, leur efficience ou encore la satisfaction liée à l'usage, ce qui conduit parfois à des échecs. SIGMA Informatique et ses partenaires affichent ainsi une capacité et une volonté encore restreintes d'adopter pleinement une conception centrée sur l'utilisateur, nécessitant des changements organisationnels profonds pour valoriser et intégrer efficacement l'UX dans les projets IA. Cette problématique est peut-être commune aux différents projets de conception de solutions informatique sur mesure que propose SIGMA Informatique, mais elle est d'autant plus marqué sur les projets IA où il est nécessaire de s'assurer de l'adhésion des utilisateurs pour réduire les freins à l'usage [59] [2].

En conclusion, pour améliorer l'intégration des méthodes UX dans le processus de

conception des projets IA, il est essentiel pour SIGMA Informatique d'adopter une démarche plus centrée sur l'utilisateur. Cela implique de mobiliser des moyens et des profils métiers spécialisés dans la compréhension approfondie et continue des utilisateurs, mais également des profils experts dans l'IA, qui ont conscience de la difficulté à proposer des solutions viables techniquement tout en tenant compte des problématiques des opérateurs humains déjà en poste. Cela implique des ajustements organisationnels pour favoriser une collaboration transdisciplinaire et itérative entre les équipes de conception et les utilisateurs, ainsi qu'une volonté renforcée de mettre l'utilisateur au cœur de la conception des solutions IA.

Chapitre 5

Méthodes clés pour étudier l'effet de la confiance déclarée d'un outil prédictif sur la confiance accordée par les utilisateurs

Dans ce chapitre

Nous présentons, dans ce chapitre, une expérimentation qui découle des résultats obtenus dans l'exploration des projets IA présentés précédemment. Nous testons l'effet de la confiance déclarée d'une solution IA en ses propres résultats sur la confiance accordée par un opérateur humain. La tâche expérimentale consiste à prédire l'âge de portraits photos. Pour réaliser cette étude, nous utilisons des méthodes employées dans l'évaluation subjective de QoE. Nos recherches nous montrent 1) qu'il est possible de mesurer la confiance d'un humain envers une solution IA en termes d'accord Humain-IA et 2) que communiquer davantage d'informations pour justifier la prise de décision de la solution IA semble accentuer la confiance que l'humain lui accorde.

5.1 Introduction

Dans ce chapitre, nous explorons un autre aspect de la transparence que celui exposé par les XAI dans la chapitre 1 : la capacité du modèle à communiquer le degré de confiance qu'il accorde à ses propres prédictions. En nous basant sur la littérature sur la transparence

des solutions IA, nous estimons que l'indice de confiance du modèle (ou indice de certitude) peut influencer la confiance des utilisateurs dans les résultats du modèle [182] [134] [21]. De plus, dans ce chapitre nous questionnons également les méthodes traditionnellement utilisées pour mesurer la confiance que les utilisateurs accordent aux prédictions des outils d'aide à la décision. Plutôt que de nous appuyer uniquement sur des mesures subjectives de la confiance (via questionnaire par exemple), nous utilisons l'accord entre les prédictions de l'utilisateur et celles du modèle comme indicateur. Pour ce faire, nous transposons dans notre contexte expérimental des méthodes utilisées dans le cadre de l'évaluation subjective de la qualité de l'image [8] [9] [99]. L'objectif de notre étude est donc de contribuer aux recherches sur l'impact de la transparence des modèles IA sur la confiance que leur accordent les humains. Mais aussi, de mobiliser une diversité de méthodes afin de mesurer la confiance dans les solutions IA.

5.2 Contexte théorique

La littérature en IHM met de plus en plus en avant qu'il est nécessaire de favoriser la confiance pour s'assurer de l'adoption des solutions IA [158] [157] car il est déterminant pour l'usage volontaire de ces outils d'en comprendre les tenants et aboutissants pour s'y fier. Si l'utilisateur ne fait pas confiance à la solution IA, alors il aura tendance à être récalcitrant et à ne pas vouloir s'en servir. Cette vision s'est constituée au regard que l'IA génère encore aujourd'hui beaucoup de craintes entre les cas de mésusages récurrents et la difficulté toujours présente à clairement la définir [187] et à définir ce qu'elle permet et ne permet pas de faire.

Pouvoir interagir efficacement avec une solution IA, tout en conservant son agentivité, nécessite de comprendre comment l'outil fonctionne et prend des décisions [86]. Cela devient un réel impératif, face auquel a été mise en avant la notion de transparence [21] [148] [32] [125]. La transparence implique de comprendre le fonctionnement de la solution, mais aussi de pouvoir vérifier qu'elle fonctionne de manière robuste, éthique et en conformité avec nos législations et réglementations. La transparence permet à l'humain de comprendre les décisions prises par la solution IA, renforçant ainsi la confiance en ce type de système.

5.2.1 Évaluation de la confiance

Au-delà d'aborder le concept de transparence et de son potentiel impact sur la confiance que lui accorde l'humain, nous nous intéressons à comment mesurer la confiance des utilisateurs envers une solution IA. L'évaluation de la confiance est une tâche complexe qui peut impliquer une variété de méthodes, allant des méthodes d'enquête traditionnelles (entretiens, questionnaires) [56] à des méthodes plus objectives et basées sur le comportement de l'utilisateur [182] [185]. Après avoir exploré la méta-analyse des méthodes de mesures de la confiance en les solutions IA, telle que par Vereschak et ses collaborateurs (2021) [172], nous nous sommes intéressés aux mesures comportementales. Les auteurs proposent de décomposer ces méthodes en plusieurs aspects que nous détaillons dans le tableau 5.1. Chaque méthode fournit un éclairage différent sur le niveau de confiance des participants envers le système. Dans le cadre de notre étude, nous nous sommes intéressés à la conformité et à l'accord Humain-IA comme mesure de la confiance. L'idée est que si un utilisateur est d'accord avec les prédictions d'un modèle, il est probable qu'il ait confiance dans ce modèle. Ces méthodes ont l'avantage d'être relativement simples à mettre en œuvre et de fournir une mesure quantitative de la confiance.

Pour évaluer l'accord entre les prédictions de l'utilisateur et celles du modèle prédictif, nous utilisons des méthodes utilisées pour faire de l'évaluation subjective de la qualité d'expérience (QoE). Ces méthodes ont été largement utilisées dans les études d'évaluation quantitative de l'UX (terminologie similaire à celle de QoE comme nous avons pu le voir dans le chapitre 2) et ont prouvé leur efficacité notamment dans des études d'évaluation de qualité perçue d'images [8] [9] [100]. Dans notre cas, nous les utilisons pour évaluer la qualité des prédictions de notre modèle à partir du point de vue de l'utilisateur. Cependant, il est important de noter que cette méthode a ses limites. Tout d'abord, elle repose sur l'hypothèse que l'utilisateur est capable d'évaluer correctement la qualité des prédictions du modèle. Si cette hypothèse ne tient pas, c'est-à-dire que l'humain n'est pas capable de juger correctement les prédictions, les mesures de confiance obtenues pourraient être biaisées. De plus, l'accord entre les prédictions de l'utilisateur et celles du modèle ne capture qu'un aspect de la confiance ("accord", d'après la classification de Vereschak et al. [172]) et pourrait ne pas refléter d'autres facteurs importants tels que la compréhension du modèle ou sa satisfaction à son égard.

Malgré ces limites, nous supposons que cette approche peut tout de même fournir des informations précieuses sur la confiance des utilisateurs dans les solutions IA. Et

Méthodes		Description
Temps de décision		*Vitesse d'acceptation d'une recommandation de la part de la solution IA.*
Conformité	Conformité appropriée	*Acceptation des recommandations correctes et rejet des recommandations incorrectes.*
	Sur-conformité	*Acceptation des recommandations incorrectes.*
	Sous-conformité	*Rejet des recommandations correctes.*
Dépendance	Dépendance appropriée	*Demande bénéfique à la solution IA.*
	Surdépendance	*Demande coûteuse à la solution IA.*
	Sous-dépendance	*Absence de demande bénéfique à la solution IA.*
Accord		*Acceptation immédiate de la recommandation de la solution IA.*
Désaccord		*Rejet immédiat de la recommandation de la solution IA.*
Niveaux de questionnement		*Nombre de demande de recommandations supplémentaires.*
Taux de changement		*Nombre de fois où un participant initialement en désaccord a finalement suivi la recommandation de la solution IA.*

TABLE 5.1 – Méthode des mesures de l'interaction Humain-IA selon Vereschak et ses collaborateurs (2021) [172]

en combinant cette approche avec d'autres méthodes d'évaluation de la confiance, nous pouvons obtenir une image plus complète de la manière dont les utilisateurs interagissent avec et font confiance à la solution IA.

5.2.2 Utilité des méthodes utilisées pour faire de l'évaluation subjective de la qualité d'expérience

L'utilisation des méthodes utilisées pour faire de l'évaluation subjective de QoE présente plusieurs avantages pour évaluer la confiance des utilisateurs envers les modèles prédictifs. Premièrement, elles offrent une mesure perceptuelle basée sur l'interaction réelle de l'utilisateur avec le système. Contrairement aux mesures auto-déclarées, où le participant déclare faire confiance par exemple, qui peuvent être sujettes à divers biais, l'évaluation subjective de QoE capte la réaction immédiate de l'utilisateur face aux prédictions du modèle, des données plus facilement objectivables. De plus, elle permet de capturer une évaluation

plus nuancée de la confiance, qui va au-delà des mesures binaires de confiance ou méfiance, en permettant d'identifier des cas précis qui nuisent à la confiance ou qui la renforcent. Les utilisateurs peuvent exprimer leur degré de confiance en évaluant la qualité des prédictions du modèle sur une échelle continue, ce qui permet une analyse plus fine de la confiance.

Enfin, méthodes utilisées pour faire de l'évaluation subjective de QoE offre l'avantage de fournir une mesure implicite de la confiance, qui peut être moins sujette à l'influence de facteurs externes. Par exemple, dans une situation où les utilisateurs sont conscients qu'ils sont en train d'auto-évaluer leur confiance dans un système, ils peuvent modifier inconsciemment ou consciemment leur comportement, ce qui pourrait fausser les résultats. En revanche, en demandant aux utilisateurs d'évaluer la qualité des prédictions du modèle, nous pouvons mesurer leur confiance d'une manière qui est moins susceptible d'être influencée par ce type de biais. Il est important de noter que méthodes utilisées pour faire de l'évaluation subjective de QoE font partie d'un large panel de méthodes disponibles pour évaluer la confiance dans les solutions IA, et que différentes méthodes peuvent être plus appropriées en fonction du contexte spécifique et des objectifs de l'étude.

5.3 Méthodologie

5.3.1 Participants et tâche

Dans le cadre de notre expérimentation, des portraits de personnes sont présentés aux participants. La consigne est d'estimer si la personne du portrait est dans une tranche d'âge définie. Chaque participant doit effectuer 60 prédictions : 20 estimations d'âge sans modèle pour établir la stratégie mise en place initialement (condition contrôle), 20 estimations d'âge avec les prédictions du modèle et 20 estimations d'âge avec les prédictions du modèle et son indice de confiance en ses propres prédictions (voir figure 5.1).

Pour cette étude, 30 participants de tout horizon ont été recrutés, ils ont entre 18 et 60 ans. Il n'y avait aucun prérequis particulier pour réaliser la tâche, si ce n'est que de se présenter avec sa correction visuelle si besoin. Pour limiter tout biais lié aux portraits affichés, les participants ont été répartis en trois groupes d'appartenance, pour qui les ensembles de portraits ont été intervertis entre les modalités de la condition expérimentale (voir tableau 5.2). C'est-à-dire que les portraits affichés avec l'indice de confiance pour un groupe sont affichés sans l'indice de confiance pour un autre groupe et sans recommandation

FIGURE 5.1 – Interface de l'expérimentation pour la 3ème modalité de la variable intrasujet : Recommandation du modèle (RM) (Légende : 1ère modalité : il n'y a que le portrait et la consigne ; 2ème modalité : la prédiction du modèle est affichée ; 3ème modalité : l'indice de confiance du modèle est affiché)

pour le troisième groupe.

La première condition expérimentale concerne la recommandation du modèle (RM) : aucune (condition contrôle) VS. recommandation du modèle VS. recommandation et indice de confiance (faible et élevé). La deuxième condition expérimentale désigne l'écart entre l'âge réel de la personne et la prédiction du modèle (EARPM) : même âge VS. écart faible (7 ans) VS. écart élevé (14 ans). Et la troisième condition est la tranche d'âge proposée à l'utilisateur (TAP) : soit la mauvaise tranche avec un écart de +/- 7 ans à l'âge réel, soit la mauvaise tranche avec un écart de +/- 14 ans à l'âge réel. Ces conditions expérimentales

	Groupe d'appartenance		
Condition RM	**Groupe 1**	**Groupe 2**	**Groupe 3**
Estimation seul.e	Set d'images A	Set d'images B	Set d'images C
Estimation avec les recommandations du modèle	Set d'images B	Set d'images C	Set d'images A
Estimation avec les recommandations du modèle et indice de confiance	Set d'images C	Set d'images A	Set d'images B

TABLE 5.2 – Répartition des sets d'images et des groupes d'appartenance des participants en fonction des conditions expérimentales

Tranche d'âge proposée pour le portrait	**Recommandation IA**	**Indice de Confiance**
Bonne tranche	Âge exact	Fort
		Faible
	Âge proche (+/- 7 ans)	Fort
		Faible
	Âge éloigné (+/- 14 ans)	Fort
		Faible
Mauvaise tranche proche (+/- 7 ans)	Âge exact	Fort
		Faible
	Âge proche (+/- 7 ans)	Fort
		Faible
	Âge éloigné (+/- 14 ans)	Fort
		Faible
Mauvaise tranche éloignée (+/- 7 ans)	Âge exact	Fort
		Faible
	Âge proche (+/- 7 ans)	Fort
		Faible
	Âge éloigné (+/- 14 ans)	Fort
		Faible

TABLE 5.3 – Tableau récapitulatif des conditions expérimentales et de la répartition des portraits photos en fonction de ces conditions

sont présentées dans le tableau 5.3 pour lesquelles nous nous sommes efforcés de répartir les portraits afin d'équilibrer les conditions croisées. Comme expliqué avec le tableau 5.2, tous les participants vont explorer les portraits répartis sur chacune des modalités, mais les mêmes portraits ne bénéficieront pas des mêmes informations du modèle IA selon le groupe d'appartenance.

L'expérimentation s'est déroulée sur le site de la Halle 6 Ouest de Nantes Université. La Halle 6 Ouest représente un vecteur d'innovation avec son *Experience Lab*, consacré à l'évaluation de l'expérience humaine, en mettant à disposition des salles d'expérimentations avec des équipements de pointe [1].

5.3.2 Matériel

Pour réaliser notre expérimentation, nous utilisons la solution de questionnaire SphinxOnline, qui permet d'intégrer une interface dans laquelle les participants peuvent faire

1. Casques EEG, capteurs physiologiques, oculomètres...

leurs prédictions, mais aussi observer les recommandations de la solution IA. Les 60 portraits de personnes à évaluer sont extraits d'un jeu de données d'images de visages annotés avec leur âge réel [146]. La sélection des portraits a été faite par tirage aléatoire, suivi d'un nettoyage des données afin d'exclure les portraits ne convenant pas aux prérequis de notre tâche. Les critères de refus sont associés à :

— une prise de vue ne permettant pas d'identifier la totalité du visage de la personne (trop éloignée, personne de biais), avec plusieurs personnes ou de trop mauvaise qualité ;

— la personne sur le portrait qui ne doit pas être mineure, porter un signe ostentatoire, porter un chapeau, un casque, etc. qui masquent une trop importante partie du visage

— une non-parité de genre pour chaque set d'images ;

— des fichiers ne laissant apparaître aucun personnage.

Un oculomètre est également utilisé pour mesurer objectivement le parcours visuel des participants sur l'écran (voir chapitre 6). Puis nous avons administré deux questionnaires, le premier basé sur le TAM de Davis [43] et le deuxième est le questionnaire *Trust in Automation* (TiA) de Korber [104]. Les deux questionnaires sont détaillés dans le chapitre 7. Et enfin, nous utilisons un guide d'entretien post-expérimentation.

5.3.3 Mesures

Les données récoltées lors de l'expérimentation sont :

— Les réponses des participants. Ces données brutes sont exploitées, mais aussi transformées en proportion d'accord entre l'opérateur humain et le modèle prédictif (voir sous-section 5.4.2) ;
— Les temps de fixation bruts des portraits photo et des recommandations du modèle prédictif (voir chapitre 6) ;
— Les réponses aux questionnaires TAM et TiA (voir chapitre 7)
— Les réponses des participants à l'entretien post-expérimentation.

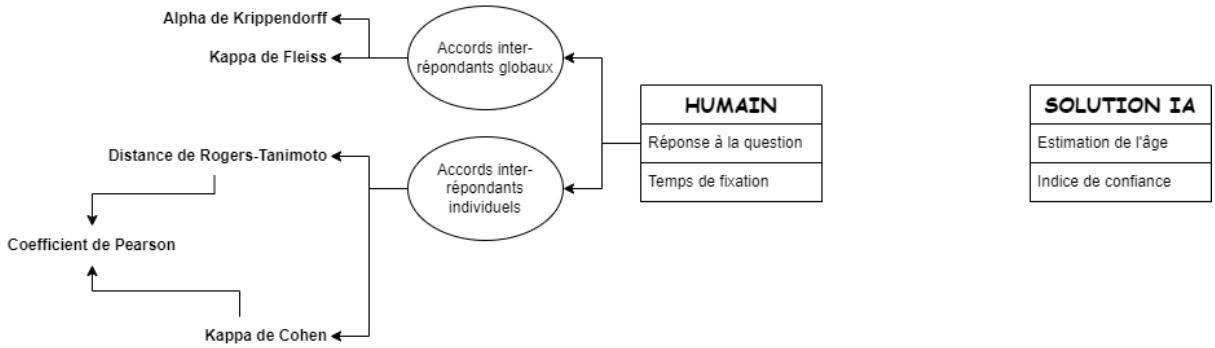


FIGURE 5.2 – Méthodes employées pour calculer les accords inter-répondants par paires de participants et globaux

5.4 Est-ce que les réponses des participants fluctuent en fonction des informations du modèle IA

5.4.1 Évaluation de l'accord inter-répondants

Pour comparer les réponses des participants, nous avons commencé par mesurer l'accord global entre tous les répondants en utilisant deux indices bien établis : l'alpha de Krippendorff [76] et le Kappa de Fleiss [62]. Ces outils statistiques ont été employés pour donner une évaluation précise et complète du degré de consensus parmi l'ensemble des participants à notre étude.

Ensuite, nous avons approfondi notre analyse en introduisant deux métriques distinctes visant à quantifier le niveau d'accord entre chaque paire unique de répondants. L'objectif de cette démarche était de 1) mesurer la subjectivité de la tâche en fonction du désaccord entre participants, 2) évaluer les différences de comportements d'une condition expérimentale à une autre et 3) identifier des comportements irréguliers. Pour cela nous utilisons tout d'abord du coefficient Kappa de Cohen pour chaque combinaison unique d'observateurs. Il s'agit d'une approche largement utilisée pour évaluer l'accord inter-observateurs [84] [173] [180] [8]. Et enfin, nous avons utilisé une métrique de similarité binaire, appelée distance de Rogers-Tanimoto (ou distance de Jaccard pondérée). Cette métrique permet d'évaluer la similarité entre deux vecteurs binaires [89] [8]. Les méthodes analytiques utilisées pour déterminer les accords inter-répondants sont représentées dans la figure 5.2.

Pour éviter de biaiser nos résultats avec les recommandations du modèle et leurs indices de confiance, nous avons dans premier temps calculé les accords inter-répondants par

groupe, mais uniquement lorsque le modèle ne faisait pas de prédictions.

Accords inter-répondants globaux

Alpha de Krippendorff Dans un premier temps, nous avons donc utilisé l'alpha de Krippendorff sur leurs réponses (Oui/Non) pour chacun des groupes. Cet indice est utilisé comme un indicateur de fiabilité inter-juges. Il mesure l'accord entre deux répondants ou plus qui doivent classifier des éléments dans des catégories exclusives [76] [101].

$$\alpha = 1 - \frac{D_o}{D_e} \tag{5.1}$$

, où D_o correspond au désaccord observé et D_e correspond au désaccord supposément attribué au hasard (voir annexe 10.5).

	Groupe d'appartenance		
	Groupe 1	**Groupe 2**	**Groupe 3**
Alpha de Krippendorff	- .023	.002	$<$ .001

TABLE 5.4 – Alpha de Krippendorff par groupe

L'alpha de Krippendorff de chacun des groupes est très proche de 0, ce qui semble indiquer que l'accord entre les répondants de chaque groupe n'est pas meilleur que ce qui serait attendu par hasard (voir tableau 5.4).

Kappa de Fleiss Nous avons ensuite utilisé le kappa de Fleiss qui vise à mesurer l'accord global entre plus de deux observateurs. Le point fort de cet indice est qu'il représente l'accord entre différents observateurs tout en prenant en compte la proportion d'accord potentiellement dû au hasard [62] [57]. Cet indice se rapproche donc de l'alpha de Krippendorff. Pour interpréter cet indice, nous nous sommes appuyé sur les travaux de Landis et Koch (1977) [105]. Selon ces auteurs, l'indice kappa s'interprète de la manière suivante : plus l'indice est proche de 1 et plus les évaluations concordent vers un accord élevé. Á l'inverse, plus l'indice est proche de -1 et plus les évaluations concordent vers un désaccord élevé des évaluateurs. Et enfin plus l'indice est proche de 0 et plus la concordance sera faible, nous aurons tendance à considérer que les accords sont principalement dus au hasard.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{5.2}$$

, où P_o correspond à la proportion observée d'accords entre les observateurs et P_e correspond à la proportion attendue d'accords dus au hasard (accord attendu par le hasard, calculé en moyennant les proportions d'observations pour chaque catégorie individuelle et en prenant en compte le nombre d'observateurs).

	Groupe d'appartenance		
	Groupe 1	**Groupe 2**	**Groupe 3**
Kappa Fleiss	- .028	- .003	- .005

TABLE 5.5 – Kappa Fleiss par groupe

Tout comme pour l'alpha de Krippendorff, le Kappa de Fleiss de chaque groupe est également très proche de 0, indiquant un accord inter-répondants qui n'est pas meilleur que ce qui serait attendu par le hasard.

Accords entre chaque paire de répondants

Kappa de Cohen Étant également un indice kappa, le Kappa Cohen (KC) fonctionne de manière similaire au kappa Fleiss, et permet de mesurer le degré de concordance entre des répondants en tenant compte de la proportion d'accord, potentiellement due au hasard. Cependant, le kappa Cohen ne mesure l'accord qu'entre deux observateurs [105].

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{5.3}$$

, où P_o est la proportion d'accord observée entre les répondants et P_e est la probabilité d'un accord dû au hasard.

Un test de Kruskal-Wallis a ensuite été réalisé pour déterminer si des différences significatives existaient dans les Kappa Cohen entre les différents groupes d'appartenance. Les résultats semblent indiquer qu'il n'y a pas de différence significative dans les Kappa Cohen entre les groupes ($H(2) = 2.74$, $p = .254$). Ces résultats suggèrent que l'appartenance au groupe n'a pas d'effet significatif sur les scores du coefficient kappa de Cohen. Nous avons donc décidé par la suite de traiter les participants uniformément (voir figure 5.4 et tableau 5.6).

Les résultats obtenus par l'analyse de l'accord inter-juges via le coefficient Kappa de

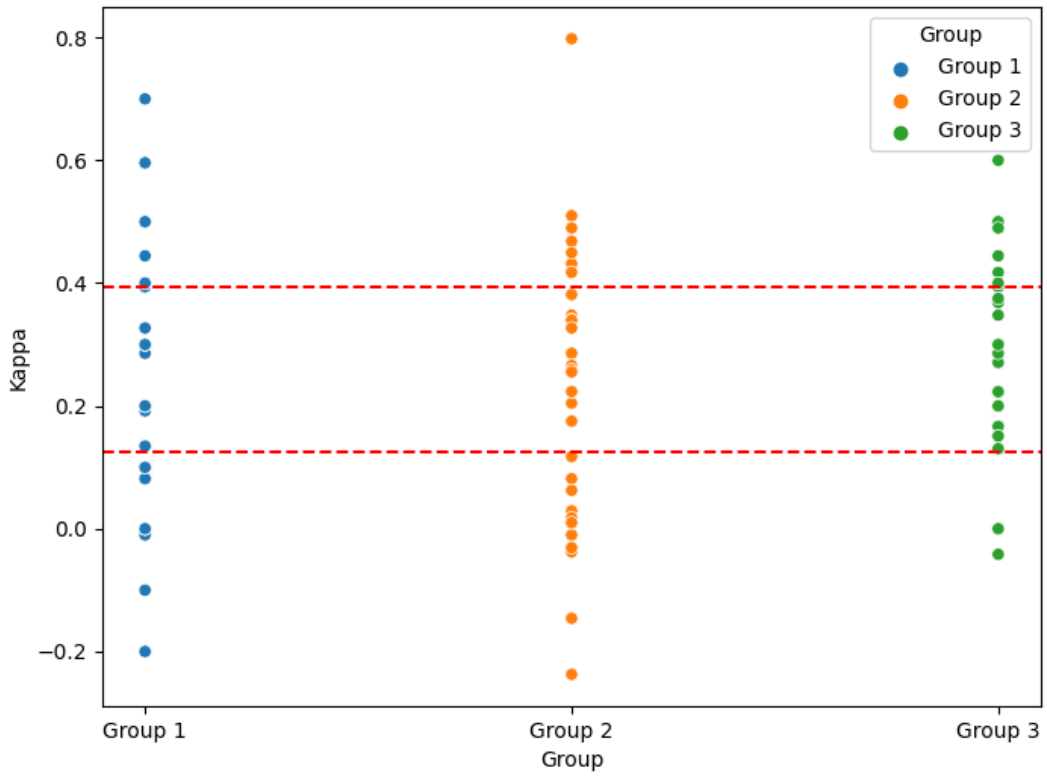


FIGURE 5.3 – Coefficients Kappa de Cohen par groupe d'appartenance (dans ce graphique, chaque point correspond à la concordance des réponses entre une paire de participants du même groupe d'appartenance)

Moyenne	**Écart-type**	**[Min ; Max]**	**Écart interquartile**
.250	.3191	[- .234 ; .79]	[.12 ; .39]

TABLE 5.6 – Description des coefficients Kappa de Cohen sans distinction de groupe d'appartenance (les traits horizontaux sur le graphique représentent l'écart interquartile qui s'étend de 0.12 à 0.39)

Cohen indiquent une variabilité considérable dans la capacité des participants à estimer correctement l'âge des personnes sur les portraits. Avec un coefficient Kappa de Cohen moyen de 0.25, nous pouvons considérer que l'accord entre les participants est relativement faible. Bien qu'éloignée de zéro, cette valeur indique que les réponses des participants ne sont que légèrement mieux que ce qui serait attendu par hasard, soulignant une certaine

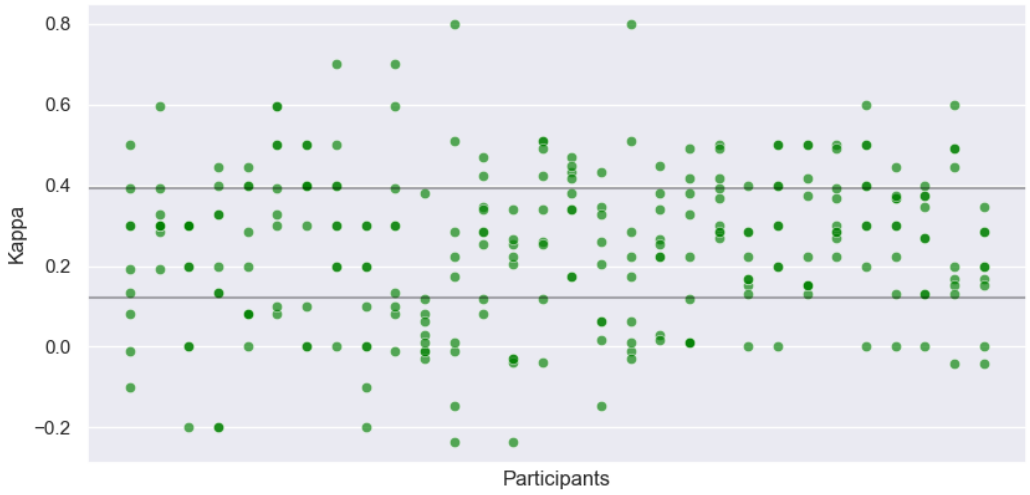


FIGURE 5.4 – Coefficients Kappa de Cohen sans distinction de groupe d'appartenance (dans ce graphique, chaque point correspond à la concordance des réponses entre une paire de participants sur l'ensemble de l'échantillon)

subjectivité dans la tâche d'estimation de l'âge [105].

La valeur minimale de -0.23 montre que le binôme de participants le moins en accord de tout notre échantillon reste sur un désaccord relativement faible. A contrario, le binôme le plus en accord avec un coefficient Kappa Cohen de 0.79 nous montre que certains juges sont capable d'un accord fort [105]. Ce qui nous conforte dans le fait que certains participants font des estimations d'âge similaire.

Pourtant l'écart interquartile s'étend de 0.12 à 0.39, ce qui confirme notre hypothèse d'une variabilité dans l'accord. Cet écart indique que la majorité des valeurs kappa se situent dans un domaine où l'accord est considéré de faible à modéré. Ce même écart est finalement relativement étroit autour de valeurs faibles de kappa, ce qui reflète une certaine consistance dans le niveau de subjectivité perçue par les participants lors de l'évaluation de l'âge.

A l'issu de ces résultats, nous estimons que notre tâche d'estimation d'âge à partir de portraits porte une part de subjectivité significative, rendant l'accord entre différents évaluateurs difficile à atteindre, et ce de manière consistante. Cette tâche, tel qu'elle a été

structurée, semble trop dépendante des perceptions personnelles et potentiellement des préjugés culturels ou individuels pour permettre un consensus fiable. Nous allons, dans la suite de cette sous-section, essayer de confirmer les résultats obtenus. En identifiant le degré d'accord inter-juge, et par extension le degré de subjectivité de notre tâche, nous souhaitons identifier si les recommandations du modèle IA auront un impact sur le jugement des participants et donc sur l'accord de jugement.

Dissimiliraté de Rogers-Tanimoto La distance de Rogers-Tanimoto (RT) est un indice inter-répondants qui mesure la similarité ou la différence (dissimilarité) entre les préférences par paire des répondants qui sont représentés en deux ensembles binaires [145]. Par exemple, cet indice peut servir à dire que pour un portrait la réponse "Oui, cette personne est dans la tranche d'âge" est préférée à la réponse "Non, cette personne n'est pas dans la tranche d'âge". Cet indice est souvent privilégié pour sa robustesse face à des échantillons de taille variable. Nous avons choisi d'utiliser la dissimilarité de Rogers-Tanimoto, qui se concentre sur la proportion d'éléments non correspondants entre les deux échantillons binaires. Un score proche de 0 indique que les deux échantillons sont identiques (les participants sont d'accord entre eux), tandis qu'un score proche de 1 indique qu'ils n'ont aucun élément en commun (les participants ne sont pas d'accord entre eux) (voir figure 5.5 et tableau 5.8).

$$D_{RT}(x, y) = \frac{2(s_{01} + s_{10})}{(s_{00} + s_{11} + 2(s_{01} + s_{10}))} \tag{5.4}$$

Cette formule tient compte de deux séquences x et y (respectivement deux participants dans notre contexte), où :

- s_{00} est le nombre de fois où les deux participants ont répondu "oui", alors on attribue la valeur de 0 ;
- s_{11} est le nombre de fois où les deux participants ont répondu "non", alors on attribue la valeur de 1 ;
- s_{01} est le nombre de fois où le premier participant a répondu "oui" et le deuxième participant a répondu "non" , alors on attribue la valeur de 1 ;
- s_{10} est le nombre de fois où le premier participant a répondu "non" et le deuxième participant a répondu "oui" , alors on attribue la valeur de 0.

Pour une meilleure lisibilité, nous avons représenté les composantes de l'équation dans le tableau 5.7.

Les scores de dissimilarité de Rogers-Tanimoto (D_{RT}) apportent des informations pré-

		A (participant 1)	
		0 ("oui")	**1 ("non")**
B (participant 2)	**0 ("oui")**	s00	s10
	1 ("non")	s01	s11

TABLE 5.7 – Composantes de l'équation de la dissimilarité de Rogers-Tanimoto

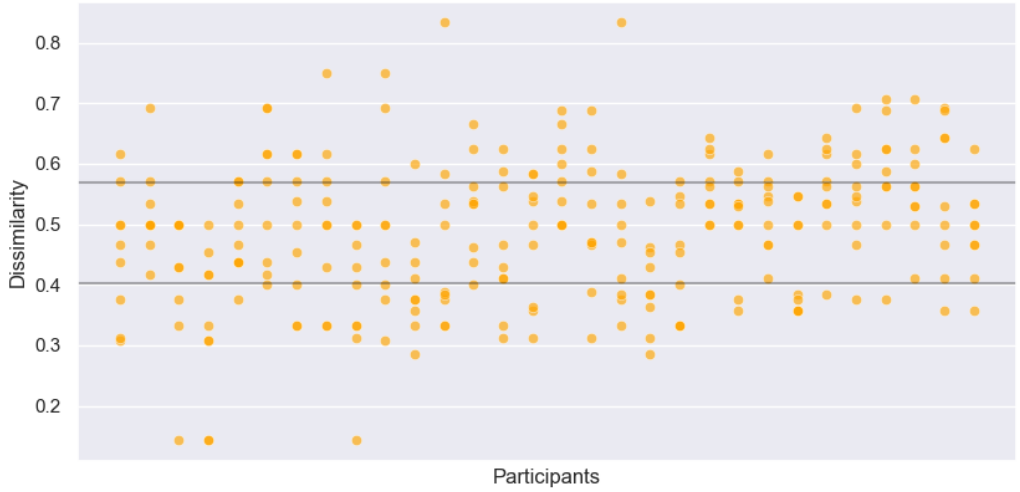


FIGURE 5.5 – Valeurs de dissimilarité de Rogers-Tanimoto sans différenciation de groupe d'appartenance (les traits horizontaux sur le graphique représentent l'écart interquartile qui s'étend de 0.4 à 0.56)

Moyenne	**Ecart-type**	**[Min ; Max]**	**Ecart interquartile**
.489	.191	[.143 ; .83]	[.4 ; .56]

TABLE 5.8 – Scores de dissimilarité de Rogers-Tanimoto

cieuses sur la variabilité des des réponses des participants. Le score moyen de dissimilarité de 0.489 indique que, en moyenne, l'estimation d'âge à partir de portraits photos présente une dissimilarité modérée. Cette valeur moyenne, située au milieu de l'échelle d'interprétation (entre 0 et 1), suggère que les réponses des participants ne sont ni extrêmement

similaires ni extrêmement différentes les unes des autres. L'écart interquartile, s'étendant de 0.4 à 0.56, indique que 50% des scores se concentrent dans cette gamme, signifiant une dissimilarité modérée pour la majorité des paires évaluées. Cet écart interquartile relativement étroit autour de 0.5 montre une certaine cohérence dans le degré de dissimilarité perçu.

Ces résultats suggèrent que les réponses des participants ne sont ni extrêmement similaires ni extrêmement différentes. Pour autant, au-delà de ces résultats globaux et cet écart interquartile restreint, nous pouvons observer une certaine variabilité des résultats avec un écart-type que nous considérons assez élevé (presque 0.2) et une valeur minimale de 0.143 qui revoie à une grande similarité entre les deux évaluateurs concernés et une valeur maximale de 0.83 qui renvoie à une grande dissimilarité entre les deux évaluateurs concernés. Cette inconsistance peut refléter une variation des jugements qui pourrait être influencée par des facteurs contextuels ou subjectifs spécifiques aux participants ou aux portraits évalués. Pour la suite de nos analyses, nous allons estimer si les méthodes utilisées précédemment (Kappa de Cohen et dissimilarité de Rogers-Tanimoto) expriment les mêmes tendances en termes d'accord inter-répondants.

Corrélation entre le Kappa Cohen et la dissimilarité RT L'étude de la corrélation entre les méthodes suggérées a été réalisée à travers le coefficient de corrélation de Pearson. Le graphique 5.6 montre la répartition des deux mesures de consensus entre les répondants.

Dans leur étude, Ak et ses collaborateurs (2021) ont utilisé la même méthode et ont découvert une corrélation négative forte entre le coefficient de Kappa Cohen et la dissimilarité de Rogers-Tanimoto dans l'une de leurs expérimentations [8]. Cela signifie que, dans des tâches où la subjectivité est faible (c'est-à-dire où il y a moins d'espace pour l'interprétation personnelle), les deux mesures d'accord ont tendance à se déplacer dans des directions opposées : lorsque l'une augmente, l'autre diminue, et vice versa.

Dans notre cas, le coefficient de Pearson est de -0.84 ($p < .001$), ce qui montre également une corrélation négative forte entre les Kappa Cohen et les scores de similarité de Rogers-Tanimoto. Cela suggère que lorsque l'accord entre les participants (mesuré par le kappa de Cohen) est plus élevé, la similarité entre leurs réponses (mesurée par la similarité de Rogers-Tanimoto) est généralement plus faible, et inversement. Cela peut indiquer que les tâches dans notre étude avaient également un faible taux de subjectivité.

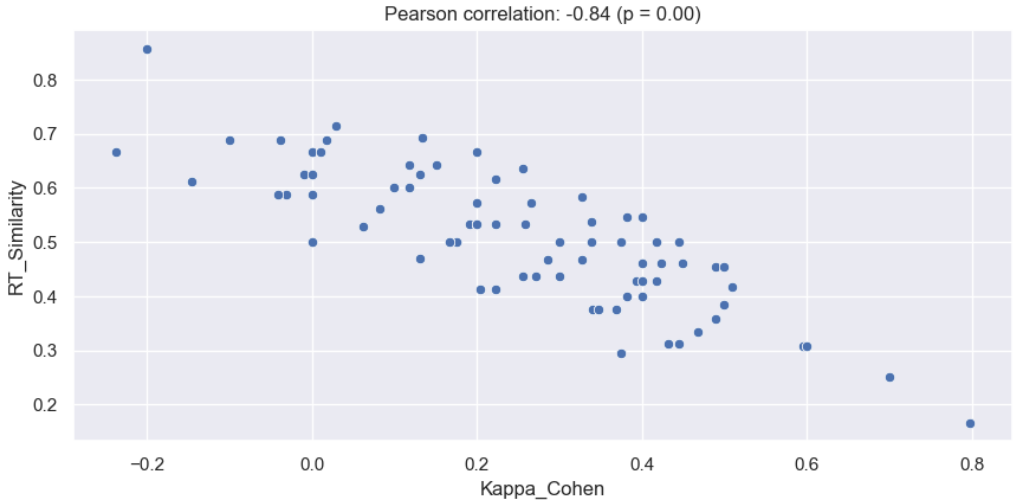


FIGURE 5.6 – Comparaison des valeurs de Kappa de Cohen avec les valeurs de dissimilarité RT.

5.4.2 Tests d'indépendance des accords Humain-IA par rapport aux informations du modèle IA

Dans cette sous-section, nous nous intéressons aux relations de dépendance entre les accords "participants et modèle" (appelé accord Humain-IA) et les différentes variables étudiées. Les accords Humain-IA sont considérés ici comme une source de confiance de l'opérateur humain envers le modèle comme développé dans la sous-section 5.2.1. Dans le cadre de notre expérimentation, les réponses des participants (oui/non) sont de nature nominale alors que les réponses de l'IA (âge) sont de nature ordinale. Dans cette sous-section nous intéressons donc particulièrement à l'accord Humain comme source de confiance de l'opérateur humain envers le modèle comme développé dans la sous-section 5.2.1. Nous avons donc adapté notre considération de l'accord Humain-IA qui doit répondre à un des deux critères suivants 1) les réponses Oui des participants lorsque l'âge proposé par le modèle est dans la tranche d'âge et 2) les réponses Non des participants lorsque l'âge proposé par le modèle est hors de la tranche d'âge (voir figure 5.8 et tableau 5.9). Voici les hypothèses explorées dans cette partie :

— H1 : les accords Humain-IA dépendent des informations fournies par le modèle (avec les recommandations du modèle VS. avec les recommandations du modèle et l'indice de confiance) ;

— H2 : les accords Humain-IA dépendent de la distance entre la prédiction du modèle

et l'âge réel du portrait (âge réel VS. +/- 7 ans VS. +/- 14 ans).

		Âge prédit par le modèle IA	
		Dans la tranche d'âge proposée	**Hors de la tranche d'âge proposée**
Réponse du participant à la question	**Dans la tranche d'âge proposée ("oui")**	Accord Humain-IA	Désaccord Humain-IA
	Hors de la tranche d'âge ("non")	Désaccord Humain-IA	Accord Humain-IA

TABLE 5.9 – Représentation des accords et des désaccords Humain-IA dans notre contexte expérimental (les accords Humain-IA sont les réponses "Oui" des participants lorsque l'âge proposé par le modèle est dans la tranche d'âge, et les réponses "Non" des participants lorsque l'âge proposé par le modèle est hors la tranche d'âge)

Les méthodes analytiques utilisées pour comparer les accords Humain-IA en fonction des informations mises à disposition par le modèle IA sont représentées dans la figure 5.7. Et les données sont décrites dans le tableau 5.10.

Nous avons effectué différents tests d'indépendance représentés dans le tableau 5.11.

Au vu des résultats obtenus et à un seuil de significativité statistique de 5%, il semble que l'accord Humain-IA est dépendant des informations mises à disposition par le modèle. Pour le même seuil de significativité, l'accord Humain-IA semble dépendant de la distance entre l'âge du portrait et l'âge prédit par le modèle.

Nous avons, ici, multiplié les tests d'indépendance car 1) seul le test Chi² peut être appliqué à des matrices 3x2, 2) le test de Fisher est le test le plus robuste selon la littérature, mais 3) le test de Barnard est parfois plus puissant que le test de Fisher pour des échantillons de petite taille.

Le Kappa de Cohen comme indicateur d'accord Humain-IA Le Kappa de Cohen servant à mesurer l'accord entre deux observateurs, nous avons choisi de le réemployer ici. Seulement, au lieu de s'en servir pour mesurer l'accord Humain-Humain, nous nous en sommes servis pour mesurer l'accord Humain-IA (voir tableau 5.12 et figure 5.10). De la même manière que pour les précédentes analyses, nous avons représenté la méthode employée dans la figure 5.9.

	Nombre d'accords Humain-IA	Nombre de désaccords Humain-IA
Toutes conditions confondues (avec le modèle)	671	529
Avec les recommandations du modèle (condition 2)	303	297
Avec les recommandations du modèle et les indices de confiance (condition 3)	368	232
Avec les recommandations du modèle et les indices de confiance faible uniquement	190	130
Avec les recommandations du modèle et les indices de confiance forte uniquement	178	102

TABLE 5.10 – Description des accords et désaccords Humain-IA en fonction des informations mises à disposition par le modèle

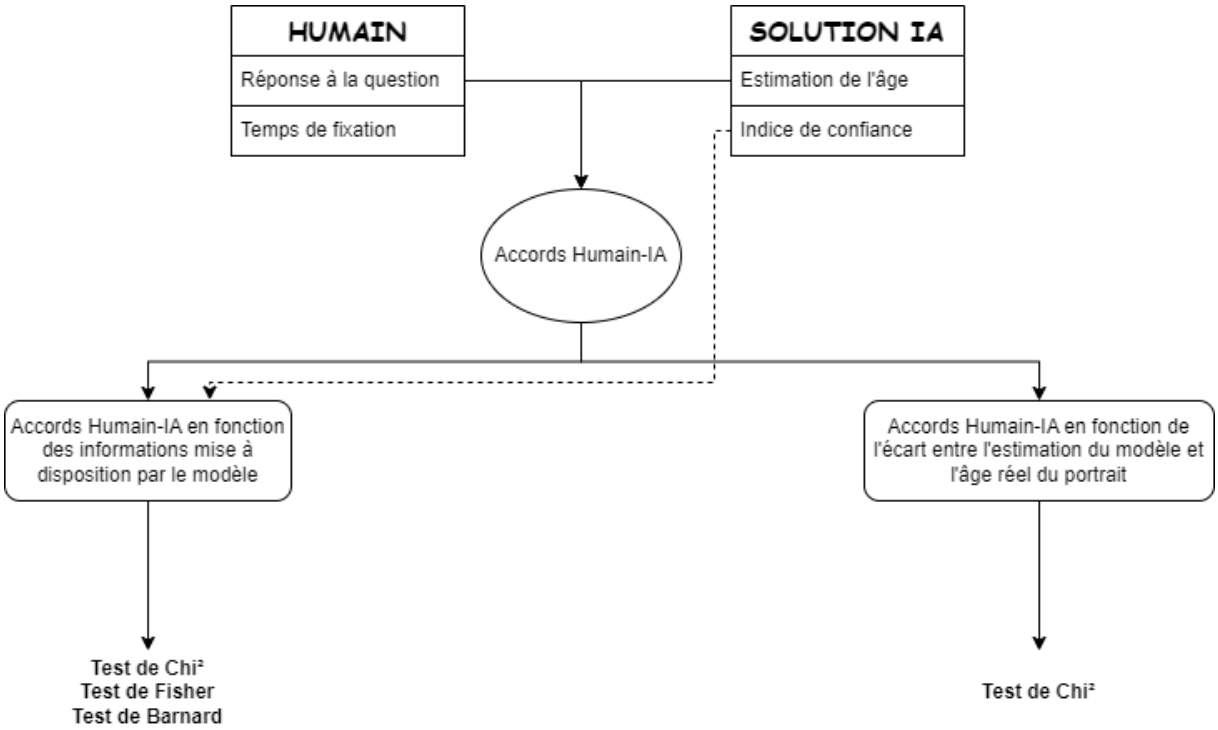


FIGURE 5.7 – Méthodes employées pour comparer les accords Humain-IA en fonction des informations mises à disposition par le modèle IA

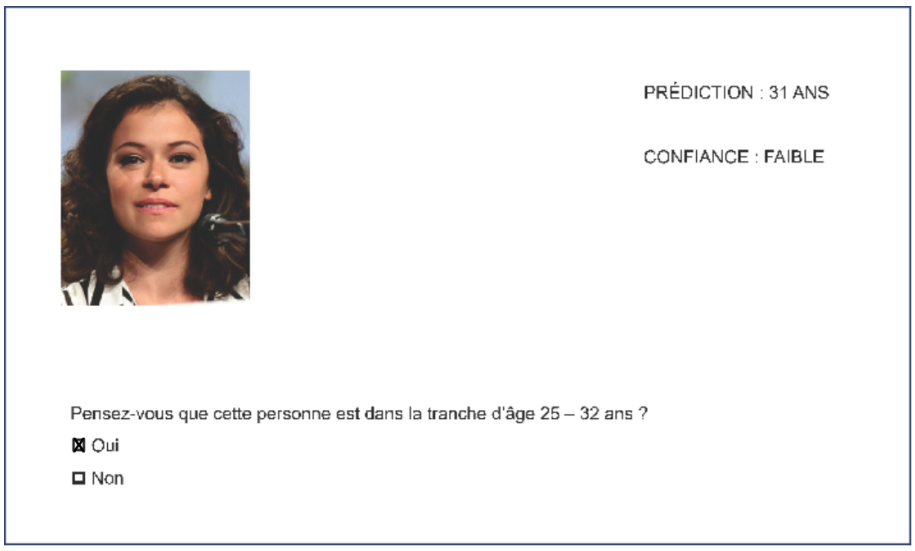


FIGURE 5.8 – Exemple d'accord Humain-IA pour un portrait (ici, nous avons un accord Humain-IA car le participant a répondu "oui" et que la proposition du modèle IA de 31 ans est dans la tranche d'âge)

Hypothèse	Test de Chi² : valeur statistique (*valeur p*)	Test de Fisher : valeur statistique (*valeur p*)	Test de Barnard : valeur statistique (*valeur p*)
H1	13.848 ($< .001$)	.643 ($<.001$)	- 3.779 ($< .001$)
H2	23.467 ($< .0001$)	-	-

TABLE 5.11 – Résultats des tests d'indépendance entre les accords Humain-IA et les différentes variables étudiées (H1 - informations fournies par le modèle ; H2 - distance entre la prédiction du modèle et l'âge réel du portrait)

	Moyenne	**Écart-type**	**[Min ; Max]**	**Écart interquartile**
Accord H-IA avec recommandations du modèle	-.01	.25	[-.596 ; .490]	[-.182 ; .156]
Accord H-IA avec recommandations du modèle et indice de confiance	.208	.21	[-.319 ; .604]	[.07 ; .365]

TABLE 5.12 – Tableau descriptif des accords Humain-IA (Kappa de Cohen) en fonction des informations fournies par le modèle

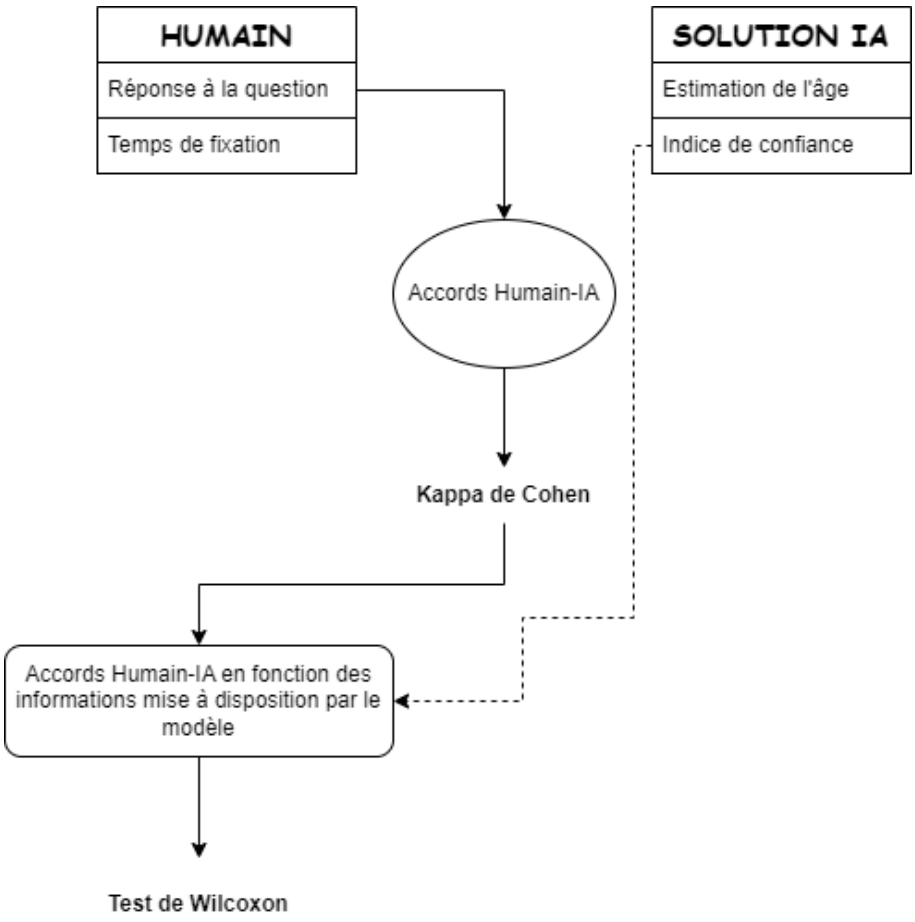


FIGURE 5.9 – Méthodes employées pour comparer les accords Humain-IA (scores Kappa de Cohen) en fonction des informations mises à disposition par le modèle IA

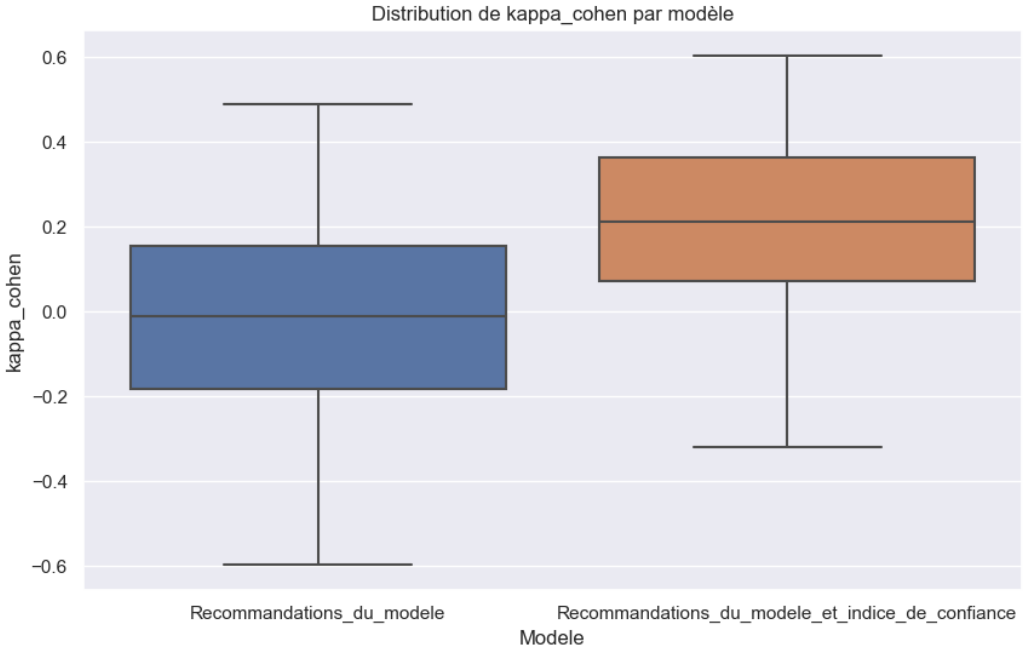


FIGURE 5.10 – Distribution des scores Kappa de Cohen pour les accords H-IA en fonction des conditions expérimentales

Nous avons ensuite utilisé le test de rang signé de Wilcoxon sur nos nouveaux indicateurs d'accord Humain-IA (Kappa de Cohen). Cette méthode non paramétrique sert à comparer deux ensembles de scores. Plus précisément, le test compare les différences de médianes entre deux échantillons appariés. Nous avons ainsi pu retester notre hypothèse H1 (voir tableau 5.13). De manière similaire à notre première analyse, il semble que les informations mises à disposition par le modèle (recommandations VS. recommandations et indice de confiance) ont un effet sur l'accord Humain-IA (mesuré par Kappa Cohen) au seuil de significativité statistique de 5% ($W = 86$, $p < .01$).

Hypothèse	**Test des rangs signés de Wilcoxon : valeur statistique** (*valeur p*)
H1	86 (*.005*)

TABLE 5.13 – Résultats des tests des rangs signés de Wilcoxon sur les accords Humain-IA (Kappa de Cohen) en fonction des informations mises à disposition par le modèle

	Prédiction du modèle	
Réponse du participant	Dans la tranche d'âge	Hors de la tranche d'âge
Dans la tranche d'âge	VRAI (Vrai positif)	FAUX (Faux négatif)
Hors de la tranche d'âge	FAUX (Faux positif)	VRAI (Vrai négatif)

TABLE 5.14 – Tableau de contingence 2x2 pour synthétiser une expérimentation à instances positives et instances négatives en fonction de l'accord Humain-IA

5.5 Est-ce que des critères objectifs permettent de prédire la confiance accordée au modèle IA ?

Dans cette partie, nous abordons la mesure et la comparaison de la performance des participants en utilisant différents facteurs objectifs (FO). Ce que nous appelons ici facteur objectif correspond à un moyen quantitatif d'évaluer l'efficacité et la pertinence des réponses des participants par rapport aux informations qu'ils ont à leur disposition. Cette approche vise à faciliter l'analyse systématique de l'accord Humain-IA basée sur des critères mesurables et non influencés par des jugements subjectifs. La distribution d'accords et désaccords des participants avec le modèle est séparée en utilisant le critère de la distance euclidienne entre leurs réponses (subjectives) et des scores objectifs comme la validité de la réponse de la solution IA ou encore si la tranche d'âge proposée englobe ou non l'âge réel de la personne sur le portrait. Les stimuli sont ensuite classés en deux catégories : ceux qui correspondent à la prédiction de la solution et ceux qui ne correspondent pas (voir tableau 5.14).

Ces FO nous permettent de construire des courbes ROC (*Receiver Operating Characteristic*) pour évaluer la précision de nos prédictions en fonction de la manière dont nous objectivons notre situation de test. Les analyses ROC servent initialement à évaluer la performance de modèles de classification binaire ("accord Humain-IA VS. désaccord Humain-IA dans notre cas)en mesurant leur capacité à distinguer entre deux classes. Elle permet de comparer différents modèles en fonction d'un prédicteur (nos FO) et de quantifier la performance globale à travers l'AUC (aire sous la courbe) [100]. Un AUC élevé indique une meilleure performance du modèle.

L'objectif principal est d'identifier à quel point nous pouvons être précis dans notre

capacité à prédire le nombre d'accords entre les participants et le modèle selon les différents FO. Les FO sont définis de la manière suivante :

— FO 1 - la plus faible distance entre l'âge réel et les bornes de la tranche d'âge proposée en valeur absolue, si l'âge réel est dans la tranche d'âge, alors le FO est négatif :

$$FO1 = \min(|a - b_1|, |a - b_2|); négatif si a \in [b_1; b_2] \tag{5.5}$$

— FO 2 - la distance entre l'âge réel et la valeur médiane de la tranche d'âge proposée :

$$FO2 = a - m([b_1; b_2]) \tag{5.6}$$

— FO 3 - la plus faible distance entre l'âge prédit par le modèle et les bornes de la tranche d'âge proposée en valeur absolue , si l'âge prédit par le modèle est dans la tranche d'âge proposée, alors le prédicteur est négatif :

$$FO3 = \min(|c - b_1|, |c - b_2|); négatif si c \in [b_1; b_2] \tag{5.7}$$

— FO 4 - la distance entre l'âge prédit et la valeur médiane de la tranche d'âge :

$$FO4 = c - m([b_1; b_2]) \tag{5.8}$$

— FO 5 - la distance entre l'âge réel et l'âge prédit par le modèle :

$$FO5 = a - c \tag{5.9}$$

, où a correspond à l'âge réel du portrait, b_1 et b_2 correspondent respectivement aux bornes inférieure et supérieure de la tranche d'âge, m correspond à la médiane de la tranche d'âge proposée et c correspond à l'âge prédit par le modèle.

Toutes les analyses réalisées dans cette section sont représentées dans la figure 5.11.

L'interprétation visuelle de l'AUC se fait principalement à travers la courbe ROC.

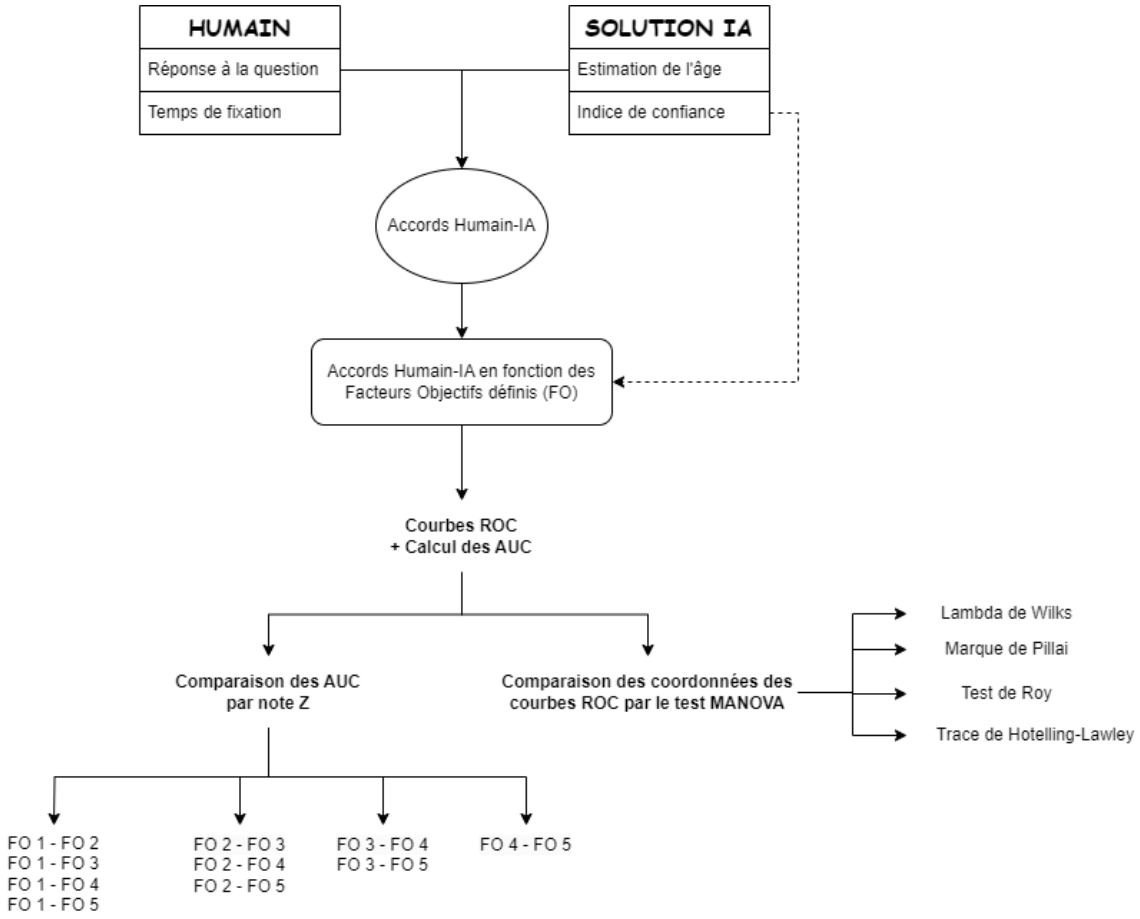


FIGURE 5.11 – Méthodes employées pour comparer les capacités à prédire les accords Humain-IA en fonction des facteurs objectifs appliqués

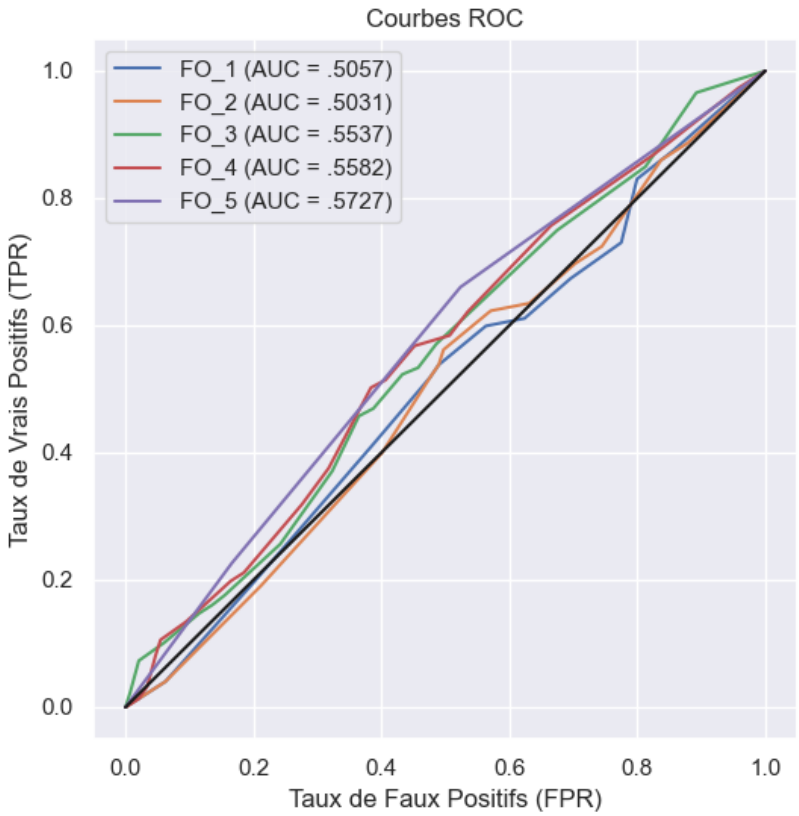


FIGURE 5.12 – Courbes ROC en fonction de chacun des FO

Facteur Objectif (FO)	**Aire sous la courbe (AUC)**
FO1	.5057
FO2	.5031
FO3	.5537
FO4	.5582
FO5	.5727

TABLE 5.15 – Aires sous la courbe (AUC) en fonction des facteurs objectifs

Une courbe qui se rapproche du coin supérieur gauche du graphique désigne une bonne capacité prédictive. Cela signifie que le modèle a une sensibilité élevée et une spécificité élevée aux différents seuils calculés. Une courbe qui se rapproche de la diagonale (du coin inférieur gauche au coin supérieur droit) indique une performance qui n'est pas mieux que le hasard, comme nous pouvons le constater sur les différentes courbes de notre

graphique. Et enfin, une courbe qui est en dessous de la diagonale et qui se rapproche du coin inférieur droit signifie que le modèle fait pire que le hasard, ce qui peut signifier que l'on prédit à l'envers. Dans notre cas de figure, toutes les AUC sont assez proches de la diagonale (tracé en noir sur la figure 5.12) ($min_{AUC} = .5031$ et $max_{AUC} = .5727$). Ce qui signifie qu'avec nos données et les facteurs objectifs que nous avons sélectionnés, nous ne prédisons pas forcément mieux que le hasard. Cela peut notamment s'expliquer par une des hypothèses que nous avons faites au début de la sous-section 5.4.1, à savoir que l'estimation d'âge est une tâche trop subjective pour correctement y estimer l'accord Humain-IA.

Il ne semble pas non plus y avoir de grandes différences entre les courbes (voir figure 5.12 et tableau 5.15). Pour confirmer ce constat, nous allons mobiliser deux types de méthodes de comparaison de courbes ROC : la méthode de comparaison d'AUC par Note z de Hanley et McNeil [71] et le test d'analyse multivariée de variance (MANOVA) des coordonnées des courbes ROC.

5.5.1 Comparaison des facteurs objectifs par note Z

Pour comparer les différents facteurs objectifs, nous avons d'abord utilisé le test de comparaison d'aire sous la courbe (AUC) des courbes ROC, en nous basant sur le score z [99]. Proposée par Hanley et McNeil en 1983 [71] en alternative à la méthode de Delong [49], cette méthode permet déterminer s'il y a une différence entre des AUC d'une même distribution en fonction des manières de l'objectiver (FOs) et ce, de manière robuste et fiable en évaluant la performance des modèles prédictifs. Dans le cadre de notre étude, nous employons cette méthode sur un seul prédicteur (les réponses des participants par rapport aux réponses du modèle) selon différentes manières d'objectiver notre situation de test (les FO). En comparant les AUC des courbes ROC, nous pouvons donc identifier les FO qui offrent la meilleure performance prédictive et ainsi orienter nos efforts pour améliorer la qualité de nos prédictions d'accord entre l'humain et le modèle, et leur applicabilité dans des contextes réels. Voici la formule de la note Z :

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{SE_{AUC_1}^2 + SE_{AUC_2}^2 - 2rSE_{AUC_1}SE_{AUC_2}}} \tag{5.10}$$

La formule mathématique de la note Z et la méthode pour déterminer si cette différence est statistiquement significative sont détaillées dans l'annexe 10.5.

Au vu des résultats du test, il semblerait qu'il n'y pas de différence significative entre les

	FO1	FO2	FO3	FO4	FO5
FO1	-				
FO2	.41(*.68*)	-			
FO3	- 7.95(*< .001*)	-8.41 (*< .001*)	-		
FO4	-8.81 (*< .001*)	-9.27 (*< .001*)	-0.80 (*.42*)	-	
FO5	-7.09 (*.006*)	-7.38 (*.001*)	-2.06 (*.039*)	-1.58 (*.11*)	-

TABLE 5.16 – Notes Z calculées en comparant les AUC de notre distribution des accords Humain-IA en fonction de nos facteurs objectifs avec la méthode de McNeil et Hanley [71] (Interprétation : si la note Z est positive et élevée, cela suggère que l'AUC_1 est significativement plus grande que l'AUC_2. Á l'inverse, si la note Z est négative et basse, cela suggère que l'AUC_1 est significativement inférieure que l'AUC_2. Et si la note Z est proche de 0, alors cela suggère qu'il n'y a pas de différence entre les AUC.

AUC des courbes obtenues en fonction de l'écart à la tranche ou à la médiane, cela concerne les binômes "FO 1 - FO 2" et "FO 3 - FO 4", qui sont statistiquement similaires ($p > .05$). De plus, nous pouvons observer qu'il n'y a pas de différence significative entre le FO 4 et le FO 5 ($p > .05$). Donc prédire l'accord humain-IA en fonction de - la distance entre l'âge prédit par le modèle et la valeur médiane de la tranche d'âge proposée (FO4) - ou de - la distance entre l'âge réel et l'âge prédit par le modèle (FO5) - est statistiquement similaire au vu de nos résultats (voir tableau 5.16). Les autres comparaisons sont statistiquement significatives au seuil critique de 5%.

Pour savoir quel FO est le meilleur prédicteur, nous pouvons directement nous référer à la valeur de l'AUC, sans avoir à réaliser de tests post-hoc puisque l'AUC représente la performance de prédiction de notre FO et que nous savons lesquels sont statistiquement différents. Ainsi, les facteurs d'objectivation - écart entre l'âge réel et l'âge prédit par le modèle - (FO 5, AUC = .5727) et - distance entre l'âge prédit et la valeur médiane de la tranche d'âge - (FO 4, AUC = .5582) sont les meilleurs moyens d'étudier l'accord Humain-IA au vu de leur niveau de performance. Ceci peut s'expliquer par le fait que plus le modèle prédictif donne un résultat éloigné de l'âge réel de la personne, plus cela semblera incohérent pour le participant qui aura tendance à ne pas le croire. De la même manière que pour FO 4, si la tranche d'âge proposée semble proche de l'âge de la personne, mais que la prédiction semble éloignée, alors le participant aura tendance à ne pas suivre l'avis du modèle.

Variable	**Statistique W** (*valeur p*)
TPR	.969 (*.799*)
FPR	.972 (*.132*)

Table 5.17 – Test de Shapiro-Wilk pour la normalité de la distribution

Variable	**Statistique F** (*valeur p*)
TPR	.50 (*.736*)
FPR	.28 (*.892*)

Table 5.18 – Test de Levene pour l'homogénéité des variances

5.5.2 Comparaison des facteurs objectifs à partir du test de MANOVA

Pour confronter les résultats de nos comparaisons d'AUC obtenus avec la note Z, nous avons testé une autre méthode : comparer les taux de vrais positifs (TPR) et de faux positifs (FPR) de notre distribution en fonction des FO. Pour cela, nous utilisons le test d'Analyse Multivariée de la Variance (MANOVA). Le test de MANOVA est un test statistique paramétrique particulièrement utile pour étudier la relation entre une ou plusieurs variables indépendantes (VI) et un ensemble de variables dépendantes. Dans notre cas, la variable indépendante est le type de FO et les variables dépendantes sont les TPR et FPR.

Comme le test de MANOVA est un test paramétrique, nous devons d'abord tester la normalité de la distribution des scores TPR et FPR pour chaque FO et l'homogénéité des variances. La normalité de la distribution est calculée à l'aide du test de Shapiro-Wilk, pour vérifier si les données suivent une distribution normale, et l'homogénéité des variances est évaluée par le test de Levene, pour déterminer si les variances des scores TPR et FPR sont égales entre les groupes de prédicteurs.

Au seuil de significativité statistique de 5%, les tests de Shapiro-Wilk pour les résidus des variables TPR et FPR montrent que les résidus suivent une distribution normale. Au même seuil critique, les tests de Levene indiquent l'absence de différence significative entre les variances. Cela signifie que les variances des groupes sont statistiquement équivalentes pour les deux variables dépendantes. Cela valide l'hypothèse de normalité nécessaire pour

	Statistique	Valeur	F	ddl1	ddl2	Valeur p
Prédicteur	Marque de Pillai	0.317	3.01	8	128	0.004
	Lambda de Wilks	0.684	3.30	8	126	0.002
	Trace de Hotelling-Lawley	0.462	3.58	8	124	0.001
	Test de Roy	0.461	7.37	4	64	0.001

TABLE 5.19 – Test de MANOVA

l'application de tests paramétriques tels que le MANOVA.

Le test de MANOVA utilise plusieurs statistiques pour tester l'hypothèse nulle que les vecteurs moyens des groupes sont égaux dans un espace multivarié. Chaque statistique offre certains avantages et limites, nous allons donc nous appuyer sur les résultats des quatre statistiques pour voir si elles tendent toutes vers le même résultat.

Comme nous pouvons le voir dans le tableau 5.19, au risque d'erreur de 5%, le type de FO a un effet significatif sur les variables dépendantes (TPR et FPR) selon nos différentes statistiques du test de MANOVA [2]. De manière globale, les résultats du test de MANOVA indiquent donc que les variations dans les valeurs TPR et FPR de notre distribution peuvent être en partie expliquées par les différents FO. Ces constatations vont dans le sens de ce que nous avons déjà pu explorer avec notre test de comparaison d'AUC.

5.6 Discussion

Les analyses réalisées à l'aide de la méthode de Hanley et McNeil [71] et du test de MANOVA offrent des perspectives complémentaires sur l'impact des facteurs objectifs (FO) dans l'évaluation de l'accord humain-IA. La méthode de Hanley et McNeil, centrée sur la comparaison des AUC des courbes ROC pour différents FO, n'a pas révélé de différences significatives entre certains binômes de FO, indiquant une performance similaire dans la prédiction de l'accord humain-IA pour ces paires. Cela suggère que, pour certaines façons d'objectiver les données, il n'existe pas de supériorité marquée d'un FO sur un autre en termes de capacité à prédire cet accord. Cependant, en examinant l'effet des types de FO

2. Lambda de Wilks : $F = 3.30$, $p = .002$; Marque de Pillai : $F = 3.01$, $p = .004$; Trace de Hotelling-Lawley : $F = 3.58$, $p < .001$; test de Roy : $F = 7.37$; $p < .001$

sur les variables dépendantes TPR et FPR, le test de MANOVA a montré que le choix du FO a un impact significatif sur ces mesures. Cette différence dans les résultats peut s'expliquer par le fait que le MANOVA prend en compte l'interaction entre les variables dépendantes et offre une vue d'ensemble de l'effet des FO, tandis que la méthode de Hanley et McNeil se concentre sur la comparaison individuelle des performances des FO à travers les AUC. Selon nous, ces résultats soulignent l'importance de considérer à la fois la performance prédictive des modèles (comme mesurée par l'AUC) et les spécificités des taux de réponse (TPR et FPR) lors de l'évaluation de l'accord humain-IA. Ils suggèrent également que, bien qu'il puisse ne pas y avoir de différence significative dans la capacité de certains FO à prédire l'accord humain-IA de manière isolée, la manière dont ces FO interagissent avec les métriques de performance peut varier de manière significative.

Bien que ces méthodes offrent des informations importantes sur l'impact des facteurs objectifs dans l'évaluation de l'accord humain-IA, plusieurs limites doivent tout de même être prises en compte pour interpréter nos résultats. Tout d'abord, nous nous sommes concentrés sur un ensemble spécifique de FO pour évaluer l'accord humain-IA. Mais il est possible que d'autres FO pertinents n'aient pas été inclus, ce qui limite la généralisation de nos conclusions. Inclure un éventail plus large de FO pourrait révéler des aspects supplémentaires et pertinents de l'accord humain-IA, que nous n'avons pas capturés par notre étude. De plus, les différentes AUC calculées sont relativement faibles (toutes très proches de 0.5), ce qui signifie qu'avec les FO que nous avons sélectionnés et les données recueillies nous n'arrivons pas à prédire bien mieux que le hasard. Comme pour toute étude scientifique, les résultats pourraient être influencés par la taille et la diversité de notre échantillon de participants. Ici, nous ne sélections que 30 participants sans spécificités, mais un échantillon plus grand et avec une plus grande diversité pourrait fournir des résultats plus généralisables et révéler davantage de nuances dans les accords Humain-IA. Les performances en fonction de FO peuvent aussi varier en fonction de la complexité des modèles d'IA utilisés. Dans notre étude, nous avons volontairement induit des portraits pour lesquels le modèle avait une performance moyenne (taux d'erreur d'environ 50%) pour identifier si les participants pourraient identifier si le modèle se trompait et si dans ces cas de figure, ils suivaient quand même ses recommandations. Il serait tout de même intéressant de valoriser ces méthodes dans un contexte où le modèle propose un taux de performance bien plus élevé pour identifier si dans le contexte d'un outil efficace et efficient, les participants trouvent ses réponses suffisamment acceptables pour s'y fier.

Et enfin, le contexte d'application de ces méthodes est l'estimation d'âge des portraits, mais nous savons maintenant que l'accord humain-IA peut varier considérablement en fonction des besoins et contraintes de l'utilisateur, du contexte ou encore de la tâche définie. Nos résultats, obtenus dans un contexte spécifique, pourraient ne pas être directement transférables à d'autres domaines d'application sans une évaluation supplémentaire. C'est pour cela qu'il serait intéressant de réitérer l'utilisation de cette méthodologie dans un contexte professionnel pour en renforcer la validité. Ces résultats permettent tout de même de mieux appréhender l'impact des différentes conditions expérimentales sur la performance des participants et la confiance qu'ils accordent au modèle prédictif. Ces méthodes de comparaison offrent de la visibilité sur la manière de choisir un facteur en fonction de ses données en évaluant la précision des prédictions.

Chapitre 6

Au-delà des chiffres : La valeur mésestimée des résultats non significatifs

Dans ce chapitre

Faisant suite au chapitre précédent, ce chapitre souligne l'importance des résultats non significatifs. Nous y présentons des méthodes de mesures et d'analyse ayant donné lieu à des résultats statistiques non significatifs mais utiles pour comprendre comment s'intéresser à la confiance des opérateurs humains envers des solutions IA.

6.1 Introduction

Dans ce chapitre, nous continuons d'examiner les données présentées précédemment dans le chapitre 5. Nous abordons ici les analyses statistiques qui n'ont pas produit de résultats significatifs. Les méthodes d'analyse que nous avons employées offrent diverses possibilités pour interpréter nos données. Il est essentiel de noter que l'absence de significativité statistique peut être due à plusieurs facteurs. Premièrement, la taille de l'échantillon peut être insuffisante. Deuxièmement, comme nous l'avons observé dans le chapitre 5 avec les accords inter-répondants, la tâche est très subjective, ce qui peut entraîner une variabilité des données collectées, rendant difficile la détection de tendances statistiques. Il se peut également que les données exploitées ne soient tout simplement pas adaptées aux hypothèses que nous souhaitons tester.

Dans un premier temps, nous nous attarderons sur les tests d'indépendance qui n'ont

abouti à aucune différence significative en fonction des conditions expérimentales étudiées. Nous y examinerons comment nous avons fait le lien entre les conditions expérimentales et les accords Humain-IA. Dans un second temps, nous regarderons si le temps passé par les participants à fixer les portraits et les recommandations du modèle peuvent être des métriques exploitables pour comprendre l'accord Humain-IA en fonction des conditions expérimentales. Nous analyserons les temps de fixation avec des tests d'indépendance puis nous examinerons les biais et les inconsistances dans ces mêmes données.

6.2 Influence de l'indice de confiance du modèle IA sur les accords Humain-IA

Dans cette section, nous nous intéressons aux relations entre les accords Humain-IA et les informations mises à disposition par le modèle. Nous posons ici l'hypothèse selon laquelle il existe un lien entre les accords Humain-IA et le degré de confiance attribué par le modèle à ses propres prédictions, classé en : confiance faible versus confiance forte. La description de ces données est observable dans le tableau 5.10. Pour tester notre hypothèse, nous avons recours au test du Chi-carré, au test de Fisher et au test de Barnard comme illustré dans la figure 6.1. Chacun de ces tests, utilisés pour l'étude de variables catégorielles, offre un éclairage particulier sur les données récoltées.

Test de Chi² : valeur statistique (*valeur p*)	Test de Fisher : valeur statistique (*valeur p*)	Test de Barnard : valeur statistique (*valeur p*)
.939 (*.333*)	.838 (*.314*)	- 1.053 (*.3*)

Table 6.1 – Résultats des tests d'indépendance des accords Humain-IA par rapport aux indices de confiance du modèle IA

Les résultats obtenus (voir tableau 6.1) semblent réfuter notre hypothèse, montrant une absence de lien statistique entre les accords Humain-IA et le niveau de confiance du modèle. En effet, dans notre étude, les accords Humain-IA semblent indépendants de la valeur de l'indice de confiance du modèle au risque d'erreur de 5%.

Bien que les analyses du chapitre 5 nous ont montré un effet du degré d'informations mises à disposition par le modèle sur les accords Humain-IA, les résultats que nous venons d'obtenir nous montrent que la valeur de l'indice de confiance en elle-même n'aurait pas d'influence sur l'accord Humain-IA. Malgré les résultats du chapitre 5, il semble

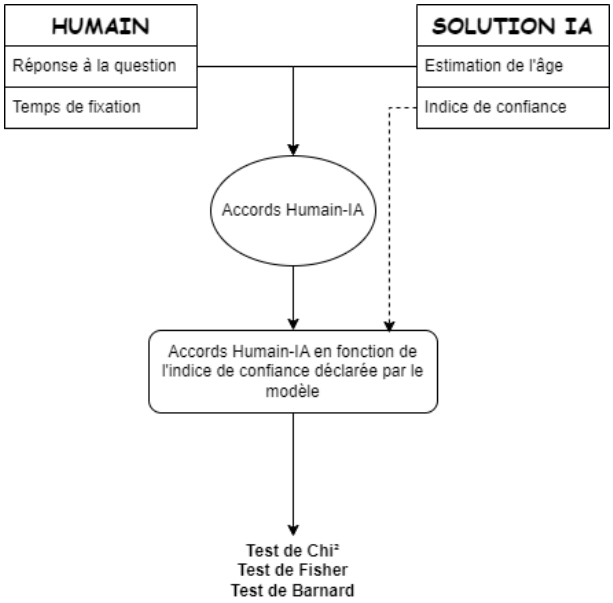


FIGURE 6.1 – Méthodes employées pour comparer les accords Humain-IA en fonction de l'indice de confiance déclarée par le modèle IA

qu'importe que la confiance soit faible ou forte, l'impact sur les accords Humain-IA tient de la présence d'informations est à disposition de la part du modèle. Cette conclusion nous amène à réévaluer la relation entre la perception humaine des recommandations d'une solution IA et l'utilité de la confiance déclarée par le modèle. Ce qui souligne la complexité des interactions Humain-IA et la nécessité de poursuivre les investigations pour mieux comprendre les facteurs influençant ces accords.

6.2.1 Le Kappa de Cohen comme indicateur d'accord Humain-IA

Comme dans le chapitre 5, nous avons ensuite voulu confirmer nos premiers résultats en transformant les accords Humain-IA en valeurs de Kappa de Cohen (voir tableau 6.2 et figure 6.3). Nous utilisons ensuite le test de rangs signés de Wilcoxon sur les valeurs de Kappa de Cohen pour retester notre hypothèse selon laquelle il y a une différence significative d'accords Humain-IA en fonction de la valeur de l'indice de confiance déclarée par le modèle (voir figure 6.2).

Les données du tableau 6.2 et de la figure 6.3 nous montrent des moyennes relativement différentes avec des écarts-type assez importants (.249 et .299). Cette différence reste à être confirmée ou non avec le test de Wilcoxon.

Le test de Wilcoxon nous montre finalement des résultats similaires aux premiers

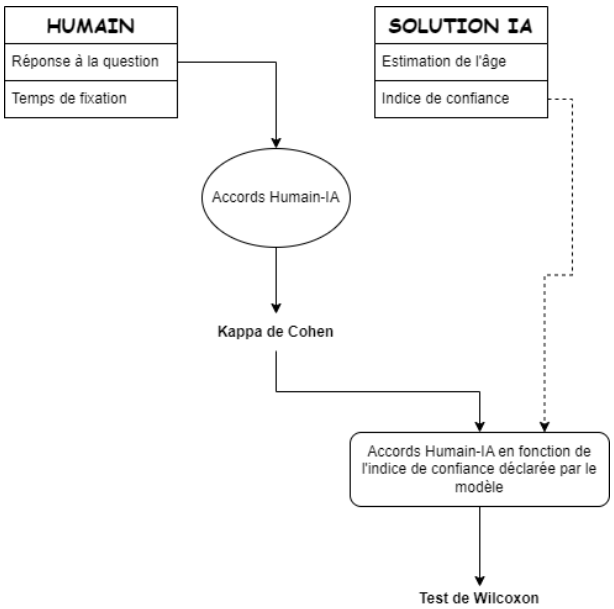


FIGURE 6.2 – Méthodes employées pour comparer les accords Humain-IA (scores Kappa de Cohen) en fonction de l'indice de confiance déclarée par le modèle IA

	Moyenne	**Écart-type**	**[Min ; Max]**	**Écart inter-quartile**
Accords Humain-IA pour l'indice de confiance faible	.156	.249	[-.296 ; .8]	[0 ; .339]
Accords Humain-IA pour l'indice de confiance forte	.266	.299	[-.667 ; .750]	[.019 ; .475]

TABLE 6.2 – Tableau descriptif des accords Humain-IA en fonction de l'indice de confiance déclaré du modèle

Test des rangs signés de Wilcoxon : valeur statistique (*valeur p*)
140 (*.096*)

TABLE 6.3 – Résultats du test des rangs signés de Wilcoxon sur les accords Humain-IA (scores Kappa de Cohen) en fonction de l'indice de confiance déclaré du modèle

tests réalisés (voir tableau 6.3). De manière similaire à notre première analyse, l'accord Humain-IA (mesuré par Kappa de Cohen) ne semble toujours pas influencé par la valeur de l'indice de confiance (faible VS. forte) au risque d'erreur de 5% ($W = 140$, $p > .05$).

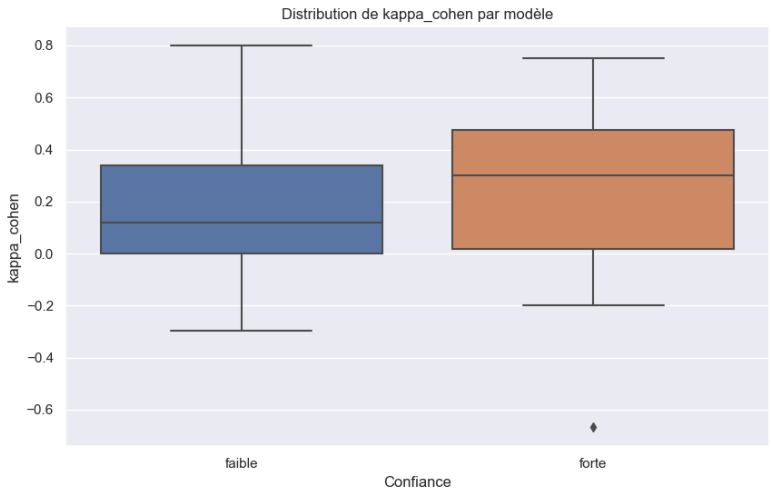


FIGURE 6.3 – Accords Humain-IA (scores Kappa de Cohen) en fonction des modalités de l'indice de confiance du modèle

6.3 Est-ce que le temps de fixation des portraits et des recommandations du modèle sont de bonnes métriques de confiance en la solution IA ?

En complément des accords Humain-IA, nous avons également mesuré les temps de fixation des portraits et des recommandations du modèle à l'aide d'un oculomètre. Nous avons posé l'hypothèse principale d'un changement de comportement (temps passé à fixer les informations à disposition) en fonction des conditions expérimentales. Nous avons testé notre hypothèse à l'aide des tests de Kruskal-Wallis, de Wilcoxon et de Mann-Withney. Nous avons utilisé ces tests pour étudier les variations de temps de fixation en fonction de deux conditions expérimentales : les recommandations du modèle (RM) et la confiance du modèle (CM) (voir figure 6.4).

Nous avons découpé nos hypothèses en 4 sous-hypothèses qui sont les suivantes :

— H1 : Il existe une différence significative dans le temps de fixation des portraits en fonction de RM.

— H2 : Il existe une différence significative dans le temps de fixation des portraits en fonction de CM.

— H3 : Il existe une différence significative dans le temps de fixation de la recommandation du modèle en fonction de RM.

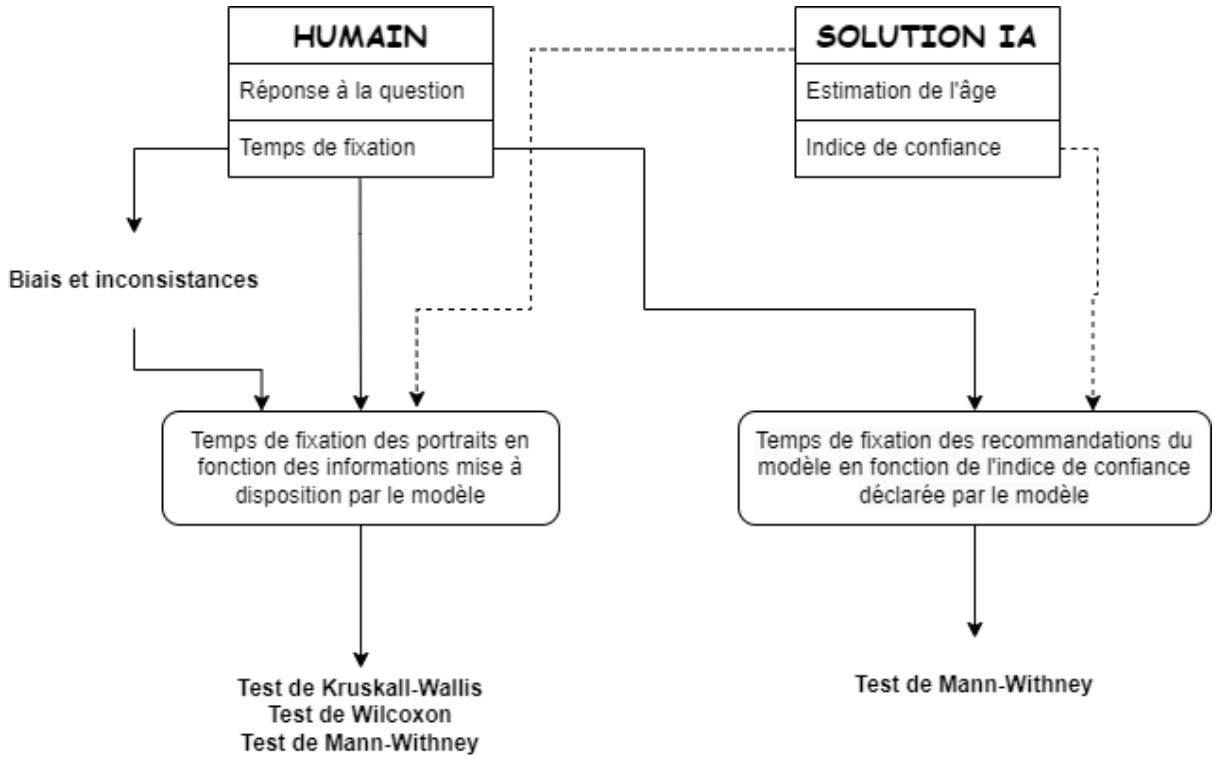


FIGURE 6.4 – Méthodes employées pour comparer les temps de fixation en fonction des informations mises à disposition par le modèle IA

— H4 : Il existe une différence significative dans le temps de fixation de la recommandation du modèle en fonction de CM.

Les données recueillies sont décrites dans le tableau 6.4 pour les temps de fixation des portraits et le tableau 6.5 pour les temps de fixation des recommandations du modèle.

Au vu des résultats obtenus (voir tableau 6.6), nous n'avons observé aucune différence significative des temps de fixation du portrait en fonction des conditions (RM) au risque d'erreur de 5%, nous avons donc rejeté les hypothèses selon lesquelles les informations mises à disposition par le modèle auraient une influence sur le temps total passé à fixer les portraits ou les recommandations du modèle. Au même risque d'erreur, nous n'avons pas non plus observé de différence significative des temps de fixation des portraits ni des recommandations du modèle en fonction de la valeur de l'indice de confiance. Ainsi, que le modèle n'affiche pas d'indice de confiance ou qu'il ait une confiance forte ou faible, il n'y aurait aucune différence significative du temps total passé à fixer le portrait ou les recommandations du modèle. Pour confirmer nos résultats, nous avons choisi d'explorer les notions de biais et d'inconsistance dans les données de chaque participant que nous

Temps de fixation brut des portraits (en ms)	**Moyenne (*Écart-type*)**	**[Min ; Max]**	**Écart Inter-quartile**
Pour toutes les conditions confondues	3869.85 (*3411.59*)	[0.00 ; 40308.00]	[;]3248.50
Sans les recommandations du modèle (condition 1)	3817.55 (*3631.91*)	[0.00 ; 40308.00]	[1692.00 ; 4940.50]
Avec les recommandations du modèle (condition 2)	4014.79 (*3602.63*)	[235.00 ; 36105.00]	[1573.00 ; 4659.50]
Avec les recommandations du modèle et les indices de confiance (condition 3)	3777.22 (*2958.64*)	[130.00 ; 27173.00]	[1811.00 ; 5019.00]
Avec les recommandations du modèle et les indices de confiance faible uniquement	3559.39 (*2629.29*)	[242.00 ; 16401.00]	[1815.50 ; 4790.25]
Avec les recommandations du modèle et les indices de confiance forte uniquement	4026.17 (*3282.15*)	[130.00 ; 27173.00]	[1809.50 ; 5073.50]

TABLE 6.4 – Statistiques descriptives des temps de fixation bruts des portraits (en ms)

allons présenter dans la sous-section 6.3.1.

6.3.1 Évaluation de biais et inconsistance des temps de fixation des portraits en fonction de RM

Dans cette sous-section, nous avons recours à l'outil Sureal (Subjective Recovery Analysis), développé par Netflix [68] [111], pour détecter les préjugés et les incohérences dans les durées de fixation des stimuli chez les participants selon les recommandations du modèle (RM). Sureal permet de mobiliser un estimateur du maximum de vraisemblance (MLE, pour *Maximum Likelihood Estimator*) qui est une méthode statistique conçue pour déterminer les paramètres sous lesquels les données observées deviennent les plus probables [112]. Cette approche vise principalement à estimer les valeurs de paramètres qui augmentent la chance que notre situation de test explique fidèlement les données recueillies. En faisant cela, le MLE aide à évaluer et à mesurer deux sortes d'erreurs spécifiques : les biais et les incohérences.

Dans notre contexte, le biais fait référence à la tendance systématique à sous-estimer ou surestimer une mesure donnée. Ainsi un participant peut avoir tendance à passer plus de

Temps de fixation	**Moyenne (Écart-type)**	**[Min ; Max]**	**[Q1 ; Q3]**
Pour toutes les conditions (incluant le modèle)	472.21 (496.65)	[0.00 ; 5333.00]	[165.00 ; 498.00]
Avec les recommandations du modèle (condition 2)	474.96 (528.77)	[0.00 ; 5333.00]	[165.00 ; 498.00]
Avec les recommandations du modèle et les indices de confiance (condition 3)	469.46 (462.72)	[0.00 ; 3292.00]	[166.00 ; 499.00]
Avec les recommandations du modèle et les indices de confiance faible uniquement	459.35 (434.46)	[0.00 ; 3275.00]	[250.00 ; 600.25]
Avec les recommandations du modèle et les indices de confiance forte uniquement	481.02 (493.58)	[0.00 ; 3292.00]	[253.00 ; 551.75]

TABLE 6.5 – Statistiques descriptives des temps de fixation bruts des recommandations du modèle (en ms)

Hypothèse	**Test de Kruskall Wallis : valeur statistique** (*valeur p*)	**Test de Wilcoxon : valeur statistique** (*valeur p*)	**Test de Mann-Withney : valeur statistique** (*valeur p*)
H1	3.219 (*.2*)	-	-
H2	-	-	48050.0 (*.125*)
H3	-	75886.5 (*.795*)	-
H4	-	-	44746.0 (*.980*)

TABLE 6.6 – Résultats des tests statistiques utilisés pour tester les hypothèses d'effet des conditions expérimentales sur les temps de fixation des portraits

temps à fixer les portraits lorsqu'il est confronté à certaines recommandations du modèle, indépendamment des autres facteurs. L'inconsistance, elle, renvoie à la variabilité des réponses d'un participant face aux mêmes conditions. Par exemple, si un participant a des temps de fixation très différents pour le même portrait présenté dans le même contexte à plusieurs reprises, cela serait considéré comme une inconsistance.

Après avoir obtenu les valeurs de biais et d'inconsistance de chaque participant, nous avons utilisé le test de Kruskall-Wallis pour retester notre sous-hypothèse H1. Cette sous-hypothèse désigne l'estimation selon laquelle il existe une différence significative dans

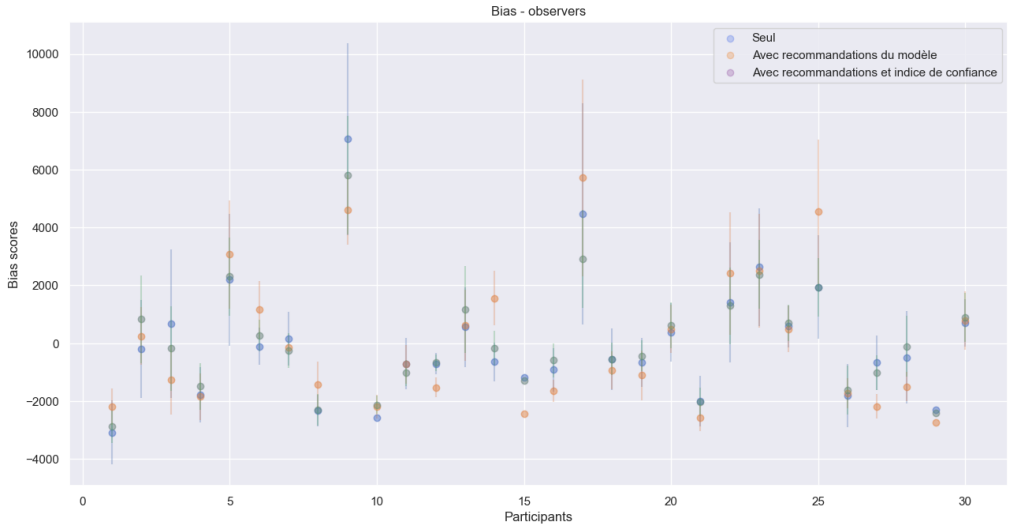


FIGURE 6.5 – Valeurs de biais des participants en fonction des conditions expérimentales

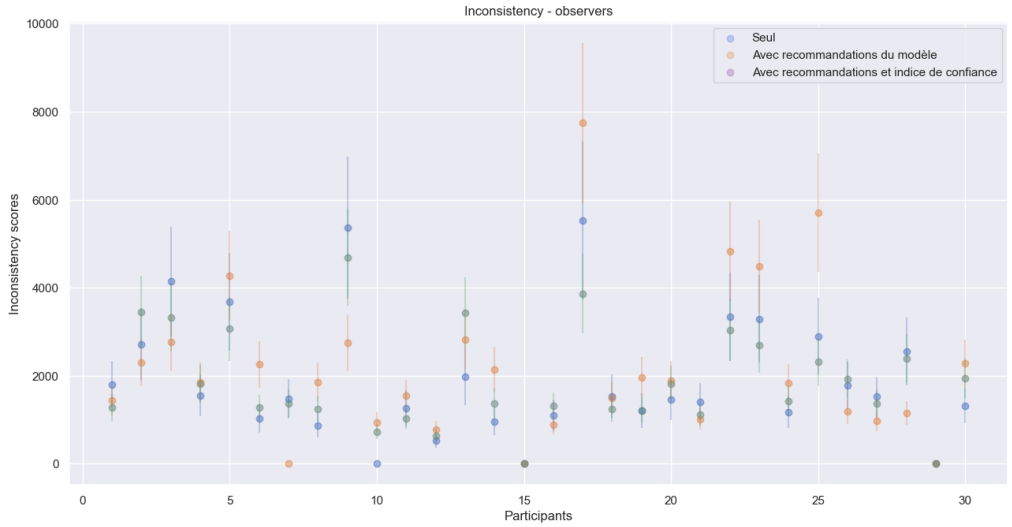


FIGURE 6.6 – Valeurs d'inconsistance des participants en fonction des conditions expérimentales

Variable dépendante	**Test de Kruskall Wallis : valeur statistique** (*valeur p*)
Biais mesurée chez chaque participant	.390 (*.823*)
Inconsistance mesurée chez chaque participant	.245 (*.885*)

TABLE 6.7 – Résultats des tests de Kruskall-Wallis sur les valeurs de biais et d'inconsistance des participants en fonction des conditions expérimentales

le temps de fixation des portraits en fonction des conditions expérimentales.

Au risque d'erreur de 5%, nous n'observons pas de différence significative de biais (voir figure 6.5) ni d'inconsistance (voir figure 6.6) dans les temps de fixation des portraits en fonction des conditions expérimentales (voir tableau 6.7). Il ne semble donc pas y avoir d'impact des informations du modèle sur le temps passé par les participants à fixer les portraits affichés.

Limite de l'utilisation du MLE dans notre cadre expérimental La méthode MLE est décrite dans la littérature comme suffisamment robuste pour estimer les paramètres d'un modèle. Son apport dans notre cadre expérimental est indéniable, pour autant il nous faut prendre en compte un certain nombre de limites. Tout d'abord, le MLE est une méthode d'estimation ponctuelle qui ne fournit pas d'informations qualitatives sur les valeurs de biais et d'inconsistance estimées, ce qui peut être problématique dans notre contexte, où nous nous intéressons à des cas spécifiques pour essayer d'en tirer des conclusions (par exemple, que se passe-t-il quand le modèle fait une prédiction juste avec une tranche d'âge éloignée). Ensuite, le MLE peut être sensible aux valeurs aberrantes, ce qui signifie que quelques observations atypiques peuvent avoir un fort impact sur les estimations de paramètres. Dans le contexte de notre étude, cela pourrait signifier que quelques temps de fixation particulièrement longs ou courts pourraient biaiser nos estimations de la tendance centrale. Il est donc important de prendre ces limites en compte lors de l'interprétation de nos résultats. Cependant, cette méthode nous permet tout de même de mobiliser une approche moins traditionnelle de prise en compte de ce type de données en complément ou remplacement des tests d'indépendance qui ont été employés précédemment.

6.4 Conclusion

Dans ce chapitre, nous avons exploré les résultats non significatifs de notre étude sur la confiance accordée à un modèle prédictif en fonction des informations qu'il met à disposition : prédiction et indice de confiance en cette même prédiction. Les résultats, bien que non significatifs, nous montrent des possibilités méthodologiques en termes d'analyse de données avec une approche parfois moins traditionnelle, par exemple en explorant les biais et inconsistance dans les temps de fixation. Ces approches nous offrent donc une nouvelle perspective sur la manière d'explorer le concept de confiance en un système prédictif. La prise en compte de ces approches alternatives encourage une approche plus diversifiée et objective tout en restant centrée sur l'utilisateur dans l'évaluation des solutions IA.

Chapitre 7

Méthodes d'enquête pour une approche plus holistique

Dans ce chapitre

Pour finaliser notre expérimentation sur la confiance accordée à une solution IA sur une tache d'estimation d'âge, nous présentons ici les résultats de notre travail d'enquête (questionnaires et entretiens). Les méthodes de sondage employées ici sont complémentaires à celles des méthodes précédemment présentées. Elles permettent ainsi d'avoir une vision plus holistique de la confiance et de l'acceptabilité des solutions IA dans notre contexte expérimental.

7.1 Introduction

Dans ce chapitre, nous avons choisi de mobiliser des méthodes explicites pour mesurer la perception des participants à l'égard de la solution IA qui fait des recommandations d'âge, en affichant un indice de confiance en ses prédictions, dans le cadre de l'expérimentation présentée au chapitre 5. Après avoir réalisé la tâche donnée, nous avons administré deux questionnaires puis effectué un entretien semi-directif avec chaque participant. Les questionnaires sont basés sur deux outils théoriques : le TiA (pour *Trust in Automation*) de Korber et al. [104] et le TAM (pour *Technology Acceptance Model*) de Davis [43], présenté précédemment (voir chapitre 2. Ces deux questionnaires servent à mesurer respectivement la confiance accordée au modèle prédictif et l'intention d'usage de cet outil par nos participants. Les items des deux questionnaires sont évalués sur une échelle de Likert (1 - Pas du tout à 5 - Totalement). Dans le tableau 7.1, nous présentons les différentes dimensions explorées par les deux questionnaires et les résultats associés.

Questionnaire	**Dimensions explorées**	**Score obtenu (sur 5)**
TAM	Utilité perçue	3.083
	Facilité d'utilisation perçue	3.489
	Intention d'usage	2.25
TiA	Fiabilité/Compétence	2.52
	Compréhension/prévisibilité	3.22
	Familiarité	1.8
	Intention des développeurs	3.7
	Propension à la confiance	2.56
	Confiance dans l'automatisation	2.73

TABLE 7.1 – Résultats aux questionnaires TiA et TAM

L'importance de ces deux aspects ne peut être sous-estimée dans le cadre de notre étude. Comme présenté précédemment, nous considérons la confiance en l'IA, ici évaluée grâce au questionnaire TiA, comme un indicateur déterminant dans l'adoption des technologies. Elle traduit le degré de confort des utilisateurs vis-à-vis des technologies automatisées et leur propension à se reposer sur elles. Parallèlement, le questionnaire TAM nous a fourni des indications précieuses sur l'intention d'usage des technologies. Il s'agit là d'un paramètre fondamental pour prédire le taux d'adoption de ces technologies. En effet, même si une technologie est techniquement robuste et répond à un besoin, si les utilisateurs potentiels ne sont pas prêts ou disposés à l'utiliser, son adoption sera limitée.

Au cours de l'analyse, nous examinerons de près les scores de ces deux questionnaires. Cela nous permettra de mieux comprendre les attitudes et les intentions de nos participants envers le modèle utilisé, et par conséquent, de fournir des recommandations utiles pour l'implémentation et l'adoption de technologies à l'avenir.

7.2 Analyse des questionnaires

Sur l'évaluation de la confiance accordée au modèle, le score de familiarité avec ce type d'outil est relativement faible (1.8/5), sans surprise au vu de la spécificité de l'outil (prédiction d'âge de portrait photo). Pour l'intégralité des participants, il s'agissait de la "première fois [qu'ils faisaient] ce type d'exercice". Les scores de fiabilité de l'outil, de propension à faire confiance à l'outil et la confiance déclarée dans l'automatisation de ce type de tâche sont légèrement en dessous de la moyenne (entre 2.5 et 3). Les participants déclarent le plus souvent ne pas se sentir "particulièrement influencé[s]" et trouver légitime de "ne pas faire confiance au modèle étant donné qu'il lui arrive d'avoir une confiance faible, mais aussi de se tromper alors que sa confiance est forte". Le fait que le modèle puisse être moins sûr de ses propres prédictions incitait les participants à être plus vigilants vis-à-vis de ce qu'il proposait, et à questionner ses propositions. La compréhension de ce que propose le modèle a atteint un score de 3.22 sur 5, ce qui est plutôt moyen et peut s'expliquer par le fait qu'il était "facile de comprendre ce que proposait le modèle", mais beaucoup moins de "comprendre comment il s'est trompé". Les participants ont tendance à anthropomorphiser les erreurs du modèle (attribuer des caractéristiques humaines aux comportements de l'outil), considérant que ces erreurs auraient pu être commises par des humains pour les mêmes raisons qu'ils attribuent au modèle (ex. un visage marqué par des rides, un athlète, des vêtements). Le score d'intention des développeurs est le plus élevé (3.7/5), pour rappel cette dimension se réfère à la motivation des développeurs à concevoir ce type d'outil pour le bon usage des utilisateurs. Lors de l'accueil des participants, nous leur avons déclaré que le modèle était conçu uniquement à but de recherche sans donner plus de détails. Bien que certains participants trouvaient louable le fait de concevoir ce type de modèle pour voir s'il pouvait aider l'humain sur ce type de tâche, d'autres étaient plus sceptiques, présentant la conception de ce type d'outil comme "dangereux s'il arrive à faire ce que fait l'humain". Au vu des précédents éléments, il n'est pas surprenant de retrouver des scores moyens de propension à faire confiance à l'outil et de confiance en l'automatisation de ce type de tâche (respectivement 2.56/5 et 2.73/5).

Comme exposé précédemment, la tâche est très spécifique (estimation d'âge de portraits photo) et à faible degré d'importance. Si nous regardons les résultats du TAM, ceci pourrait expliquer que les scores d'utilité perçue et l'intention d'usage oscillent entre 2 et 3.5 (respectivement 3.083/5 et 2.25/5). La facilité d'utilisation perçue a un score plus élevé

(3.489/5). Notre hypothèse concernant ce résultat est qu'il n'y a pas besoin d'interagir avec l'outil pour qu'il produise une prédiction, il l'affiche nativement. Cependant, il faut aussi prendre en compte la charge cognitive engendrée lorsque 1) la confiance est faible et que l'on doit questionner le jugement de l'outil ou 2) lorsque l'on pense que l'outil se trompe.

7.3 Cartographie des unités de sens

En complément de cette analyse des résultats des questionnaires et de la mise en lien avec les discours recueillis lors des entretiens, nous avons examiné les discours en effectuant une cartographie des unités de sens. Cette méthode d'analyse qualitative de discours consiste à explorer la structure sémantique et les relations entre les éléments d'un discours [144] [53]. Elle nécessite d'identifier, extraire et organiser les unités de sens significatives du discours des participants. Dans le cadre de notre expérimentation, nous nous sommes intéressés aux discours des participants sous le prisme de l'acceptabilité et de la confiance en l'IA. Cette cartographie nous permet d'examiner le discours des participants graphiquement en faisant ressortir comment ces unités de sens interagissent et contribuent à la compréhension des comportements et perceptions des participants. Le discours des participants gravite fortement autour de la confiance qu'ils accordent ou non à l'IA et sur un comparatif avec l'humain.

Comme nous pouvons le voir sur la figure 7.1, les dimensions les plus saillantes des discours sont reliées aux différents propos récurrents dans les discours des participants. La confiance dans le modèle prédictif apparaît comme un thème central. D'une part, elle représente un risque selon les utilisateurs de créer des biais dans leur jugement. Les recommandations du modèle peuvent influencer les décisions des utilisateurs. De manière extrême, ils pourraient s'appuyer trop fortement sur les recommandations de l'outil, ce qui soulève la question de la dépendance et de la responsabilité dans la prise de décision. D'autre part, les utilisateurs apprécient tout de même "l'honnêteté" du modèle, lorsqu'il manifeste sa propre incertitude, avec un taux de confiance déclarée faible. Cela souligne la valeur de la transparence de l'IA en termes de renforcement de la confiance des utilisateurs. Les utilisateurs expriment également un certain degré de confusion et de doute, notamment lorsque les prédictions de l'IA diffèrent de leurs estimations personnelles, ou lorsqu'ils observent une erreur de la part du modèle alors que sa confiance déclarée est élevée. Ces moments de divergence conduisent à une remise en question de l'outil et à une perte de confiance, soulignant la nécessité de développer des solutions IA qui peuvent calibrer de

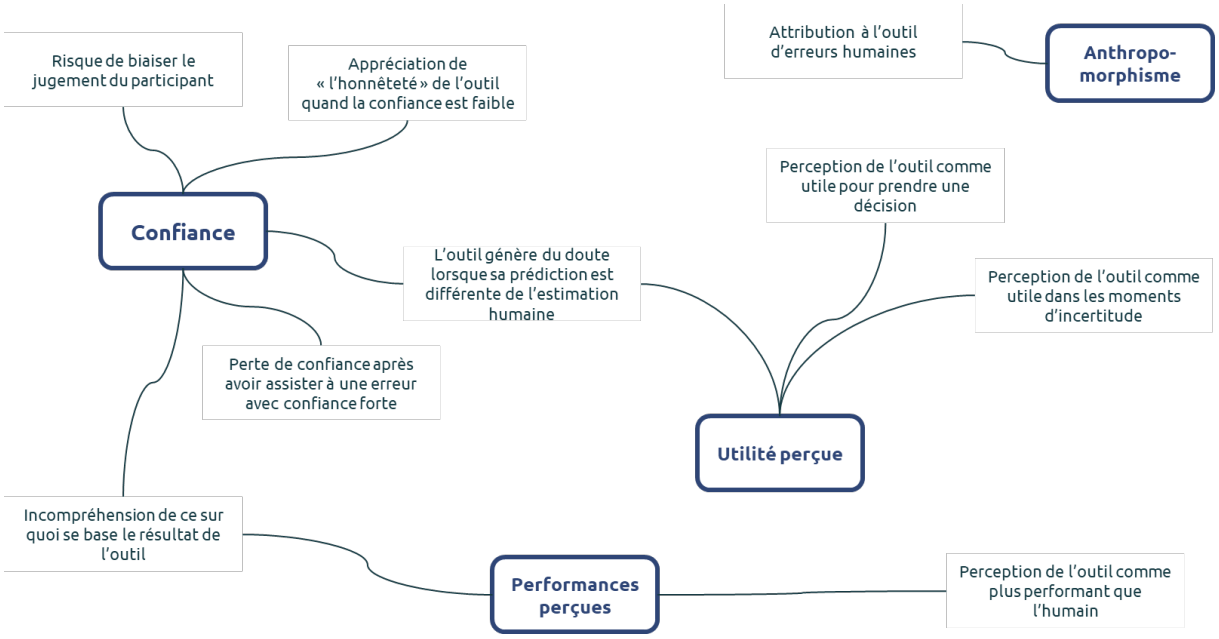


FIGURE 7.1 – Cartographie des unités de sens

manière précise et compréhensible leur niveau de confiance affiché. De plus, la complexité inhérente à certains modèles peut laisser les utilisateurs perplexes quant aux fondements des décisions de l'outil. La démystification des processus de ces outils semble donc d'autant plus cruciale pour bâtir une confiance éclairée.

L'anthropomorphisme est un autre élément évoqué. Les participants ont tendance à attribuer au modèle des caractéristiques humaines. Cela peut à la fois faciliter son acceptation et aussi poser des problèmes si les utilisateurs s'attendent à ce que le modèle agisse et réagisse comme un être humain. Cette personnification peut créer des attentes irréalistes et des malentendus concernant les capacités de l'outil.

L'utilité perçue de l'outil en tant qu'assistant de prise de décision est également soulignée. L'outil est considéré comme particulièrement utile pour aider à prendre des décisions. Il permet ainsi, lorsque les utilisateurs sont confrontés à l'incertitude, de bénéficier du support que le modèle peut offrir, ce qui peut renforcer son adoption.

Enfin, les performances perçues de l'outil par rapport à celles des humains varient. Certains utilisateurs perçoivent le modèle comme surpassant les capacités humaines, ce qui peut refléter une reconnaissance de la supériorité technique des solutions IA sur l'humain

dans certains domaines. Cela peut augmenter la probabilité d'une adoption plus large si cette perception est maintenue et renforcée par des expériences utilisateur positives et des résultats précis.

Chacun de ces éléments fournit un aperçu de la perception et des attitudes complexes envers un outil d'aide à la décision et souligne l'importance de considérer ces dimensions lors de la conception et du déploiement de ce type d'outil.

7.4 Conclusion

En conclusion, ce chapitre met en lumière les multiples facteurs qui influencent la confiance et l'acceptabilité des solutions d'IA en entreprise, spécifiquement dans le contexte de la prédiction d'âge à partir de portraits photo. L'analyse des résultats des questionnaires TiA et TAM, combinée à la cartographie des unités de sens issues des entretiens, révèle une complexité dans les perceptions des utilisateurs. Les dimensions telles que la fiabilité, la compréhension et la transparence de l'outil jouent un rôle crucial dans l'établissement de la confiance. Les utilisateurs expriment une certaine méfiance lorsqu'ils observent des incohérences ou des erreurs, surtout si l'outil affiche une haute confiance dans ses prédictions erronées. Cependant, la transparence de l'IA, notamment lorsqu'elle communique ses incertitudes, semble renforcer la confiance des utilisateurs, soulignant ainsi l'importance de développer des solutions IA capables de calibrer et d'afficher de manière précise et compréhensible leur niveau de confiance.

Conclusion de la Partie II

Dans le cadre de cette thèse, nous nous sommes intéressés aux solutions intégrant de l'Intelligence Artificielle, notamment comment les déployer en contexte professionnel et les rendre acceptable par les employés dont les postes de travail vont être impactés. Dans un contexte d'entreprise, il n'est pas possible de s'intéresser qu'aux utilisateurs finaux pour déployer des solutions IA acceptables. Il existe tout un réseau de parties prenantes à ces projets IA qui est concerné, et impactant sur cette acceptabilité. Il faut une considération plus large des besoins. Bien évidemment nous nous sommes donc intéressés aux utilisateurs, mais également aux équipes commerciales, aux managers, aux décideurs de la mise en place de ce type d'outil. Notre état de l'art de la Partie I de ce manuscrit a permis de faire remonter qu'il n'est pas seulement affaire de performance lorsque nous parlons d'acceptabilité, mais qu'il est aussi affaire de considération de l'humain dans les choix de conception, d'inclusivité dans leur conception et intrinsèquement de la confiance que les parties prenantes arrivent à accorder à ces solutions IA.

Pour nos recherches, nous nous sommes focalisés sur SIGMA Informatique et ses partenaires qui ont lancé des projets IA en identifiant ce domaine comme un axe de développement 1) économique, 2) de leur savoir-faire, 3) de leur productivité, 4) des services à mettre à disposition à leur clientèle. L'expertise en IA n'étant que très peu présente chez ces entreprises, la principale stratégie mise en place est de solliciter des prestataires spécialisés dans l'exploitation de données, la conception de modèles IA et leur déploiement dans des outils informatiques. Cela permet de proposer des produits qui, techniquement, répondent aux attentes initiales : pour une tâche donnée, elles sont, pour la plupart, efficaces (atteinte du but) et efficientes (performantes). Cependant, elles n'offrent que peu de satisfaction globale au réseau de parties prenantes à ces projets, dont les utilisateurs, en raison d'un décalage dans les attentes et perceptions de ces solutions. Ainsi nous avons pu constater une certaine volonté, de la part de SIGMA Informatique, à créer une démarche inclusive de conception de solutions IA. Mais cette volonté nécessite un réel investissement et une montée en compétence pour pouvoir réellement faire effet. L'inclusivité des utilisateurs dans la démarche de conception s'en retrouve la plupart

du temps partiel, avec des décisions parfois arbitraires de la part des concepteurs et/ou décideurs. Cette démarche est ainsi, bien souvent, trop technocentrée, là où il y a un réel besoin d'accentuer la prise en compte de l'UX ce qui peut conduire une adoption imposée de certaines solutions IA pour lesquelles les utilisateurs sont récalcitrants et/ou en font un mésusage.

Nous avons pu constater des besoins essentiels exprimés par les utilisateurs parfois ignorés car ils n'étaient pas en adéquation avec les attentes des concepteurs ou décideurs quant à l'intérêt du projet. Les conséquences peuvent parfois être minimes comme avec l'optimisation du Capacity Planning où l'utilisateur apprécie et utilise la solution IA mais certains de ses besoins ignorés par les concepteurs montrent la nécessité d'itérer pour améliorer l'ergonomie de l'interface utilisateur et des fonctionnalités propres au modèle IA déployé dans l'outil (linéarité des prédictions) (voir chapitre 4). A l'inverse, il y des situations où le trop grand décalage entre les attentes et les perceptions des parties prenantes au projet IA a causé un abandon du projet. Dans le cas du Support augmenté, le SI ne croyait pas en l'intérêt de l'outil, les concepteurs craignaient une trop grande charge de travail pour proposer un outil performant et les utilisateurs n'avaient jamais entendu parler de l'outil.

Le projet de commande de marchandises a été déterminant dans notre volonté de réaliser une expérimentation en milieu contrôlé. Les utilisateurs finaux [1] ont pu tester la mise à jour IA de leur outil de commande et ont ensuite exprimé un besoin de bénéficier de plus de transparence sur comment le modèle IA faisait ses recommandations. Plus particulièrement, ils expliquaient vouloir savoir à quel point le modèle renvoie un indice de confiance en ses propres prédictions. Ce qui nous a conduit à réaliser une expérimentation pour étudier l'effet de la confiance déclarée d'un modèle de prédiction en ses propres prévisions sur la confiance accordée par l'opérateur humain qui prend la décision finale. Les résultats nous ont montré que plus l'opérateur humain bénéficie d'informations du modèle sur sa prise de décision, plus il est susceptible d'être en accord avec ce dernier quant bien même la prédiction est fausse dans certains cas. Cet accord Humain-IA qui nous sert de métrique pour estimer la confiance accordée par l'opérateur humain. Cependant, la valeur de l'indice de certitude (classé en faible ou forte dans notre cas) ne semblait avoir d'impact direct sur les décisions de l'opérateur humain. Il a besoin d'avoir plus d'informations

1. gestionnaires des stocks de marchandises

mais la valeur de cette information ne semble statistiquement pas impactante. En faisant verbaliser les participants à notre expérimentation, nous nous rendons compte que les opérateurs humains sont consciemment plus à l'aise avec un outil qui déclare ne pas être sûr de ses prédictions. Un tel outil invite l'utilisateur à être plus attentif dans le choix de la décision finale. En revanche, quand l'outil affiche une certitude forte en ses propres prédictions, l'humain ne lui laisse pas le droit à l'erreur. Ce deuxième cas semble être très impactant. Les participants nous ont déclaré perdre toute confiance dans les décisions du modèle IA après l'avoir vu commettre des erreurs grossières de jugement avec un indice élevé de confiance.

Nous avons eu l'occasion d'utiliser des méthodes issues de l'évaluation subjective de la QoE comme moyen d'estimer cette confiance. Mais l'objectif de cette thèse est, et a toujours été, de proposer la démarche la plus holistique possible afin de rendre compte du phénomène d'acceptabilité des solutions IA en entreprise de la manière la plus globale et compréhensible possible. C'est pour cela qu'il nous apparaît très important d'également mobiliser des méthodes d'enquête, essentielles pour collecter des données qualitatives et quantitatives directement auprès du réseau de parties prenantes et des utilisateurs cibles. Dans notre expérimentation, les questionnaires TiA et TAM ont permis de recueillir rapidement des informations standardisées auprès de ces personnes qui sont essentielles pour estimer la confiance accordée à la solution IA et l'intention de s'en servir. Les entretiens réalisés [2] nous ont offert une compréhension plus approfondie de la perception des salariés et de ce qui faisait qu'elles trouvaient les solutions IA acceptables ou non et pour quelles raisons.

L'aboutissement des deux premières parties de cette thèse est que l'acceptabilité des solutions IA en entreprise repose sur la performance réelle de la solution par rapport à ce qui en était attendu par les différentes parties prenantes au projet IA et les utilisateurs cibles. Mais elle repose aussi sur la considération des perceptions et de la confiance de ces salariés. De plus, la place qui leur est accordée dans la conception et le déploiement de ces solutions, est également impactante pour leur faire adopter ces solutions IA de manière consentie.

2. De manière exploratoire au sein de SIGMA Informatique, à des fins d'approfondissement ou encore en post-expérimentation

Depuis plusieurs années, le domaine de l'IA promet aux entreprises de pouvoir automatiser une partie, voire l'entièreté de certaines tâches et de permettre aux employés d'être plus performant et de bénéficier de davantage de bien-être sur leur poste de travail. Cela passe par un gain de productivité, une accélération de l'innovation et une réduction de l'erreur humaine. C'est pour ces raisons-là que la décision de déployer des solutions IA sur les postes de travail est presque toujours revenue à la hiérarchie. Mais depuis début 2023, nous constatons un bouleversement, une forte tendance des employés à importer eux-mêmes des solutions IA sur le poste de travail. Ces solutions sont les outils génératifs. Ces outils sont directement amenés par les employés, demandant une adaptation de la part de l'organisation face à ce comportement. Ce changement de façon de faire dans le déploiement de solutions IA en entreprise pour les outils génératifs nous amène à penser qu'il peut y avoir un bouleversement dans la manière de considérer l'acceptabilité de ce type de solutions en particulier.

Troisième partie

L'émergence des outils génératifs : des facteurs d'acceptabilité différents ?

Chapitre 8

Les LLM dans la collaboration Humain-IA

Dans ce chapitre

Ce chapitre explore l'arrivée des outils génératifs dans le contexte professionnel. Nous y explorons notamment les spécificités de ces outils avec l'arrivée au grand public. Ensuite, nous nous intéresserons aux méthodes employées pour estimer la performance des outils génératifs. Et enfin, nous nous intéresserons aux opportunités à intégrer ces outils dans le développement informatique.

8.1 Introduction

L'évolution rapide de l'IA a conduit à une prolifération de technologies innovantes qui redéfinissent progressivement notre façon de travailler. Au coeur de cette évolution, les modèles de langage à grande échelle, abrégés LLM (pour *Large Language Model*), sont apparus. Ces modèles sont conçus pour traiter et générer du contenu en réponse à des requêtes humaines (prompts). Les LLM se distinguent par leur pertinence et leur précision accrues. D'un point de vue technique, cela est permis par l'utilisation de très grande quantité de données lors de l'entraînement, combiné à des architectures de réseaux de neurones spécialisés dans le traitement du langage (NLP, pour *Natural Language Processing*), telles que les *Transformers*.

Contrairement aux précédentes approches comme les réseaux de neurones récurrents (RNN, pour *Recurrent Neural Network*), les architectures de type *Transformers* n'effectuent pas des calculs séquentiels. Le calcul séquentiel consiste à partir d'un état de départ qui

est mis à jour progressivement avec les informations entrées dans chaque séquence. En revanche, les architectures *Transformers* utilisent le calcul parallèle, ce qui leur permet de traiter de l'information simultanément en profitant de la puissance de calcul offerte par les GPU (*Graphic Processing Unit*). De plus, les mécanismes d'attention de ces architectures, qui permettent au modèle de sélectionner les informations les plus pertinentes des données d'entrée pour accomplir la tâche donnée, améliorent considérablement leur performance. Leur prise en compte du contexte est bien plus fine et efficace en pondérant chaque mot du prompt sous la forme de *tokens*. Les architectures de type *Transformers* permettent ainsi aux LLM de traiter le langage de manière différente par rapport aux approches antérieures dans le domaine du NLP. Elles sont utilisées pour des tâches variées telles que la traduction automatique, la génération de texte, et même la création d'images. Les LLM obtiennent ainsi des capacités accrues en termes d'efficacité en accentuant la recherche des informations les plus pertinentes. Enfin, ces architectures contribuent également à donner plus de flexibilité au modèle, qui est capable de répondre à un vaste éventail de tâches. Ainsi, les LLM peuvent comprendre et de générer du contenu de manière cohérente, tout en permettant des interactions plus naturelles et constructives entre les humains et les solutions IA [33].

En termes d'adoption au travail, les outils génératifs [1] semblent susceptibles de présenter des facteurs spécifiques d'acceptabilité professionnelle avec une interaction Humain-IA très proche des interactions Humain-Humain. Avec l'arrivée de ces outils grand public tels que ChatGPT de OpenAI [136], nous posons l'hypothèse qu'ils sont introduits par les employés eux-mêmes sur le poste de travail plutôt qu'à la suite d'une décision hiérarchique. Cette hypothèse tient du fait qu'une réelle effervescence se tient autour des outils génératifs et que les employés en perçoivent une réelle amélioration de leur confort de travail et de leur performance [192]. Nous détaillerons ces points au fur et à mesure de ce chapitre en faisant un focus sur le domaine de l'informatique. Cette reconfiguration de l'introduction des solutions IA sur les postes de travail amène de nouvelles questions et de nouveaux défis pour les organisations, nous incitant également à reconsidérer la façon dont la confiance et l'acceptabilité des solutions IA sont mesurées [98].

Dans ce chapitre, nous mettrons l'accent sur le cas de ChatGPT, une référence en termes d'outil génératif. Il fait partie des premiers outils accessibles au grand public et

1. solutions informatiques servant d'interface pour interagir avec les LLM

reste à l'heure actuelle l'un des plus utilisés. Ensuite, nous explorerons dans ce chapitre les méthodes actuelles d'évaluation des LLM qui sont mises en avant dans la littérature. Au vu de notre contexte de recherche au sein de SIGMA Informatique, nous proposons ensuite une attention particulière sur l'évaluation des LLM pour des tâches de développement informatique. Ces évaluations sont cruciales pour comprendre les atouts et les limites de ces technologies. Mais aussi pour fournir des recommandations sur leur intégration dans les environnements de travail pour une collaboration Développeur-LLM efficiente. Et enfin, nous discuterons des opportunités et des limites associées à l'utilisation des LLM en contexte professionnel, explorant comment ces outils peuvent être à la fois des catalyseurs d'efficacité et des sources potentielles de risques [69]. Ainsi, ce chapitre vise à enrichir notre compréhension des implications des LLM dans la dynamique de collaboration Humain-IA, et à mettre en lumière les facteurs qui peuvent influencer l'acceptabilité de ces technologies émergentes dans le contexte professionnel.

8.2 L'avènement de ChatGPT : vers de nouvelles opportunités et préoccupations en entreprise ?

Arrivé fin 2022, ChatGPT de OpenAI est le premier outil génératif grand public [136]. Il est capable de comprendre le contexte d'une conversation pour générer des réponses appropriées. De plus, l'outil de OpenAI est capable de générer des réponses dans différents styles (formel comme informel) pour pouvoir ajuster sa manière d'interagir avec l'utilisateur. ChatGPT est un outil génératif sous la forme d'un chatbot qui utilise le LLM GPT-3.5 (pour *Generative Pre-Trained*) et ses versions ultérieures. Conçus pour diverses applications et ayant fortement gagné en visibilité depuis début 2023, ChatGPT et les outils génératifs similaires sont très prisés dans de nombreux domaines.

Bénéfices liés à l'usage de ChatGPT L'étude de Deng et Lin, réalisée fin 2022, met en avant plusieurs avantages liés à l'utilisation d'outils tels que ChatGPT. Ils soulignent notamment son efficacité, sa précision et son potentiel de réduction des coûts dans la réalisation d'une multitude de tâches [50]. ChatGPT est présenté comme capable d'améliorer les processus d'interactions, pour faciliter les échanges, permettant un gain significatif de temps et de ressources. Grâce à ses compétences linguistiques avancées, l'interaction avec ChatGPT se fait en langage naturel (similaire à l'interaction Humain-Humain) et il

est capable de tenir une conversation en retenant les informations localement (tous les prompts et réponses envoyés en retour dans une même conversation), ce qui contribue à sa capacité de compréhension très précise et améliore donc l'expérience d'usage. L'outil génère des réponses adéquates en se basant sur un large corpus de connaissances qu'il possède. Á partir de ces connaissances, il peut saisir le contexte d'une conversation et proposer des réponses pertinentes de manière réaliste et engageante [2], ce qui améliore constamment sa précision et son adaptabilité à de nouveaux contextes.

Toujours d'après Deng et Lin, la démocratisation des outils génératifs tels que ChatGPT est décrit comme une opportunité de réduction des coûts pour les entreprises en générant instantanément des réponses semblables à celles des humains. Ce qui permet dans un cadre professionnel réduire le recours à des opérateurs humains, dont l'action peut être coûteuse en ressources humaines, en temps, en effort, etc. De plus, avec une faculté d'apprentissage qui s'améliore avec le temps, le besoin de mises à jour manuelles et onéreuses est réduit, rendant ChatGPT une solution efficace et économique. Ainsi, sans trop d'effort lié à l'interaction, l'opérateur humain pourrait rapidement obtenir des informations, et ce, de manière plus ou moins élaborée en fonction de la granularité souhaitée dans la réponse.

Inquiétudes liées à l'usage de ChatGPT En retour, les chercheurs soulignent également plusieurs défis à surpasser dans l'utilisation des outils tels que ChatGPT, incluant des inquiétudes liées à la sécurité et à des capacités restreintes [3] [50]. Concernant la sécurité, les outils comme ChatGPT peuvent présenter des vulnérabilités. La possibilité que ChatGPT soit exploité pour propager de fausses informations ou de la propagande peut contribuer à des risques sociétaux. De plus, la capacité de ChatGPT à créer des textes semblables à ceux d'un humain peut faciliter l'usurpation d'identité. Les entreprises et organisations doivent donc examiner ces risques avec attention et adopter des mesures adéquates lorsqu'elles emploient ChatGPT ou des technologies similaires. Ce n'est pas tout puisque malgré la puissance de ChatGPT en tant qu'outil génératif, il ne peut produire du contenu qu'à partir d'informations qu'il a déjà apprises. Le solliciter sur des sujets qu'il ne connaît pas peut contribuer à la génération d'informations fausses, mais répondant aux requêtes envoyées : on parle d'hallucinations. Plus précisément, les hallucinations chez les LLM

2. ChatGPT incite l'utilisateur à alimenter le dialogue, à être plus précis dans ses requêtes
3. ChatGPT ne génère du texte qu'en fonction des prompts reçus, n'a pas accès à des informations externes et ne navigue pas sur Internet

désignent le fait que le modèle génère des réponses qui contiennent des inexactitudes factuelles ou des informations qu'il a fabriquées [143].

Nous partageons l'idée que les outils tels que ChatGPT doivent encore surpasser un certain nombre de défis pour être qualifiés d'IA de confiance, notion définie au chapitre 1 [108] [187] [86]. D'un point de vue éthique, ces outils génératifs ne semblent pas pleinement prendre en compte les recommandations visant à valoriser certains principes clés. L'action humaine est essentielle pour une collaboration Humain-IA efficace. Actuellement, les outils génératifs, bien qu'ils soient conçus pour répondre aux prompts, peuvent réduire significativement l'implication humaine sur le résultat final. Comme le souligne Yann Ferguson [4], la transparence est également un défi majeur avec les outils génératifs tels que ChatGPT en raison de la complexité des LLM et de leur nature souvent opaque, rendant difficile pour les utilisateurs de comprendre comment les réponses sont générées [60]. Développer des interfaces et des explications qui dévoilent les mécanismes sous-jacents des LLM, y compris leurs données d'entraînement et les limites de leurs capacités, pourrait grandement améliorer la transparence.

Concernant la protection des données, il est impératif que les données envoyées aux outils génératifs soient traitées de manière à respecter la confidentialité des utilisateurs. Cela nécessite l'adoption de politiques strictes pour assurer que les données personnelles ne sont ni mal utilisées ni exploitées sans consentement, conformément au RGPD. Dans le milieu professionnel, où les employés manipulent parfois des données sensibles, il est crucial de définir clairement les données pouvant être partagées et celles qui doivent rester confidentielles, en tenant compte des enjeux de propriété et de sécurité.

L'impact social et environnemental de ces technologies soulève également des questions, notamment en raison de la consommation énergétique élevée requise par les serveurs entraînant et exécutant les LLM. Il devient donc important de comprendre si l'aspect énergivore des LLM a un impact sur leur acceptabilité selon la situation de travail. Au niveau plus global, il devient nécessaire de concevoir des LLM économes en ressources pour minimiser leur empreinte carbone face aux défis environnementaux.

Dans un contexte social, les entreprises doivent également veiller à ce que ces nouvelles

4. A l'AI ACT Day de Impact AI en 2023

technologies apportent une contribution positive aux opérateurs humains sans accentuer les disparités sociales existantes, ce qui est un fait courant dans le déploiement de nouvelles technologies selon les recherches de Lelong et ses collaborateurs [107].

Pour résumer, les outils génératifs comme ChatGPT semblent encore trop s'exempter des principes de l'IA de confiance abordés au chapitre 1. Il nous semble important de porter une attention particulière à ces aspects, en mettant l'accent sur une utilisation responsable et éthique. Cependant, cela n'empêche pas d'avoir une trajectoire d'adoption des IA génératives en forte croissance. Malgré de multiples échecs de projets d'IA en entreprise [163], des outils comme ChatGPT sont directement importés au travail par les employés. C'est ce que souligne Yann Ferguson, qui ajoute que ces outils, capables de fournir des réponses rapides et souvent exactes, offrent la capacité d'optimiser les ressources et améliorer l'efficacité opérationnelle des entreprises [60]. Nous supposons que, contrairement aux outils imposés par la hiérarchie, les employés apprécient ces outils génératifs pour leur grande permissivité et leur adaptabilité à divers cas d'usage, permettant aux employés de définir leurs propres pratiques.

Maintenant que nous avons présenté les outils génératifs à travers l'exemple de ChatGPT et exploré leur intégration dans le milieu professionnel, nous allons nous pencher vers les méthodes d'évaluation des modèles sur lesquels reposent ces outils : les LLM. Évaluer les capacités des LLM permet d'identifier s'ils répondent à des attentes de conception et s'ils peuvent s'intégrer dans des cas d'usage spécifiques. Ainsi, nous pourrons déterminer quels LLM répondent le mieux aux besoins opérationnels et éthiques des entreprises, favorisant une adoption plus efficace et responsable de ces technologies.

8.3 Comment évaluer les LLM ?

Dans le cadre de la conception des LLM, la phase d'entraînement [5] est complétée une phase d'évaluation en situation écologie, c'est-à-dire qu'après entraînement les LLM sont évalués sur des cas d'application bien spécifiques pour étudier leurs capacités réelles. Cette seconde étape est également cruciale pour avoir une vision de leur applicabilité dans des environnements interactifs [113]. Les moyens d'évaluation sont assez variables pour estimer les capacités d'un LLM, et rendre compte d'une certaine manière de sa qualité.

5. Qui comprend sa propre partie d'évaluation de compétences sur la base d'indicateurs de réussite

Pour bénéficier d'une vision globale de ces capacités, nous estimons qu'il est nécessaire de multiplier les métriques. Nos recherches montrent que les concepteurs de LLM ont tendance à présenter la performance de leur modèle à l'aide des benchmarks avec des scores obtenus grâce à des jeux de données d'évaluation [34] [126]. Comme nous le montrent Humza Naveed et ses collaborateurs (2023), il existe une multitude de jeux de données d'évaluation de LLM, et même des frameworks multitâches comme MMLU (pour *Massive Multitask Language Understanding*) [78]. Pourtant nous constatons que les six critères les plus prisés par les concepteurs de LLM pour l'évaluation de leur modèle semblent être : 1) la compréhension du monde, 2) les résultats populaires agrégés, 3) le niveau en mathématiques, 4) la capacité à générer du code informatique, 5) la compréhension de la lecture, 6) le raisonnement/bon sens. Le détail de ces dimensions est disponible dans l'annexe 10.5.

Chacune de ces dimensions offre un aperçu distinct des capacités des LLM et de leur potentiel d'application dans un contexte professionnel. Elles forment une base solide pour évaluer les forces et les faiblesses des LLM, et pour comprendre comment ces modèles peuvent être intégrés efficacement dans des environnements de travail pour améliorer la collaboration Humain-IA.

8.3.1 Dans le domaine du développement informatique

Notre exploration de la littérature nous montre deux types d'interaction entre les développeurs et les LLM pour produire du code informatique avec des LLM : de la génération intégrale de code (ex. demander au LLM de générer une fonction, voir figure 8.1) [6] et de la génération partielle (ex. le LLM complète une fonction que le développeur a commencé à rédiger, voir figure 8.2) [7]. C'est en tout cas la distinction que nous propose Chen et ses collaborateurs en 2021 [37].

Génération totale Pour les chercheurs, la génération totale fait référence à la production de code informatique pour une spécification ou un problème donné sous la forme d'une instruction (prompt). L'utilisateur transmet une requête en langage naturel et le LLM retourne du code répondant au prompt.

6. Dans cet exemple, l'utilisateur rédige sa requête et le modèle lui retourne une réponse.
7. Dans cet exemple, l'utilisateur commence à rédiger son code informatique et le LLM lui propose, en gris, de compléter ce même code

FIGURE 8.1 – Exemple de génération intégrale de code par LLM, selon Chen et ses collaborateurs [37]

Génération partielle La génération partielle tient en la complétion et/ou modification d'extraits d'un code existant en fonction du contexte dans lequel le code initial a été produit et des instructions que l'on va donner au modèle, si on lui en donne [37]. Cet usage trouve surtout son utilité dans des besoins de modification, de révision ou de débogage d'un code existant [117]. C'est notamment le cas d'outils tels que GitHub Copilot. Par exemple, le développeur commence à écrire une fonction Python en la nommant et le LLM en déduit le reste de la fonction à rédiger. Dans ce cas de figure, la génération peut être dynamique puisque le modèle propose en temps réel une génération en réponse au contenu rédigé par l'utilisateur.

Maintenant que nous avons exploré comment les utilisateurs peuvent interagir avec les LLM pour produire du code, nous allons nous intéresser à quels méthodes et outils sont employés pour estimer la capacité de ces LLM à produire du code correct.

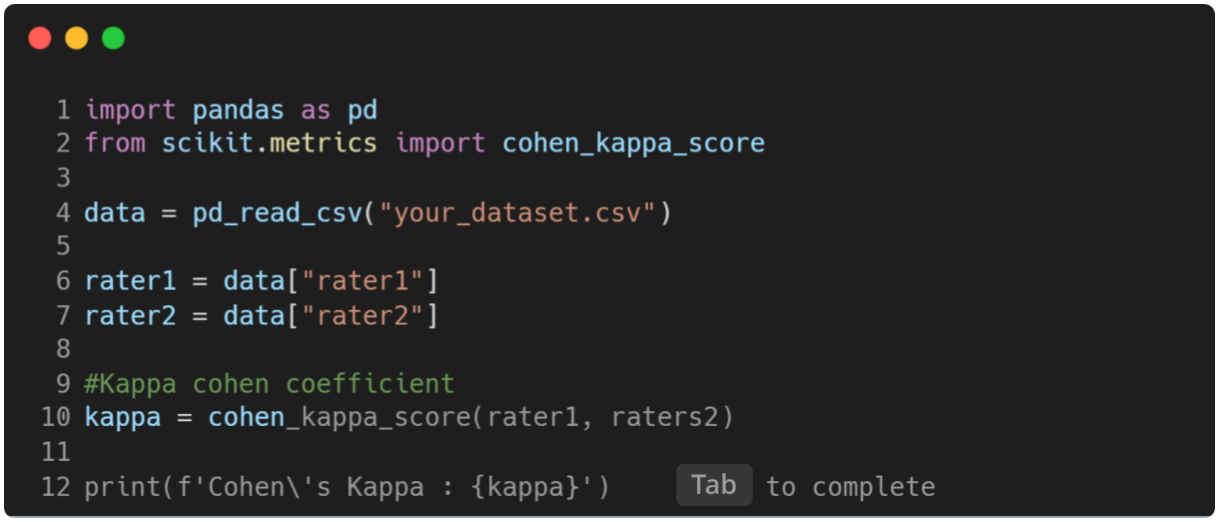


FIGURE 8.2 – Exemple de complétion de code par LLM, selon Chen et ses collaborateurs [37]

Évaluation de la capacité de production de code informatique par les LLM

D'après nos recherches bibliographiques, la capacité des LLM à produire du code correctement est principalement évaluée de manière automatisée à partir de métriques définies dans des jeux de données d'évaluation. Il existe actuellement deux jeux de données de référence, qui semblent faire autorité pour évaluer la capacité des LLM à produire du code : HumanEval [37] et MBPP [15].

Le fonctionnement de ces deux jeux de données d'évaluation, illustré à la figure 8.3, est assez similaire. Il s'agit d'extraire des problèmes de programmation, chacun accompagné de tests unitaires pour vérifier que le code généré fonctionne correctement. Une solution proposée est également fournie pour évaluer si le modèle a répondu aux attentes de la manière prévue.

HumanEval HumanEval est un jeu de données composé de 164 problèmes de programmation en Python [37]. Chaque problème est composé d'une fonction, d'une description

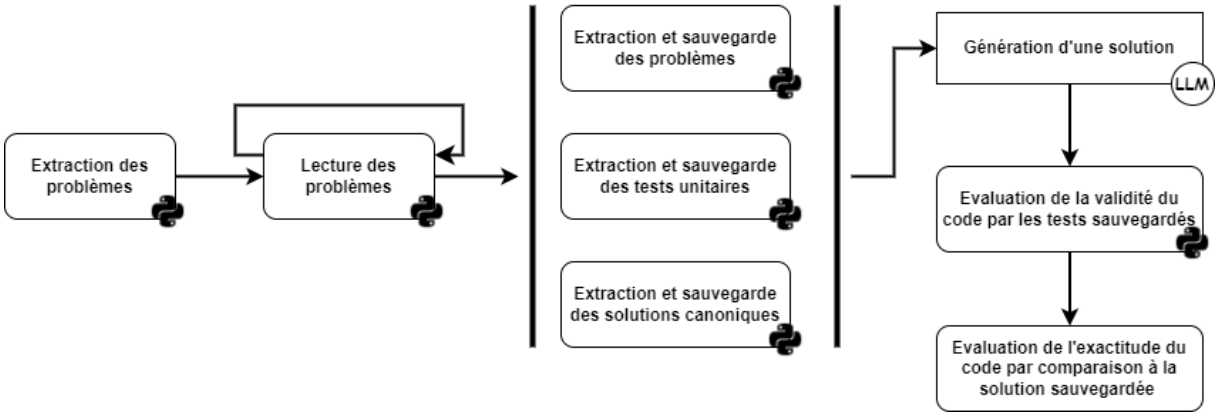


FIGURE 8.3 – Processus d'évaluation (génération du jeu de données + application) automatisée de la capacité des LLM à produire du code avec les jeux de données d'évaluation, selon les travaux de Yeticstiren et al. (2023) [183]

du problème sous la forme de docstring[8], d'un corps et de plusieurs tests unitaires. Ce jeu de données a été créé initialement pour évaluer le LLM Codex qui alimente l'outil GitHub Copilot. Comme Codex a été entraîné à partir de la plateforme GitHub, les problèmes de HumanEval ont été rédigés à la main pour limiter les biais d'évaluation avec des bouts de code qui auraient pu être utilisés pour entraîner ce même modèle.

MBPP MBPP (pour *Mostly Basic Python Problems*) est un jeu de données composé de 974 exercices courts de programmation en Python, construits en *crowdsourcing*[9] [15]. Chaque exercice de programmation est composé de 1) un court énoncé du problème sur le programme, 2) une fonction Python autonome servant à résoudre le problème spécifié, 3) trois cas de test pour vérifier que la fonction est sémantiquement exacte et 4) une solution aux trois cas de test.

La multiplication des frameworks d'évaluation dérivés de HumanEval et MBPP Le nombre de frameworks conçus pour évaluer la capacité des LLM à générer du code de qualité est depuis en pleine croissance. Ces frameworks visent à proposer une évaluation du code généré dans différents langages de programmation, on parle d'évaluation multi-

8. Chaîne de caractères utilisée pour documenter du code informatique. Le docstring est placé après la déclaration de l'entité. Il sert principalement à fournir une description du code, des paramètres que le code accepte, des exceptions potentielles et du résultat à attendre de l'exécution du code.

9. Cela signifie que les problèmes de programmation, ainsi que leurs solutions, ont été collectés à partir de contributions de nombreuses personnes, souvent via des plateformes où des utilisateurs peuvent soumettre et annoter des données.

lingue. Pour cela, les concepteurs de ces frameworks semblent souvent décliner les jeux de données HumanEval et MBPP en d'autres langages de programmation. Par exemple, le jeu de données HumanEval-X propose de nombreux problèmes qui sont des variations des problèmes initialement adressés dans HumanEval. Les problèmes de HumanEval sont réécrits manuellement dans quatre autres langages de programmation (C++, Java, JavaScript et Go) [190]. Dans chaque langage, HumanEval-X propose un même problème accompagné d'une fonction, d'une description du problème sous la forme d'un docstring, d'une instruction, d'une solution et de cas de test de manière similaire à ce que propose HumanEval. Nous pouvons également citer le framework MultiPL-E, qui de manière similaire à HumanEval-X, traduit les problèmes proposés par HumanEval, mais aussi par MBPP. Ce framework propose une déclinaison des problèmes proposés dans les deux jeux de données en 18 langages de programmation pour faire de l'évaluation multilingue [30].

Les limites de l'évaluation des LLM pour des tâches de génération de code informatique Nous constatons un certain engouement à l'utilisation des jeux de données HumanEval et MBPP pour évaluer les LLM utilisés pour produire du code. Nous avons recensé un certain nombre de LLM très prisé pour générer du code en fin 2023 dans le tableau 8.1, dont l'évaluation de leurs capacités de programmation se fait au moyen de ces jeux de données.

Cependant, il est également nécessaire de pointer les limites de ces approches. Tout d'abord, ces jeux de données servent à estimer un niveau de performance des LLM uniquement à partir de leurs capacités en Python, ce qui n'est donc pas représentatif de leur niveau dans les autres langages. Dans un deuxième temps, l'approche de traduction du Python vers d'autres langages peut faire face à des limites de traduction. Dans HumanEval, des problèmes sont parfois associés à des fonctions d'aide spécifiques à Python non traduisibles vers certains autres langages [31]. De plus, les deux jeux de données d'évaluation adressent des programmes Python très courts. Nous nous interrogeons donc sur la représentativité des résultats surtout si les LLM sont sollicités, en situation réelle, sur des programmes bien plus grands que ceux des jeux de données (ex. un code informatique qui fait des milliers de lignes). Bien que son niveau de performance puisse être adapté sur des programmes courts, le doute subsiste donc face à des programmes beaucoup plus longs.

Modèle	**Nombre de paramètres (en milliards)**	**HumanEval** (Score sur 100)	**MBPP** (Score sur 100)	**Source**
Code Llama	7 13 34	33.5 36 48.8	41.4 47 55	[147]
Llama 2	7 13 34 70	12.8 18.3 22.6.3 29.9	26.1 35.4 35.4 45.4	[164]
Mistral 7b	7	30.5	47.5	[92]
Starcoder	15	33.6	52.7	[109]
WizardCoder	15	57.3	51.8	[114]
GPT-3.5 Turbo	175	48.1	83.2	[179]
GPT-4	*	67	87.5	
CodeGeeX	13	22.9	24.4	[190]

TABLE 8.1 – Tableau récapitulatif des performances en génération de code de divers LLM, évaluées par les jeux de données HumanEval et MBPP (**valeur non communiquée, mais estimée à environ 1000 fois plus de paramètres de GPT-3.5*)

8.3.2 Vers une diversification des méthodes d'évaluation

Comme nous avons pu le présenter précédemment, il semble y avoir une forte tendance à l'évaluation automatisée des LLM en usant de jeux de données d'évaluation pour des dimensions spécifiques. L'inconvénient de cette approche est qu'elle ne rend compte des performances du modèle qu'en milieu contrôlé et sans interactivité avec l'humain, donc avec un faible potentiel de représentativité. Tout l'intérêt de ces évaluations est d'estimer les capacités des LLM afin de s'en servir en contexte réel, il est donc d'autant plus intéressant de multiplier les types d'évaluation, et notamment de valoriser la place de l'humain dans ces évaluations. A l'inverse, certains chercheurs estiment qu'avoir des évaluateurs humains créerait un biais de jugement en faveur des premiers LLM dits "forts" sur le marché tel que GPT-4 de OpenAI, qui seraient forcément préférés aux nouveaux modèles émergeant [178] [176].

Face à l'automatisation de l'évaluation des LLM, nous souhaitons nous intéresser à la valorisation de l'intervention humaine pour proposer une approche plus holistique.

Plus précisément, nous souhaitons explorer si une objectivation de l'évaluation par

les jeux de données d'évaluation est compatible à des méthodes d'évaluation centrée-utilisateurs, c'est-à-dire avec des utilisateurs finaux sur des tâches spécifiques. C'est ce que nous explorerons dans la section suivante.

8.4 Quelles opportunités de collaboration Humain-LLM dans des tâches de développement informatique ?

La problématique explorée précédemment nous semble symptomatique d'une tendance à diriger les solutions IA vers une totale automatisation. Que ce soit dans leur conception, avec l'évaluation automatisée via les jeux de données d'évaluation, ou dans leur usage en entreprise, générant des craintes de remplacement, parfois légitimes, chez les employés. Nous cultivons l'idée que faire collaborer les opérateurs humains avec les LLM est le meilleur moyen d'accroître la productivité au travail tout en contribuant au bien-être et attentes des employés.

Dans son étude sur la productivité et la qualité du code produit lors de l'utilisation de GitHub Copilot dans des tâches de développement logiciel, Imai (2022) a mené une expérience comparative [85]. Cette expérience vise à comparer les performances des binômes Humain-Humain et Humain-Copilot sur une tâche de programmation en Python d'un jeu de type Démineur. Les participants sélectionnés étaient des développeurs novices, répartis en deux groupes :

— Un groupe Humain-Humain, subdivisé en deux sous-groupes : les *drivers*, chargés de rédiger le script Python, et les *navigators*, responsables de contrôler et corriger le script rédigé.
— Un groupe Humain-Copilot, où les participants devaient collaborer avec GitHub Copilot pour produire le script.

Les binômes Driver-Navigator disposaient de 10 minutes chacun pour écrire et contrôler le programme respectivement. Les membres du groupe Humain-Copilot avaient quant à eux 20 minutes pour accomplir l'ensemble de la tâche. L'expérimentateur a évalué la qualité du code en fonction du nombre de lignes de code ajoutées ou supprimées dans les deux groupes. Pour ce faire, il a recueilli :

— Les mouvements oculaires de tous les participants à l'aide d'un oculomètre ;

— Pour le binôme Driver-Navigator, le nombre de lignes de code produites par le Driver et les modifications apportées par le Navigator.
— Pour le développeur du binôme Humain-IA : le nombre de lignes de code produites par le développeur et les modifications proposées par GitHub Copilot.

L'analyse du code a été effectuée en comparant le nombre de lignes ajoutées ou supprimées après chaque essai. Les résultats ont permis à Imai de conclure que la programmation avec GitHub Copilot générerait du code plus rapidement qu'avec le pair-programmation (Humain-Humain) sur une même période, bien que la qualité du code généré avec l'outil soit souvent inférieure.

La même année, Dakhel et ses collaborateurs ont effectué une autre étude comparative où ils testaient la qualité des solutions proposées par GitHub Copilot et par des développeurs humains (des étudiants en dernière année d'école d'ingénieur) sur un ensemble de tâches de programmation [42]. L'objectif était d'identifier si les solutions de GitHub Copilot sont compétitives avec celles des développeurs pour résoudre les problèmes de programmation, mais aussi de comprendre dans quelle mesure il peut être difficile de corriger les suggestions boguées de Copilot par rapport à celles des humains. Les deux agents (développeurs et GitHub Copilot) étaient donc en compétition et leurs productions étaient évaluées en termes de ratio de :

1. Solutions correctes ;
2. Besoin de correction des productions ;
3. Diversité dans les solutions correctes ;
4. Qualité globale des solutions.

Cette étude a mis en lumière que 1) le ratio de réponses correctes des opérateurs humains est supérieur à celui de Copilot, 2) la diversité des soumissions des opérateurs humains est supérieure à celles de Copilot, 3) le besoin/coût de correction des solutions boguées générées par Copilot est inférieur à celui des étudiants et 4) la complexité des codes générés par Copilot est inférieure à celle des étudiants. Cette étude se conclut donc sur le fait que l'outil génératif qu'est GitHub Copilot propose des scripts moins complexes et moins originaux que ceux des développeurs. Cependant il est plus facile de comprendre et corriger ses productions que celles des développeurs.

Deux ans plus tard, Ziegler et ses collaborateurs publient une étude sur l'impact de GitHub Copilot sur la productivité des développeurs [192]. Ils explorent le lien entre

l'utilisation des outils génératifs pour des tâches de complétion de code informatique et la productivité perçue ainsi que le bien-être des développeurs. D'abord, ils évaluent la productivité perçue de 2631 développeurs utilisant GitHub Copilot dans leur pratique avec un questionnaire basé sur le framework SPACE (décrit dans l'annexe 10.5). Ce framework mesure la productivité des équipes de développement en prenant en compte diverses dimensions comme la qualité du travail, le bien-être, et l'efficacité de la collaboration. Ce framework permet également d'interroger les participants sur leur compétence dans le langage pour lequel ils utilisent le plus GitHub Copilot et sur leur niveau d'expérience en programmation. Les chercheurs déclarent que les résultats de leur étude indiquent que les outils de complétion comme GitHub Copilot impactent la productivité perçue des développeurs, quel que soit le niveau de compétence. Mais les développeurs débutants seraient plus impactés que les autres. Ensuite, les résultats montreraient que GitHub Copilot réduit le temps de travail nécessaire tout en améliorant la qualité de production, la charge cognitive, le plaisir et l'apprentissage. Enfin, les chercheurs préviennent que l'exactitude des recommandations du LLM est importante, mais que l'utilité des suggestions comme point de départ pour un développement ultérieur l'est encore plus.

8.5 Conclusion

La discussion que nous avons développée au fur et à mesure de ce chapitre a pour but de réaffirmer le besoin de conserver, voire d'inclure davantage, l'humain dans la boucle de conception et d'évaluation des LLM. Cette vision que nous entretenons outrepasse la performance pure du modèle pour rendre compte de sa comptabilité avec les humains sur des tâches réelles, et de son positionnement par rapport aux employés qui sont impactés par son arrivée sur les postes de travail.

Par exemple, les méthodes de tests automatisées ne sont pas à ignorer, car elles apportent des indications précieuses sur la performance des LLM pour des dimensions données, de manière standardisée, et en réduisant les biais. Cependant, nous insistons sur l'intérêt de compléter cette approche avec de l'évaluation centrée utilisateur (ex. Une série pré-sélectionnée de LLM à l'aide de jeux de données d'évaluation pour leur score élevé qui sont ré-évalués par des développeurs sur des tâches de développement). Cette combinaison aurait un double effet : accroître la précision de la performance du modèle à partir d'un référentiel (le niveau moyen de l'échantillon humain observé) et favoriser

l'inclusion de l'humain dans la collaboration avec les solutions IA visant ainsi à rétablir des représentations de l'outil plus justes, comme préconisé dans la partie I.

Dans le chapitre 9, nous explorerons les dynamiques de SIGMA Informatique en termes de perception et d'usage des outils génératifs. Ce qui permettra, dans le chapitre 10, d'approfondir nos recommandations en présentant une méthodologie mixte qui intègre l'évaluation automatisée et l'inclusion des utilisateurs finaux dans le processus décisionnel de sélection du modèle pour un cas d'usage donné.

Chapitre 9

Quelle perception et quels usages des outils génératifs au sein de SIGMA Informatique ?

Dans ce chapitre

Ce chapitre décrit l'intégration des outils génératifs au sein de la Business Unit (BU) DIGITAL de SIGMA Informatique. Dans un premier temps, nous examinons, au moyen d'une enquête, la perception des outils génératifs par les employés. Au moment de cette enquête, début 2023, les collaborateurs SIGMA semblaient curieux de ces outils, mais aussi craintifs qu'on ne reconnaissent pas leur apport dans la réalisation de leurs tâches professionnelles s'ils sont assistés par des outils génératifs. Dans la deuxième partie de ce chapitre, nous présentons une étude comparative de la production de code Java par des développeurs avec et sans l'assistance d'un LLM. Bien que l'échantillon de participants soit restreint, nos résultats vont dans le sens de notre état de l'art du chapitre 8 avec un outil génératif qui semble réduire le temps nécessaire à la production de code informatique mais qui nécessite également un temps de vérification plus long.

9.1 Introduction

Comme étudiée dans les deux premières parties de ce manuscrit, l'IA est susceptible d'apporter de grands avantages aux employés. En automatisant les tâches simples et/ou

chronophages, elle libérerait du temps pour se concentrer sur des tâches professionnelles plus complexes et/ou à plus forte valeur ajoutée. D'après nos lectures, une particularité des outils génératifs en entreprise est qu'ils permettent aux employés de bénéficier d'un accompagnement sur une multitude de tâches avec une même solution IA. Cela leur donne ainsi la possibilité de créer leurs propres cas d'usage des solutions IA.

En nous penchant spécifiquement sur le domaine du développement informatique, la littérature souligne que les outils génératifs ont un potentiel pour améliorer le confort et la performance au travail [85] [42] [192]. En 2022, Nguyen et Nadi se sont également intéressés à la manière dont GitHub Copilot contribue à cette amélioration de la performance et facilite le travail des développeurs [127]. Les chercheurs ont soumis plusieurs problèmes de programmation à l'outil génératif et ont demandé des solutions dans quatre langages de programmation différents [1]. Ils ont ensuite analysé les codes générés selon des critères spécifiques : l'exactitude du code mesurée en termes de statut [2] et de complexité cognitive [3] et cyclomatique [4]. Ils ont conclu que GitHub Copilot constitue un bon point de départ, capable de fournir entre 60% et 91% de code correct selon le langage, sans différence significative de complexité entre les langages.

Au regarde de ces études, nous estimons que les outils génératifs offrent une expérience de travail différente, en particulier dans le domaine de l'informatique. C'est pourquoi dans ce chapitre nous explorons la perception et l'acceptabilité qu'ont les collaborateurs SIGMA des outils génératifs au travail. Dans un deuxième temps, nous nous intéressons à la collaboration développeurs-ChatGPT pour produire du code Java et et aux apports de ces outils dans leur pratique professionnelle.

1. Python, Java, JavaScript et C
2. Chaque code généré obtient un statut : Accepté : le code soumis passe tous les tests ; Mauvaise Réponse : le code soumis n'a pas d'erreurs, mais son résultat diffère de la sortie attendue pour au moins un cas de test ; Erreur de Compilation : le code soumis ne peut pas être compilé ; Dépassement de Temps Limite : le code soumis n'a pas d'erreurs, mais au moins un cas de test dépasse le temps d'exécution autorisé ; Erreur d'exécution : le code soumis échoue pour au moins un cas de test en raison d'erreurs pendant l'exécution, par exemple une division par zéro
3. Degré de compréhension du code, mesuré en termes de profondeur des boucles, de conditions imbriquées, et de structures complexes.
4. Nombre possible de chemins d'exécution du code généré

9.2 Est-ce que les collaborateurs de SIGMA Informatique s'intéressent aux outils génératifs ?

L'utilisation des outils génératifs semble devenir de plus en plus courante dans le domaine du développement informatique en raison de leur potentiel pour accompagner, voire parfois automatiser, la génération et l'optimisation de code, la création de documentation, la recherche d'informations, etc. Début 2023, nous avons donc réalisé une enquête via questionnaire pour identifier comment les collaborateurs de SIGMA informatique perçoivent les outils génératifs dans leur travail. Cette enquête a été réalisée au sein de la *Business Unit* (BU) DIGITAL de SIGMA Informatique [5] Notre enquête cherche à identifier 1) si la perception de ces outils au travail est plutôt positive ou négative, 2) le degré actuel d'intention d'usage de ces outils au travail et 3) si des facteurs intrinsèques aux employés et à leur poste ont un effet sur cette intention d'usage. En complément du questionnaire, nous avons pu échanger avec quelques participants pour approfondir certains points explorés préalablement. Dans la section suivante, nous présenterons notre méthodologie d'enquête pour explorer ces perceptions à travers le prisme de l'acceptabilité.

9.2.1 Méthodologie

Recrutement des participants Nous avons diffusé le questionnaire à l'ensemble de la BU DIGITAL (40 collaborateurs internes) et avons eu 21 répondants. Parmi ces répondants, nous avons eu 15 répondants issus des équipes de développement, 3 répondants issus de la chefferie de projet et 3 répondants issus de l'équipe commerciale.

Matériel Nous avons développé un questionnaire pour explorer la perception des employés à l'égard des outils génératifs dans leur contexte de travail. L'objectif de ce questionnaire est d'étudier la perception de ces outils du point de vue de l'acceptabilité. Dans un premier temps, ce questionnaire nous a permis d'interroger les employés (voir l'annexe 5) sur

- Leur connaissance des outils génératifs,
- Les types d'usage s'ils s'en servent déjà et à quelle fréquence,
- Les solutions numériques de travail qu'ils utilisent et qui ne sont pas mis à disposition par l'organisation

5. La BU DIGITAL est un service spécialisé dans la prestation de conception de solutions numériques sur-mesure et de modernisation de systèmes d'information.

Dans un deuxième temps, nous avons choisi d'explorer l'acceptabilité de ces outils génératifs au travail au moyen du modèle TAM [43] (voir chapitre 2), auquel nous avons ajouté des dimensions issues du modèle UTAUT2 [171] : la motivation hédonique et l'influence sociale. Nous avons inclus ces deux dimensions de l'UTAUT2, en supposant que, en raison de l'émergence récente des outils génératifs, l'attrait pour l'innovation et l'environnement social peuvent être déterminants dans l'intention d'usage (voir Annexe 5).

Hypothèses Nous avons formulé trois hypothèses concernant les facteurs susceptibles d'influencer la perception des outils génératifs au sein de la BU DIGITAL. Les hypothèses sont énoncées comme suit :

- H1 : L'expérience dans le poste peut influencer la perception de l'outil. Plus une personne a d'expérience dans son travail, moins elle a besoin d'être assistée d'un outil génératif ;
- H2 : Le type de poste peut également avoir un impact sur la perception de l'outil. Selon les tâches et responsabilités associées à un poste, l'utilisation d'un outil génératif peut être perçue différemment ;
- H3 : Le degré d'importance accordée aux tâches pour lesquelles les outils génératifs sont utilisés peut influencer la perception de ces outils. Si une personne accorde beaucoup d'importance à ces tâches, elle peut être plus critique quant à l'utilisation d'un outil pour les accomplir.

9.2.2 Résultats et Discussion

D'après nos résultats, la recherche d'informations en ligne et la demande d'aide à un collègue sont les actions les plus courantes lorsque les employés de la BU DIGITAL rencontrent des difficultés (voir figure 9.1). Ces premiers chiffres montrent un assez faible intérêt et une faible connaissance de la documentation interne, censée aider les collaborateurs. Deux des répondants ont déclaré s'aider d'une autre manière et tous deux ont affirmé déjà utiliser les outils génératifs pour s'aider. Parmi les 21 répondants, 10 utilisent des solutions numériques qui ne sont pas fournies par l'organisation pour effectuer leurs tâches de prise de notes, de documentation et de gestion du temps. Parmi ces outils, certains utilisent des outils génératifs comme ChatGPT (qui semble être le plus connu par l'échantillon) et GitHub Copilot, déjà essayé par deux des répondants. Six répondants ont quand même déclaré déjà connaître d'autres outils génératifs comme Google BARD (nouvellement GEMINI) et DALL-E.

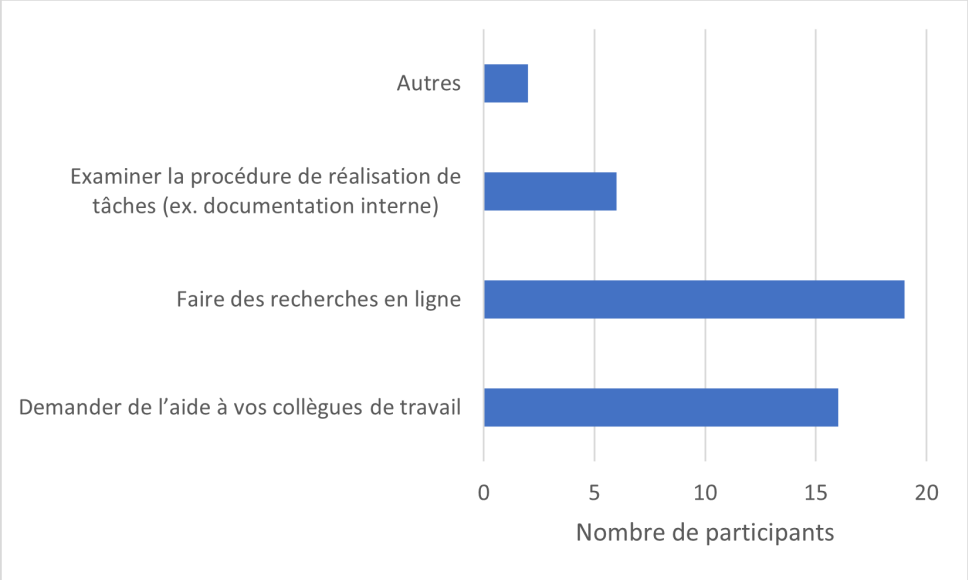


FIGURE 9.1 – Réponses à la question "Lorsque vous rencontrez des difficultés à effectuer vos tâches, vous préférez ?"

Lorsque nous avons demandé à quelle fréquence les collaborateurs utilisaient des outils génératifs, la majorité a déclaré ne jamais s'en servir, tandis que les autres participants les utilisent de manière occasionnelle. Plus de la moitié déclarent quand même s'intéresser aux outils génératifs et les considèrent potentiellement utiles pour leur travail, avec diverses utilisations telles que le soutien au développement (6 répondants), l'optimisation de la recherche d'informations (4 répondants), la rédaction d'emails (3 répondants) et les conseils préventes, de gestion et POC (*Proof of concept*) (3 répondants). Ces résultats suggèrent que les outils génératifs ont le potentiel d'améliorer l'efficacité et la productivité au travail et ils sont donc susceptibles d'être considérés comme acceptables par les utilisateurs.

Corrélation entre les dimensions étudiées Les résultats globaux des répondants quant à leur intention d'usage des outils génératifs dans leur travail sont disponibles dans la figure 9.2.

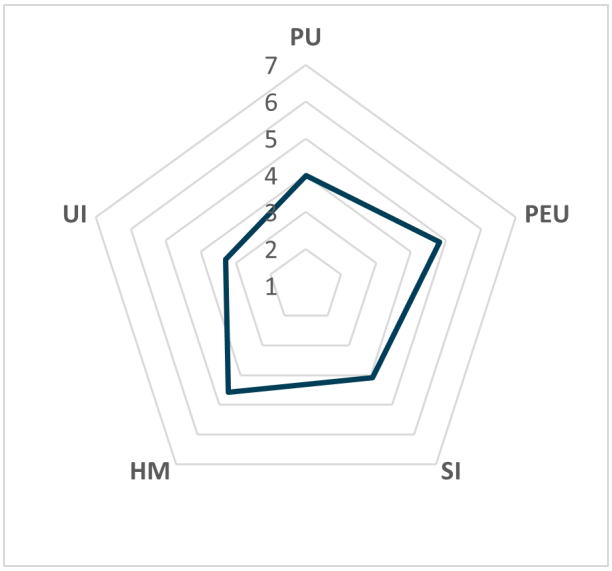


FIGURE 9.2 – Réponses des répondants à la partie du questionnaire qui s'intéresse à l'acceptabilité des outils génératifs dans leur situation de travail (Légende : PU - Utilité perçue, PEU - Facilité d'utilisation perçue, SI - Influence sociale, HM - Motivation hédonique, UI - Intention d'usage)

Á partir de ces résultats, nous avons étudié la corrélation entre les dimensions étudiées de l'intention d'usage des outils génératifs à l'aide de la corrélation de Pearson. Les résultats montrent que toutes les dimensions étudiées sont modérément à hautement corrélées entre elles, avec des corrélations allant principalement de 0,5 à 0,846, à l'exception de l'influence sociale qui a les scores de corrélation les plus bas (r entre 0,263 et 0,539) avec toutes les autres dimensions. Un corrélogramme est présenté dans la figure 9.3.

Ces résultats suggèrent que les différentes dimensions étudiées sont interconnectées et donc susceptibles d'avoir un impact les unes sur les autres. Y compris pour l'influence sociale, bien que l'effet d'interaction de cette dimension avec l'intention d'usage des outils génératifs soit légèrement plus faible ($r = 0.43$). Ceci peut indiquer que les facteurs sociaux jouent un rôle différent ou distinct dans la perception des outils LLM au travail par rapport aux autres dimensions étudiées.

Perception des outils génératifs selon le profil des répondants Afin d'explorer davantage la perception des outils génératifs, nous avons considéré notre échantillon sous différents angles. L'objectif était d'identifier s'il existe des différences dans les perceptions selon des facteurs objectifs tels que l'expérience professionnelle (voir figure 9.4), le type de poste, la fréquence d'utilisation des outils génératifs et le degré d'importance accordée aux

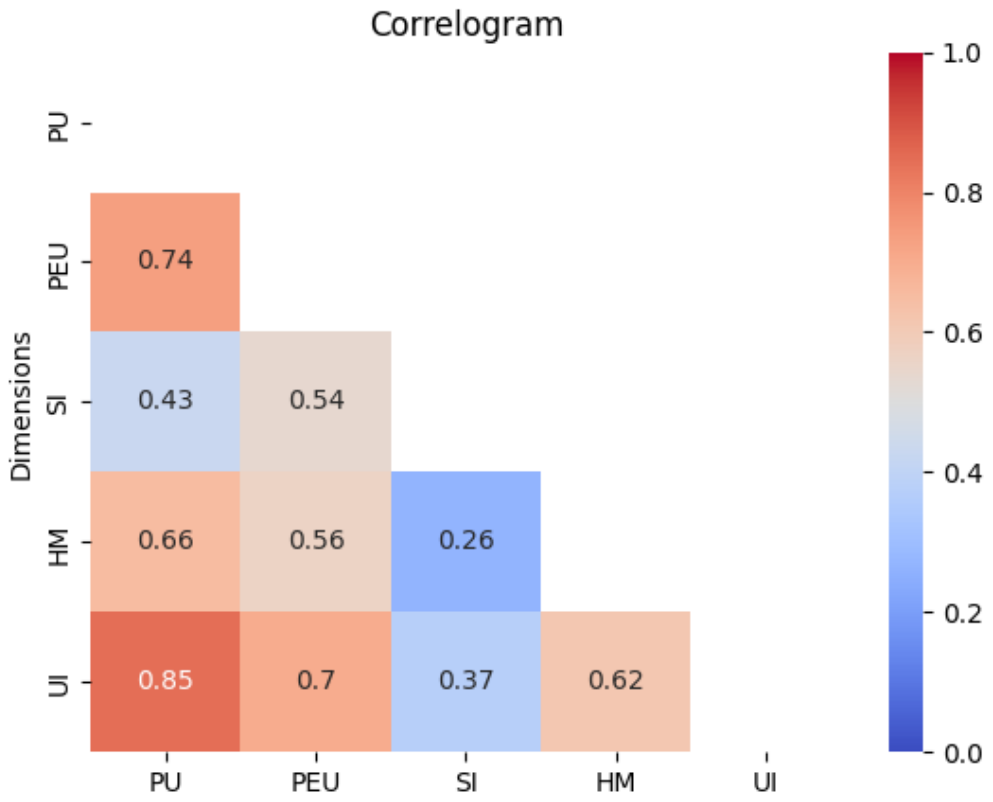


FIGURE 9.3 – Coefficients de corrélation de Pearson entre les dimensions étudiées dans la partie acceptabilité du questionnaire (Légende : plus la valeur r tend vers le rouge plus la corrélation entre les deux dimensions est élevée, à l'inverse plus la valeur r tend vers le bleu plus la corrélation entre les deux dimensions est faible)

tâches pour lesquelles les participants seraient prêts à utiliser des outils génératifs.

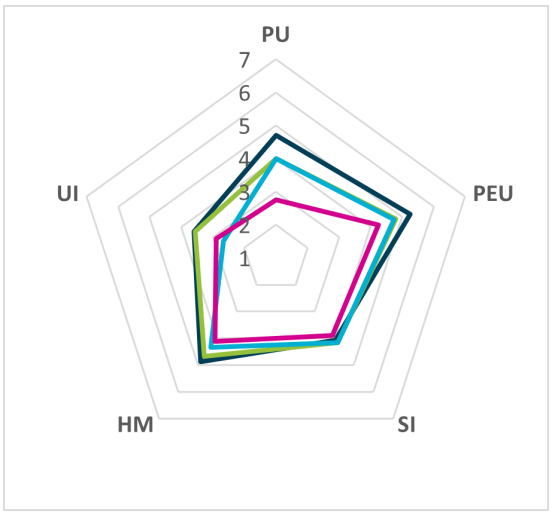
Il est intéressant, au vu de la figure 9.4, de noter certaines tendances dans les scores présentés. Par exemple, plus les répondants ont de l'expérience dans leur travail, moins ils perçoivent les outils de génération comme utiles. Ou encore, les répondants qui utilisent des outils génératifs pour des tâches à haute valeur ajoutée perçoivent ces outils comme utiles et faciles à utiliser. Cependant, l'influence sociale semble varier très peu entre les différentes conditions. En discutant avec certains des répondants, ils déclarent quand même ne pas être à l'aise avec l'idée que leur responsable hiérarchique soit au courant qu'ils utilisent des outils génératifs dans leur travail

Pour approfondir notre étude de la perception des outils génératifs dans le contexte du travail en fonction des types de profils, nous avons établi trois hypothèses :

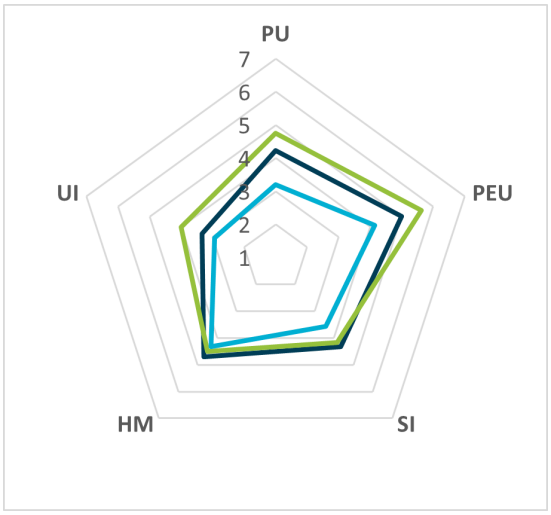
— H1 : L'expérience dans le poste a un effet sur la perception des outils génératifs au travail.
— H2 : Le type de poste a un effet sur la perception des outils génératifs au travail.
— H3 : Le degré d'importance accordée aux tâches pour lesquelles les répondants seraient prêts à utiliser les outils génératifs a un effet sur la perception des outils génératifs au travail.

Pour tester nos hypothèses, nous avons choisi d'utiliser le test de Kruskal-Wallis, qui est un test non paramétrique utilisé pour comparer les distributions de plusieurs groupes indépendants [6].

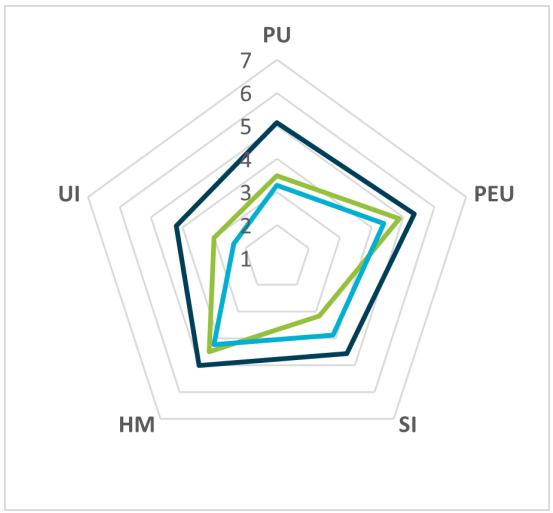
6. Pour l'hypothèse 1, nous comparons les perceptions des outils génératifs entre différents niveaux d'expérience professionnelle. Pour l'hypothèse 2, nous comparons les perceptions des outils génératifs entre les métiers. Et pour l'hypothèse 3, nous comparons les perceptions des outils génératifs en fonction de l'importance des tâches ciblées



(a) Selon l'expérience en poste (Légende : bleu foncé - moins de 2 ans d'expérience, vert - entre 2 et 5 ans d'expérience, bleu clair - entre 6 et 10 ans d'expérience, violet plus de 10 ans d'expérience)



(b) Selon le profil métier (Légende : bleu foncé - développeurs, vert - Commerce, bleu clair - chefferie de projet)



(c) Selon l'importance accordée aux tâches pour lesquelles les participants seraient prêts à utiliser des outils génératifs (Légende : bleu foncé - haut niveau d'importance, vert - moyen niveau d'importance, bleu clair - faible niveau d'importance)

FIGURE 9.4 – Résultats moyens à la deuxième partie du questionnaire en fonction des hypothèses (Légende : PU - Utilité perçue, PEU - Facilité d'utilisation perçue, SI - Influence sociale, HM - Motivation hédonique, UI - Intention d'usage)

Dimensions	**H1**	**H2**	**H3**
	H (*p value*)	*H* (*p value*)	*H* (*p value*)
Utilité perçue	5.365 (*.147*)	3.032 (*.220*)	11.79 (*.067*)
Facilité d'utilisation perçue	1.755 (*.625*)	4.710 (*.095*)	6.84 (*.336*)
Influence sociale	.310 (*.958*)	.597 (*.742*)	9.15 (*.166*)
Motivation hédonique	3.990 (*.263*	.835 (*.659*)	9.15 (*.166*)
Intention d'usage	3.925 (*.270*)	2.391 (*.303*)	10.44 (*.107*)

TABLE 9.1 – Résultats des tests de Kruskall-Wallis pour estimer si les facteurs d'acceptabilité des outils génératifs sont affectés par 1) l'expérience en poste (H1), 2) le type de poste (H2) et 3) Degré d'importance accordée aux tâches pour lesquels les outils génératifs seraient utilisés (H3).

Au seuil de significativité statistique de 5%, les résultats des tests de Kruskall-Wallis (voir tableau 9.1) indiquent que l'expérience professionnelle (H1), le type de poste (H2) et le degré d'importance accordée aux tâches pour lesquelles les répondants utiliseraient les outils génératifs (H3) n'ont pas d'influence sur l'acceptabilité de ces outils au travail d'après les données récoltées.

Cependant, nous préférons considérer ces résultats statistiques avec prudence dû :

— Au faible échantillon (21 répondants sur 40 collaborateurs dans le service) ;

— Aux entretiens réalisés dans lesquels, par exemple, des répondants avec plus de 10 ans d'expérience se disaient très à l'aise à l'idée d'admettre utiliser des outils génératifs au travail alors que les répondants avec moins de 2 ans d'expérience avaient plutôt tendance à dire qu'ils craignaient que la hiérarchie ne reconnaisse pas leurs compétences s'ils admettaient utiliser ces outils. Cependant, d'un point de vue statistique, il n'y a pas de différence significative.

Cette étude met tout de même en lumière la curiosité des collaborateurs de la BU DIGITAL quant à l'utilisation des outils génératifs au travail. Et bien que les représentations semblent éparses, nous voyons très peu de refus catégoriques à les essayer, mais plus une crainte de manque d'efficacité ou une crainte liée à l'attribution de la responsabilité de la décision finale. Par conséquent, il semblerait intéressant de créer un socle commun de connaissances pour faire parvenir aux collaborateurs ce que les outils génératifs dans leur travail peuvent réellement leur apporter ou non, et comment s'en servir. De plus, il pourrait être utile de réitérer cette étude, pour voir l'évolution de ces perceptions et de

l'acceptabilité de ces outils à l'année N+2, pour laquelle ces outils seront susceptibles d'être encore plus présents dans le monde professionnel.

9.3 Quel impact des outils génératifs sur la rédaction de code informatique par des développeurs Java ?

Á l'issue de notre enquête réalisée dans la section 9.2, nous avons décidé d'explorer plus spécifiquement l'impact de l'utilisation d'un outil génératif sur la réalisation de tâches spécifiques. Nous avons donc réalisé une étude comparative de production de code Java par les développeurs de la BU DIGITAL de SIGMA Informatique. Pour cette étude, nous leur demandons de commencer une tâche de programmation sans assistance de LLM puis de poursuivre avec l'assistance de ChatGPT de OpenAI [136], référence parmi les outils génératifs. L'objectif de cette étude est de comprendre s'il y a un impact de l'utilisation de ChatGPT sur la productivité et l'acceptabilité des développeurs Java dans un environnement d'entreprise. En évaluant l'efficacité de ChatGPT comme outil d'assistance dans la programmation, nous pourrions mieux orienter les stratégies d'implémentation de l'IA dans les pratiques de développement logiciel et contribuer à l'élaboration de stratégies de déploiement de solutions IA plus centrées sur l'utilisateur.

9.3.1 Méthodologie

Recrutement des participants Dans un premier temps, nous avons diffusé un appel à volontaire au sein de la BU DIGITAL pour participer à un exercice de développement et tester un outil génératif. Les critères de sélection étaient d'être développeur Java et d'avoir une heure de disponibilité. Six volontaires, qui répondaient aux critères, se sont présentés. Les participants provenaient de trois équipes projet différentes et avaient les niveaux d'expérience suivants : deux développeurs juniors, deux développeurs intermédiaires, et deux développeurs seniors.

Déroulé et tâche L'expérimentation était structurée en deux parties (voir figure 9.5 :

1. Les développeurs étaient invités à travailler sur un programme Java spécifique, en utilisant les méthodes "traditionnelles" d'aide au développement (ex. forums d'aide en ligne et recherches internet) pour résoudre les problèmes rencontrés.

2. Après une première session, les mêmes développeurs poursuivent le développement du programme pendant 30 minutes en utilisant ChatGPT pour obtenir de l'aide sur les mêmes tâches.

Puis les participants remplissaient un questionnaire et enfin, nous procédions à un échange libre avec eux.

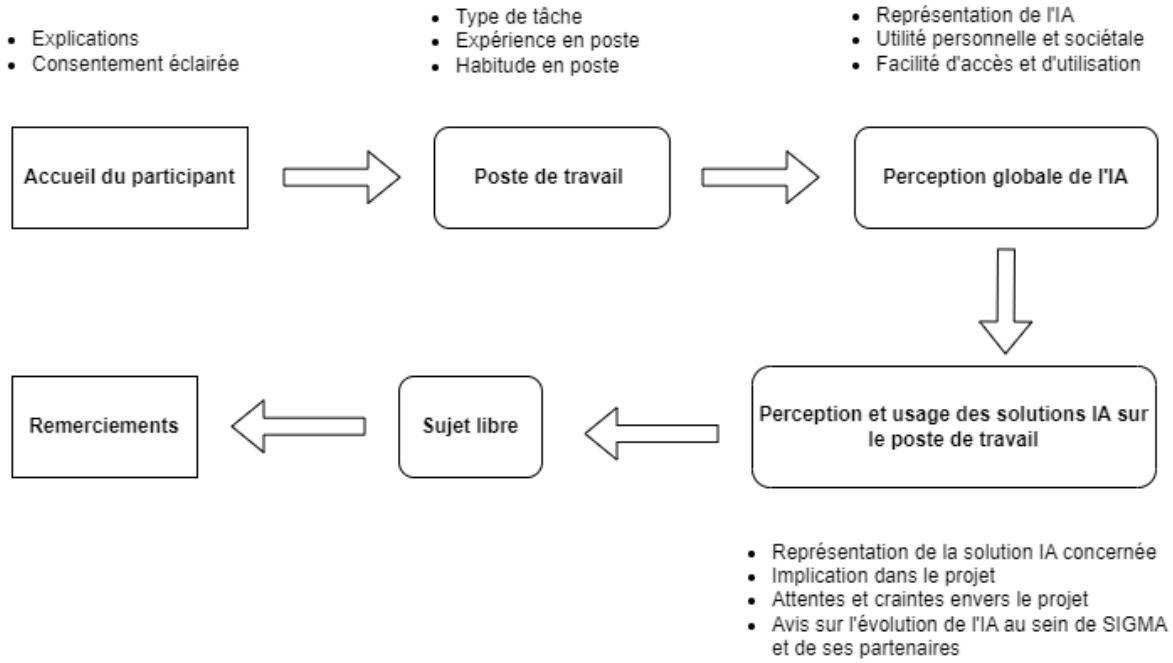


FIGURE 9.5 – Déroulé de l'expérimentation

Pour réaliser la tâche, les participants étaient devant un écran, où étaient affichés un environnement de développement (IDE) et un jeu de données dans un tableur. Ce jeu de données s'intéresse aux ventes de produits électroniques d'un revendeur pour le mois d'avril 2019. Il contient l'identifiant des commandes réalisées, le nom des articles, le type d'article, les quantités commandées, le prix unitaire des articles et la date des commandes. Á partir de ce jeu de données, il est demandé aux participants de créer un script en Java qui permet de faire ressortir les informations suivantes :

— La quantité d'articles commandés par type de produit la première semaine du mois d'avril 2019 (du 01/04/2019 au 07/04/2019 inclus)

— La quantité de smartphones commandés dont le coût est supérieur à 600 €.

Les contraintes pour cette tâche étaient que les participants ne disposez que de 20 minutes pour effectuer l'exercice. De plus, le script rédigé doit être aussi lisible que possible et les participants pouvaient utiliser les outils qu'ils souhaitaient pour réaliser la tâche (ex. naviguer sur le web), en dehors des outils génératifs type ChatGPT ou GitHub Copilot.

Dans la deuxième session, les participants bénéficiaient de 10 minutes pour finaliser le code qu'ils avaient commençaient à rédiger. Pour cela, nous leur demandions d'interagir avec ChatGPT, de la manière qu'ils souhaitaient, pour finaliser la tâche.

Collecte de données Pour analyser les usages des participants, nous avons récolté les données suivantes :

- L'efficacité du programme - chaque programme était testé pour vérifier s'il fonctionnait correctement à la fin de chaque phase.
- La nature des prompts adressés à ChatGPT - classées selon leur type pour analyser les interactions avec la solution IA.
- Le degré d'acceptabilité de ChatGPT par les participants par rapport à leurs propres compétences - mesuré avec le questionnaire TAM de Davis (voir les chapitres 2 et 7) [43]. Nous y avons ajouté la dimension de confiance en ses propres compétences pour identifier à quel point les développeurs s'estiment capables d'accomplir la tâche prescrite par eux-mêmes. Nous avons également ajouté une dimension de fiabilité perçue en ChatGPT pour mesurer à quel point les développeurs percevaient ChatGPT comme capable d'accomplir correctement la tâche prescrite (voir chapitre 7) [104].

9.3.2 Résultats et Discussion

Cette étude nous a permis de recueillir des informations détaillées sur l'impact de l'utilisation de ChatGPT dans la production de code Java par des développeurs de différents niveaux d'expérience. Tout d'abord, un développeur intermédiaire et un développeur expert ont réussi à produire un programme fonctionnel à l'issue de la deuxième phase. Cela suggère qu'en un temps très limité, l'utilisation de ChatGPT peut ne pas suffire à compenser l'écart de compétences entre les développeurs de différents niveaux d'expérience.

Ensuite, nous avons classé les prompts des développeurs (entre 1 et 4 prompts par participant), adressés à ChatGPT, en trois catégories principales : génération de script,

demande d'informations sur la syntaxe Java, et optimisation et/ou complétion d'un script existant. La répartition de ces types de prompt par niveau d'expérience est présentée dans le tableau 9.2.

Type de prompt	**Niveau d'expérience**		
	Expert	**Intermédiaire**	**Novice**
Génération de script	0	3	2
Demande d'informations sur la syntaxe Java	2	1	0
Optimisation et/ou complétion d'un script existant	1	3	2

TABLE 9.2 – Classification des prompts par niveau d'expérience

D'après notre analyse, les développeurs de niveau intermédiaire sont ceux qui ont adressé le plus de prompts de génération de bout de code (fonction, boucle, etc.). Ces résultats peuvent provenir d'une volonté de gagner du temps en générant une base fonctionnelle, en tirant parti des capacités de ChatGPT pour créer un nouveau code. Sur ce même critère, les développeurs novices arrivent en seconde position, sûrement afin d'obtenir de l'aide pour démarrer le développement de l'application. Nous ne constatons aucune demande de la part des experts, ce qui peut renvoyer à une confiance en leur propre capacité à générer un code en partant de zéro. Les développeurs experts sont pourtant ceux qui ont adressé le plus de prompts de demande d'informations, de clarification, quant à la syntaxe Java. Nous posons l'hypothèse que c'est pour approfondir leur connaissance du langage, en allant chercher des détails syntaxiques spécifiques, qui pourraient leur être cruciaux. Et enfin, les développeurs novices et intermédiaires ont respectivement utilisé plus de prompts pour optimiser/compléter leurs scripts. Cela montre une tendance à utiliser ChatGPT comme un outil pour affiner et compléter leur travail, là où il peut y avoir des points bloquants ou des incertitudes dans la rédaction du script.

Dans un deuxième temps, nous nous sommes intéressés aux résultats du questionnaire selon nos cinq dimensions : l'utilité perçue, la facilité d'utilisation perçue, la fiabilité perçue et la confiance en ses propres capacités et l'intention d'usage (voir tableau 9.3 et la figure 9.6).

Nos résultats montraient que ChatGPT était globalement perçu comme modérément utile, mais plutôt simple d'utilisation, bien que son adoption au sein de SIGMA Informatique reste limitée, voire déconseillée par les responsables hiérarchiques. Ici, les développeurs

Dimension du questionnaire	**Niveau d'expérience**		
	Expert	**Intermédiaire**	**Novice**
IU (Intention d'usage)	2.43	4.14	7.00
PEU (Facilité d'utilisation perçue)	5.00	6.25	7.00
PU (Utilité perçue)	5.00	5.00	7.00
RC (Fiabilité perçue)	4.64	4.71	2.71
SC (Confiance en ses propres capacités)	5.50	5.33	5.00

TABLE 9.3 – Score moyen des réponses des participants au questionnaire par dimension et par niveau d'expérience (sur une échelle de Likert allant de 1- Pas du tout à 7- Totalement)

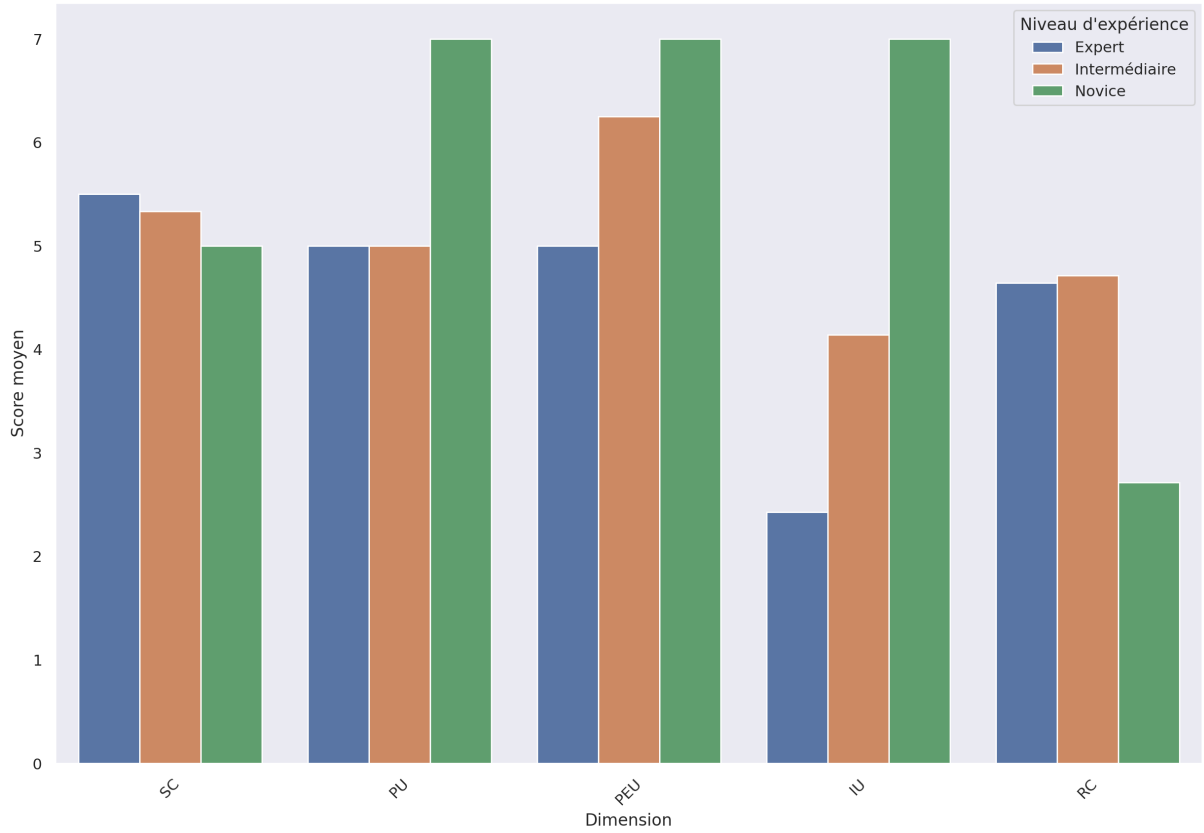


FIGURE 9.6 – Score moyen des réponses des participants au questionnaire par dimension et par niveau d'expérience (sur une échelle de Likert allant de 1- Pas du tout à 7- Totalement)

novices possédaient l'intention d'usage de ChatGPT la plus élevée et percevaient l'outil comme plus facile à utiliser et plus utile par rapport aux développeurs plus expérimentés. Cependant, ces développeurs novices considéraient également ChatGPT comme moins fiable par rapport aux autres profils. Nous supposons que c'est dû :

— soit à une moindre compréhension des capacités et des limites de l'outil.

— Soit à des prompts envoyés de nature spécifique (ex. génération de plus gros blocs, pour lesquels les erreurs de ChatGPT semblent plus probables)

Finalement c'est en échangeant avec les participants après l'expérimentation que nous en apprenons le plus. Tous les participants considéraient que ce type d'outil a le potentiel d'être utile pour produire du code plus rapidement. Mais les profils experts ne ressentaient pas le besoin de s'en servir, car ils avaient déjà leurs habitudes de travail qui leur convenaient. De plus, les profils intermédiaires et experts nous expliquaient également qu'en voyant le code produit par ChatGPT, ils étaient d'autant plus récalcitrants à s'en servir parce qu'ils constataient des erreurs qui, selon eux, n'ont pas lieu d'être. Et donc que s'ils intégraient ce type d'outil à leur pratique, ils devraient à chaque fois prendre le temps de vérifier ce que l'outil a généré, donc ils préféraient encore tout produire eux-mêmes.

Il est important de noter que cette étude a été réalisée avec un échantillon très restreint, qui n'est peut-être pas représentatif de l'acceptabilité de tous les développeurs Java de la BU DIGITAL, ni en général. Mais cette étude était l'occasion d'obtenir une première vision de comment est perçue ChatGPT dans un service de développement informatique et du degré d'acceptabilité de ce dernier dans une pratique professionnelle.

9.4 Conclusion

Les deux études présentées dans ce chapitre nous montrent une certaine curiosité pour les outils génératifs parmi les collaborateurs de la BU DIGITAL de SIGMA Informatique. L'utilisation de ces outils, notamment ChatGPT, est perçue comme potentiellement bénéfique, mais leur adoption reste limitée, probablement en raison de préoccupations liées à l'efficacité et à l'acceptabilité au sein de l'entreprise.

Bien que les outils génératifs sont très en vogue pour leurs capacités à générer du code informatique, les résultats de notre première enquête montrent qu'au sein du service, début 2023, les développeurs n'étaient pas forcément le public le plus intéressé par ces outils. Pourtant, nous avons pu constater que ceux-ci peuvent être utiles pour générer rapidement une base fonctionnelle d'application pour gagner du temps ou encore vérifier et optimiser du code déjà écrit pour améliorer la qualité de ce même code. Cette seconde étude, à prendre avec précaution au vu du faible échantillon, nous oriente tout de même sur le fait que plus l'utilisateur a d'expérience sur son poste, moins il trouve les outils génératifs utiles.

Les résultats de ces enquêtes semblent finalement corroborer notre état de l'art, indiquant que les professionnels du développement informatique montrent un intérêt pour l'introduction d'outils génératifs dans leurs pratiques, tout en exprimant certaines réserves. Tout comme les études de Imai [85], de Dakhel et al. [42] et de Ziegler et al. [192], nos recherches suggèrent que les développeurs estiment qu'utiliser des outils génératifs peut leur permettre de gagner en performance. Nous mettons également l'accent sur les développements juniors qui nous semblent les plus enclins à utiliser ces outils dans leur pratique. Tout comme notre état de l'art et au regard de notre seconde étude sur le sujet, nous avertissons également quant à une hausse du temps de vérification et de correction du code généré [85]. Néanmoins, nous nous accordons tous sur le fait que les outils génératifs ont le potentiel d'offrir un réel confort d'assistance.

Á l'issu de ces travaux, nous estimons que les outils génératifs étaient, début 2023, encore trop méconnus du public professionnel au sein de SIGMA Informatique et que les résultats obtenus étaient sûrement influencés par le choix de l'outil génératif : ChatGPT. Bien qu'étant une référence parmi les outils génératifs, ChatGPT ne nous semble à présent pas avoir été l'outil le plus adapté aux besoins spécifiques des développeurs de la BU DIGITAL. Une méthodologie de sélection de LLM plus rigoureuse et en adéquation avec les besoins, contraintes et capacités des utilisateurs cibles est nécessaire. Il faudrait décider d'un socle méthodologique, qui soit aussi approuvé et porté par l'organisation pour assurer une aisance d'usage et une intégration fluide des outils génératifs dans les pratiques quotidiennes des collaborateurs de SIGMA Informatique. L'acceptabilité des outils génératifs nécessite un soutien institutionnel pour que les développeurs se sentent à l'aise et valorisés dans l'utilisation de ces derniers. Ce point sera donc approfondi dans le prochain et dernier chapitre, où nous examinerons comment mettre en place une méthodologie centrée utilisateur dans la conception d'un assistant de développement, utilisant un LLM. Nous pensons que cette stratégie valorisera l'adoption des outils génératifs dans les processus de développement logiciel.

Chapitre 10

Méthodologie de sélection de LLM pour mettre en place un assistant de développement Python

Dans ce chapitre

Ce dernier chapitre traite de la mise en application d'une méthodologie de sélection de LLM dans la conception d'un assistant de développement au sein de SIGMA Informatique. Nous proposons une méthodologie de sélection de LLM qui répond aux enjeux de sécurité des données de la plupart des entreprises qui réfléchissent à introduire des outils génératifs dans leurs postes de travail. Cette méthodologie se base sur les préférences des utilisateurs finaux en mobilisant des méthodes utilisées pour évaluer subjectivement la QoE dans le domaine multimédia. Nous nous concentrons donc sur des LLM open sources, qui offrent de la flexibilité et qui permettrait aux entreprises de mieux contrôler l'impact écologique que peuvent avoir ce type d'outil en mobilisant des LLM, certes performants, mais aussi potentiellement plus frugaux.

10.1 Introduction

Face au nombre croissant de LLM disponibles, chacun ayant ses spécificités, il devient de plus en plus important de définir une méthodologie efficace pour sélectionner les LLM les plus pertinents en fonction du contexte d'usage et des utilisateurs-cible [29] [175] [1]. Début 2023, nous avons réalisé une enquête via questionnaire au sein de la BU DIGITAL

de SIGMA Informatique afin de comprendre comment les collaborateurs percevaient les outils qui utilisent des LLM pour générer du contenu [3]. Á cette période, il n'y avait encore que très peu d'outils génératifs grand public qui étaient connus, à part ChatGPT de OpenAI [136]. En 2024, à l'année N+1, une multitude d'outils génératifs ont fait leur apparition et proposent leurs services au monde professionnel comme Microsoft Copilot [122], Gemini [193] ou encore Mistral AI [4].

Nous avons été particulièrement attentifs aux bouleversements qu'ont amenés les outils génératifs sur les postes de travail (voir chapitre 8). Traditionnellement, les entreprises prenaient en charge la sélection des outils informatiques de travail pour les collaborateurs pour des raisons de sécurité, de gestion des coûts, d'uniformité, de formation, de contrôle des processus de travail ou encore de conformité réglementaire. Mais il semble que le caractère novateur et l'accessibilité croissante des outils génératifs au grand public ont fortement contribué à ce que les collaborateurs essaient, expérimentent et intègrent ce type d'outils sur leur poste. Cependant, cette approche accroît les risques d'un point de vue sécuritaire, opérationnel et réglementaire [124]. Les risques les plus importants que nous avons pu identifier sont en termes de sécurité des données. On y retrouve les risques de :

— Fuite - en fonction du média utilisé pour interagir avec le LLM (ex. plug-in de navigateur, site web non officiel, ou encore applications tierces), le collaborateur n'identifie pas forcément toutes les vulnérabilités de ce dernier. Le média utilisé peut parfois être lié ou soumis à des attaques malveillantes, des logiciels malveillants ou encore à une configuration non sécurisée.

— Stockage sur des serveurs externes - interagir avec le LLM via des services Web ou des modèles propriétaires (via API) nécessite d'envoyer des informations vers des services externes. Ces services peuvent ainsi stocker des fragments de données entrées dans leur système. Si les LLM sont utilisés avec des données professionnelles du collaborateur, cela pourrait conduire à une rétention de données sensibles ou confidentielles sur leurs serveurs. Ce risque est d'autant plus présent si les collaborateurs ne sont pas pleinement conscients ou formés sur les implications de l'utilisation de LLM en termes de confidentialité des données. Par exemple, ils pourraient involontairement communiquer des données sensibles ou confidentielles de l'entreprise mais aussi des clients.

Voyant se multiplier les usages d'outils génératifs sur les postes de travail et les risques associés, SIGMA Informatique s'est positionnée sur l'utilisation des outils génératifs au

travail. L'entreprise valorise le fait de voir les collaborateurs expérimenter, tester et découvrir des usages pertinents pour leurs métiers. Mais elle met également en garde les collaborateurs face aux services dont les conditions d'utilisation ne garantissent pas la confidentialité des données, parfois de manière explicite. Les données des collaborateurs pourraient être reprises pour ré-entraîner les modèles et diffusées dans les réponses apportées à l'ensemble des utilisateurs. Plus précisément, si les collaborateurs communiquent des données sensibles aux LLM, il est possible que ces données soient obtenues et réutilisées par d'autres utilisateurs. Dit autrement, si des informations de client/sensibles sont chargées, elles pourraient être fournies ultérieurement à tout utilisateur qui en ferait la demande.

C'est pourquoi SIGMA Informatique a choisi de restreindre l'accès aux outils génératifs accessible via des sites web, comme ChatGPT (le service le plus utilisé actuellement) en présentant à tout collaborateur, souhaitant l'utiliser, un message d'avertissement lors de l'accès au site. Ce qui est ici mis en avant par l'organisation est l'ambition de sensibiliser les collaborateurs à un usage respectueux des données. En ce sens, il est fortement recommandé aux collaborateurs souhaitant utiliser un LLM pour leur travail d'adopter Microsoft Copilot. Cet outil est considéré par l'organisation comme l'outil génératif le plus sécurisé. Sa licence professionnelle garantit que les données envoyées dans les prompts ne sont pas exploitées [122].

Cependant, en échangeant avec certains collaborateurs, nous nous apercevons que Microsoft Copilot, qui utilise le LLM Prometheus (lui-même issu de GPT-4), n'est pas toujours l'outil génératif le plus apprécié et/ou adapté pour les tâches des collaborateurs. Comme nous l'avons vu dans le chapitre 8, la performance des LLM est estimé selon plusieurs métriques pour lesquelles Prometheus n'a pas forcément les plus élevées partout. Nous questionnons donc la décision d'introduire cet outil pour tout type d'usage, tant que les besoins et usages des différents profils de collaborateurs n'ont pas été précisément évalués. Parfois décrit comme "moins permissif que ChatGPT", ou ayant "tendance à faire plus d'erreurs que l'outil de Google ou encore Mistral" [1], Microsoft Copilot pourrait ne pas répondre aux attentes spécifiques de tous les collaborateurs.

Dans ce contexte, nous souhaitons proposer une méthodologie de sélection de LLM

1. Ce sont des verbatims issus de nos échanges avec les collaborateurs de SIGMA Informatique suite à la recommandation d'utiliser Microsoft Copilot pour tout type de tâche.

qui permet de faire face aux risques décrits précédemment, tout en incitant à justifier sur le choix de modèle sur les préférences des utilisateurs finaux en fonction des tâches définies. Notre méthodologie de sélection de socles techniques s'intègre à une démarche centrée utilisateur car elle valorise la prise en compte des préférences et besoins des utilisateurs finaux (les collaborateurs) dans la conception, en s'appuyant sur une démarche de comparaison par pairs. Nous l'appliquons ici dans un projet de conception d'un outil génératif, qui assistera les développeurs dans la production de code Python.

10.2 Méthodologie de sélection de LLM

En accord avec la volonté de SIGMA Informatique d'amener les collaborateurs à utiliser des outils génératifs dans un cadre sécuritaire, nous proposons d'accorder un peu de flexibilité à la sélection des LLM utilisables. Notamment avec des modèles qui respecteraient la ligne directrice des équipes de conception en termes de protection des données. Nous proposons donc de s'intéresser à des LLM *open sources* pour concevoir des outils génératifs sur-mesure qui seront modifiables, transparents, coopératifs et qui permettent de réduire les coûts liés aux licences logiciels. Parmi les intérêts liés à l'utilisation de LLM *open sources*, il y a notamment le fait que :

- Les licences de ces modèles peuvent être commerciales, permettant de répondre à des enjeux business (mise à disposition des clients) mais aussi de s'assurer de répondre à des besoins de conformité avec le cadre législatif [2] (ex. les modèles Code Llama, Mistral 7b ou encore Falcon).
- Le stockage des données entrantes peut se faire localement (sur les serveurs internes) uniquement et non sur des serveurs externes. Cette approche permet de valoriser un contrôle total de la donnée (les concepteurs de l'outil sont le dernier maillon de la chaîne de diffusion de données lors de l'utilisation de l'outil). Cela permet également de mettre en place des mesures de sécurité robustes et personnalisables selon les contraintes internes et les besoins clients. Mais aussi de s'assurer de rester en conformité avec le Règlement Général sur la Protection des Données (RGPD) en termes de protection et d'exploitation des données.
- La gestion de la performance et de la latence du LLM sont gérées par SIGMA Informatique, qui peut choisir le nombre de ressources allouées pour le bon fonctionnement du LLM

2. Confidentialité, sécurité de la donnée, support client, exploitation commerciale

— Une indépendance vis-à-vis de fournisseurs d'outils génératifs, susceptibles d'exploiter les informations reçues dans les prompts envoyés au LLM.

Bien que nous réduisions l'éventail de LLM compatibles avec le contexte de SIGMA Informatique [3], tous les modèles de cette catégorie ne se valent pas. Plutôt que de se baser uniquement sur des indicateurs d'évaluation automatisée (voir section 8.3), nous avons combiné cette approche à une évaluation centrée utilisateur pour garder les utilisateurs finaux dans les choix de conception. Nous proposons donc d'utiliser la méthode de comparaison par paires qui est une approche permettant d'évaluer et comparer des concepts en les présentant deux par deux à des évaluateurs. Nous utilisons cette méthode pour évaluer quel LLM est préféré par les utilisateurs finaux en fonction du cas d'usage.

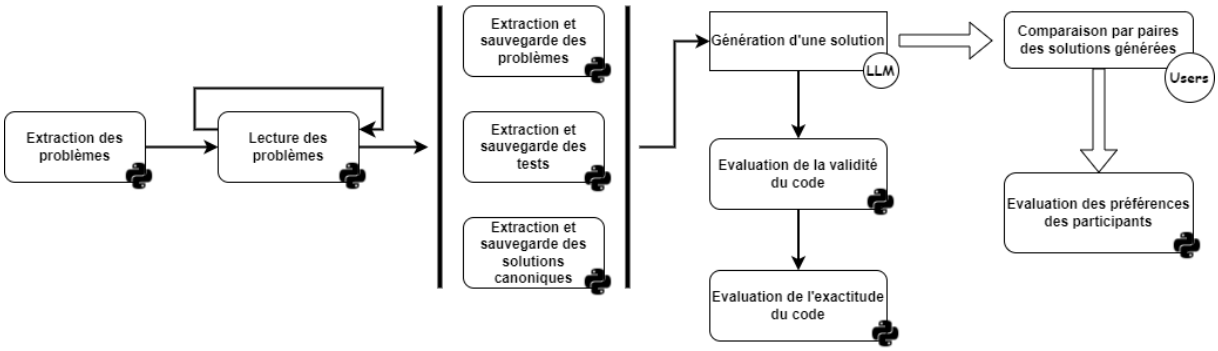


FIGURE 10.1 – Architecture d'évaluation inspirée des travaux de Yetistiren et al. [184]

Pour cette étude, nous avons choisi de nous baser sur trois modèles de langage : GPT-4 de OpenAI comme modèle de référence et deux modèles open source : Mistral 7b de Mistral AI et Code Llama 7b de Meta [4]. Les modèles open source ont été sélectionnés en fonction de leur ratio performance au jeu de données d'évaluation MBPP [15] par rapport au nombre de paramètres du modèle (voir tableau 8.1). Cette approche nous permet de comparer des modèles ayant différentes tailles et capacités tout en évaluant leur efficacité en termes de performance relative à leur complexité.

10.3 Méthodologie expérimentale

Matériel Nous avons commencé par créer les éléments à comparer par les participants. Pour cela, nous avons commencé par tirer aléatoirement 20 problèmes aléatoirement du

3. Les LLM pré-sélectionnés doivent être open sources, stockés en local et à licence commerciale
4. 7b signifie que le LLM en question a été entraîné avec 7 milliards de paramètres

jeu de données MBPP [5] [15]. Cette sélection aléatoire garantit une diversité de problèmes, couvrant différentes difficultés et types de tâches de programmation. Puis nous avons fait générer, pour chaque problème extrait de MBPP, une solution par chacun des LLM sélectionnés (GPT-4, Code LLama et Mistral 7b). Chaque LLM a donc produit une solution pour chaque problème (voir figure 10.1). Les codes sont générés par les LLM en langage Python parce qu'il s'agit du langage sur lequel les LLM sont le plus entraînés à générer du code et qu'il est le langage natif des jeux de données MBPP et HumanEval.

Participants L'étude a impliqué 17 développeurs Python, répartis en trois groupes d'expérience :

— 5 développeurs juniors (moins de 2 ans d'expérience) ;

— 10 développeurs intermédiaires (2 à 6 ans d'expérience) ;

— 2 développeurs seniors (plus de 6 ans d'expérience).

Nous avons recruté ces participants en ligne, et leur avons communiqué un lien internet pour passer l'expérimentation à distance via la plateforme de questionnaire en ligne SphinxOnline.

Tâche expérimentale Nous avons adopté une approche de comparaison par paire pour évaluer quel modèle peut être préféré par les développeurs. Pour chaque problème extraits de MBPP, nous avons créé des paires de solutions générées par les modèles (par exemple, GPT-4 vs Mistral 7b, Code Llama 7b vs Mistral 7b, GPT-4 vs Code Llama 7b). Chaque paire de solutions a été anonymisée pour éviter tout biais de la part des évaluateurs. Pour chaque paire de solutions, les développeurs doivent sélectionner la réponse qui leur semble la plus correcte, qualitative et applicable (voir figure 10.2).

10.3.1 Analyse

Pour analyser les données, nous avons déterminé les votes cumulés pour chaque comparaison de modèles (voir figure 10.4) afin de comparer les préférences des participants. Nous avons utilisé les tests statistiques suivants pour effectuer les comparaisons par paire (voir figure 10.3) :

— Test binomial, pour évaluer les préférences de chaque participant.

5. Ce jeu de données permet d'estimer les capacités des LLM à résoudre des problèmes de programmation en Python

FIGURE 10.2 – Exemple de comparaison de codes générés par deux des LLM (Pour sélectionner le code qu'il préfère, le participant clique dessus)

— Des comparaisons d'accords inter-répondants (pour plus de détails sur ces indicateurs, voir chapitre 5).

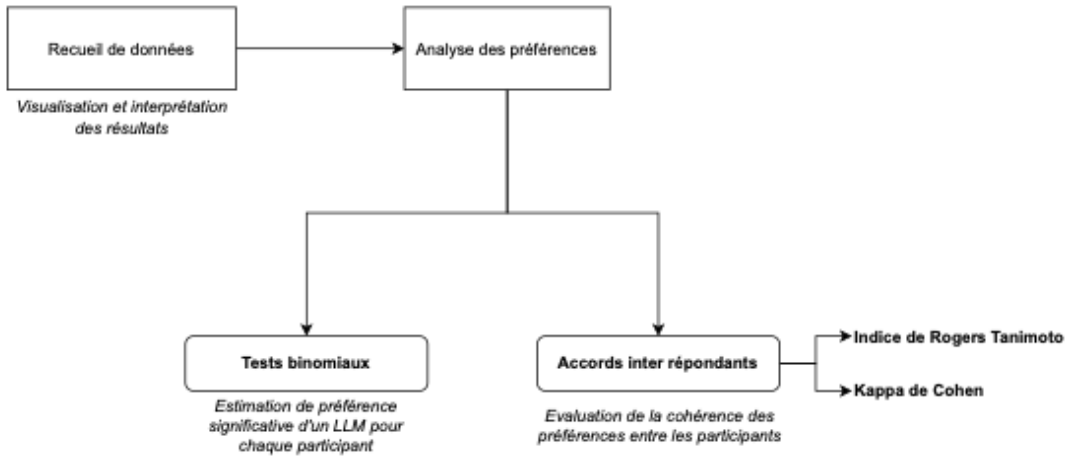


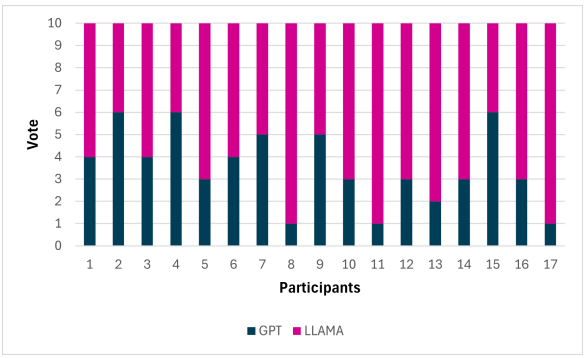
FIGURE 10.3 – Méthode d'analyse des données de comparaison par paires de codes Python produits par des LLM différents

10.4 Résultats

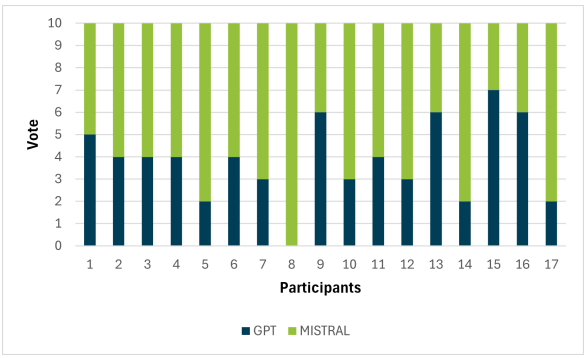
En observant nos distributions de préférences dans une tâche de comparaison par paires (voir figure 10.4), nous pouvons observer que :

— Code Llama est préféré plus de 6 fois sur 10 (64.7% du temps) à GPT-4.
— Mistral 7b est également préféré plus de 6 fois sur 10 (61,8% du temps) à GPT-4.
— Et enfin, Code Llama est préféré plus de 6 fois sur 10 (64.7% du temps) à Mistral 7b.

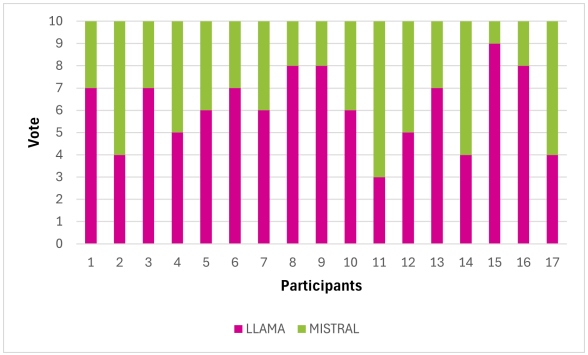
Donc au vu de nos distributions, les modèles les plus petits semblent être préférés à GPT-4, bien plus grand en termes de paramètres d'entraînement. Nous allons maintenant explorer ces résultats à l'aide des tests statistiques énoncés précédemment (voir section 10.3).



(a) GPT-4 et Code Llama 7b



(b) GPT-4 et Mistral 7b



(c) Code Llama 7b et Mistral 7b

FIGURE 10.4 – Distribution des préférences de modèles par comparaison

10.4.1 Tests binomiaux

Parmi les 12 participants sur 17 qui ont préféré Code Llama 7b à GPT-4, les tests binomiaux nous montrent que cette préférence n'est significative statistiquement que pour trois d'entre eux (voir figure 10.4a et annexe 8) [6].

6. Aucune préférence significative parmi les 5 participants qui ont préféré GPT-4

De même que parmi les 12 participants qui ont préféré Mistral 7b à GPT-4, les tests binomiaux ont montré qu'une seule préférence est statistiquement significative (voir figure 10.4b et annexe 9). Áu vu de ces résultats, la préférence pour Code Llama 7b par rapport à GPT-4 peut donc être considérée comme plus marquée que la préférence de Mistral 7b par rapport à GPT-4. Nous allons donc maintenant comparer les deux modèles à 7 milliards de paramètres.

Concernant, la comparaison entre Code Llama 7b et Mistral 7b, 11 participants ont préféré Code Llama. Mais au seuil de significativité statistique de 5%, les tests binomiaux ne montrent aucune préférence significative (voir figure 10.4c et annexe 10).

Globalement, les développeurs ont donc montré une certaine préférence pour Code Llama 7b par rapport aux deux autres LLM, bien que la différence soit très peu significative d'un point de vue statistique. Nous expliquons ces résultats par deux hypothèses :

— La concision et la qualité des réponses fournies par Code Llama 7b ont influencé les préférences. En échangeant avec certains des participants, ils nous ont expliqué apprécier les réponses plus directes et moins verbeuses que celles souvent fournies par GPT-4.

— La faible taille de l'échantillon a réduit la puissance de nos tests statistiques.

Mistral 7b n'est pas en reste, car il est également préféré à GPT-4, bien que de manière moins marquée que Code Llama 7b. Mistral 7b semble offrir un bon compromis entre performance et concision, ce qui le rend attractif pour les développeurs par rapport à GPT-4. Nos tests binomiaux montrent que les deux modèles open sources sont compétitifs et parfois plus adaptés pour des tâches de génération de code Python. Code Llama 7b a un léger avantage, puisque préféré à un LLM censé être plus performant, mais aussi plus lourd comme GPT-4.

10.4.2 Accords inter-répondants

Á partir des données récoltées, nous avons calculé les scores de dissimilarité de Rogers-Tanimoto pour déterminer si les participants sont d'accord entre eux lorsqu'ils préfèrent un modèle à un autre pour les différentes comparaisons (voir figure 10.5b et tableau 10.1). Ces scores indiquent qu'aucun LLM ne semble préféré à un autre pour les trois comparaisons. Avec des scores moyens légèrement supérieurs à 0.5, il y aurait en moyenne à peu près autant de similarités que de dissimilarités entre les deux LLM comparés à chaque

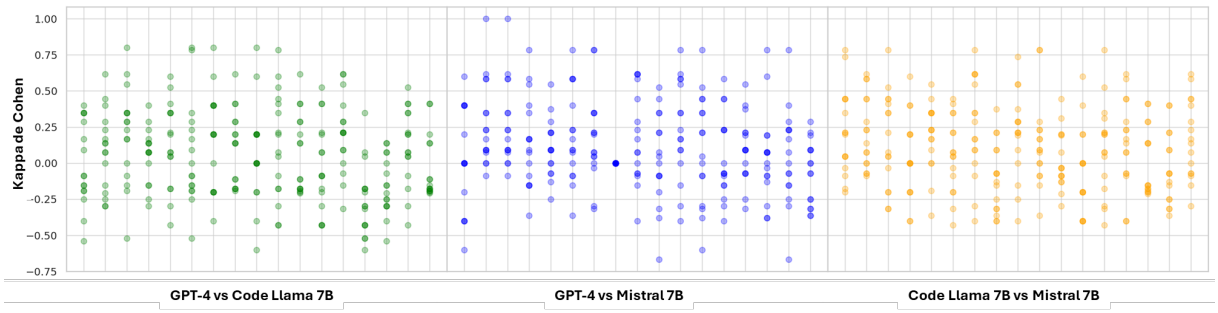
comparaison. De plus, avec un même écart interquartile pour les trois comparaisons ([.462 ; .667]), la dispersion des valeurs de dissimilarité semble modérée [7].

Les trois scores moyens de Kappa de Cohen montrent que les trois comparaisons génèrent des accords inter-répondants souvent attribuables au hasard en moyenne avec de scores proches de 0. En moyenne, c'est pour les comparaisons GPT-4 vs Mistral 7b que les accords semblent les plus élevés. Bien que faible, ceci indique une plus grande distinction entre les solutions de ces deux modèles.

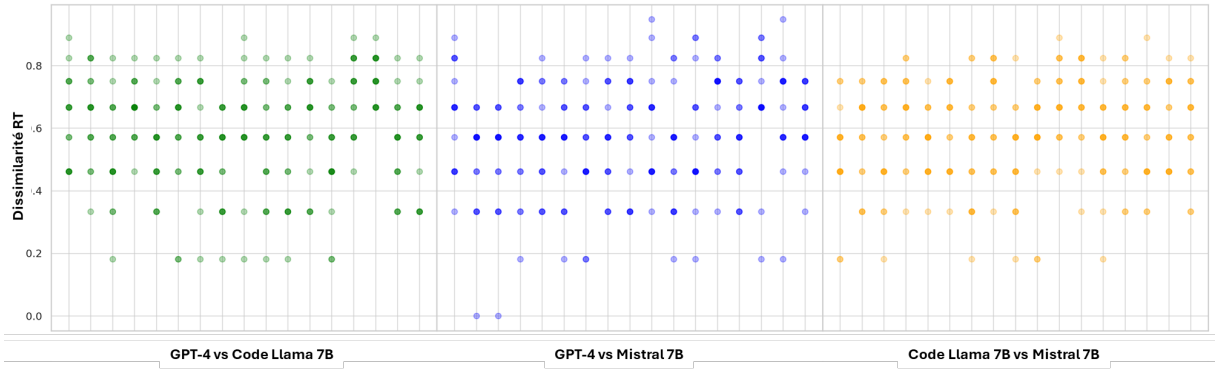
Comparaison	**Indice**	**Moyenne (*écart-type*)**	**[Min ; Max]**	**Écart inter-quartile**
GPT-4 vs Code Llama 7b	Kappa de Cohen	0.055 (0.303)	[-0.600 ; 0.800]	[-0.176 ; 0.286]
	Dissimilarité de Rogers-Tanimoto	0.588 (0.165)	[0.182 ; 0.889]	[0.462 ; 0.667]
GPT-4 vs Mistral 7b	Kappa de Cohen	0.111 (0.302)	[-0.667 ; 1.000]	[-0.071 ; 0.286]
	Dissimilarité de Rogers-Tanimoto	0.575 (0.171)	[0.000 ; 0.947]	[0.462 ; 0.667]
Code Llama 7b vs Mistral 7b	Kappa de Cohen	0.100 (0.290)	[-0.429 ; 0.783]	[-0.097 ; 0.286]
	Dissimilarité de Rogers-Tanimoto	0.584 (0.155)	[0.182 ; 0.889]	[0.462 ; 0.667]

TABLE 10.1 – Tableau descriptif des valeurs de Kappa de Cohen et de Dissimilarité de Rogers-Tanimoto entre les participants pour chaque comparaison

7. Nous avons également déterminé la présence de *outliers* (valeurs aberrantes) dans nos distributions en utilisant la méthode de Tukey [119] :
— seuil inférieur = $Q1 - 1.5 * IQR = 0.155$
— seuil supérieur = $Q3 - 1.5 * IQR = 0.975$
Seule une comparaison peut être considérée comme aberrante. Mais d'après nos tests complémentaires, l'extraction de cette valeur n'impacte pas significativement nos analyses.



(a) Kappa de Cohen



(b) Dissimilarité de Rogers-Tanimoto

FIGURE 10.5 – Accord moyen entre les participants basé sur la dissimilarité de Rogers Tanimoto et le coefficient Kappa de Cohen (chaque point représente un participant par rapport à un autre).

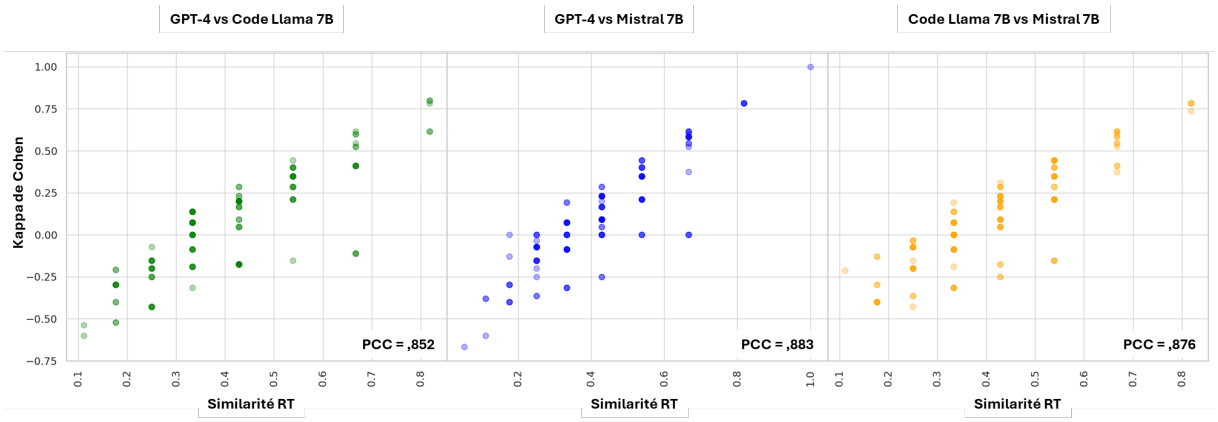


FIGURE 10.6 – Comparaison des scores de kappa de Cohen avec les scores de dissimilarité de RT. Pour la visualisation, les valeurs de similarité de Rogers-Tanimoto sont utilisées (1 - Dissimilarité RT).

Les coefficients de corrélation de Pearson entre les scores de similarité de Rogers-

Tanimoto et les scores de Kappa de Cohen sont élevés pour toutes les comparaisons (GPT-4 vs Code Llama 7b : 0.852, GPT-4 vs Mistral 7b : 0.883, Code Llama 7b vs Mistral 7b : 0.876) (voir figure 10.6. Cette forte corrélation positive indique que lorsque les solutions sont perçues comme similaires (faible dissimilarité de Rogers-Tanimoto), les développeurs ont tendance à être plus d'accords entre eux (kappa de Cohen élevé). Cela renforce l'idée que les développeurs peuvent clairement percevoir des différences entre certaines paires de modèles, en particulier face à GPT-4.

10.5 Conclusion

L'analyse des résultats dans le contexte de l'introduction de LLM pour assister les développeurs nous donne plusieurs points clés sur l'intégration de ces modèles sur les environnements professionnels.

Bien que le modèle GPT-4 soit utilisé comme modèle de référence par sa popularité, son score élevé aux datasets d'évaluation et son nombre de paramètres, nos résultats nous indiquent que les développeurs sont susceptibles de préférer les deux autres modèles présentés. Mistral 7b et Code Llama 7b sont des modèles plus petits et moins performants, mais pour autant plus frugaux et potentiellement plus sûrs en matière de partage de données. En échangeant avec certains participants de l'expérimentation, ils nous ont expliqué préférer les réponses moins fournies et plus concises que des réponses trop verbeuses avec beaucoup de commentaires (telles que GPT-4 a l'habitude de fournir).

Nous identifions cependant trois limites à notre expérimentation. Tout d'abord l'échantillon était assez restreint, il serait intéressant de poursuivre ces travaux avec un échantillon bien plus vaste pour voir si les résultats se confirment. Ensuite, les modèles que nous avons utilisés ont été rendus publics fin 2023, alors que nous pouvons constater l'apparition de nouveaux modèles au cours de 2024. Les LLM qui sont en train de faire leur apparition sont toujours plus performants aux jeux de données d'évaluation avec, pour certains, un nombre de paramètres encore plus restreints que pour Mistral 7b et Code Llama 7b car ils sont spécialisés dans des micro-tâches. Et enfin, le profil des développeurs au sein de la BU DIGITAL de SIGMA Informatique est fortement orienté Java et PHP. Bien que permettant d'éclairer l'aspect méthodologique pour choisir son LLM, nos résultats n'ont que peu d'impact direct sur la sélection du LLM pour concevoir un assistant au

développement orienté Java et/ou PHP. Nous serions donc fortement intéressés par l'idée de réitérer notre expérimentation avec un échantillon plus large, des modèles plus récents, potentiellement plus petits et sur un langage cohérent avec l'activité de la BU DIGITAL.

Conclusion de la Partie III

La dernière partie de cette thèse aura apporté des éclaircissements sur comment les LLM sont perçus au sein d'une entreprise du secteur numérique et comment ils peuvent s'intégrer dans une situation de travail en prenant en compte les enjeux des entreprises et les préférences des utilisateurs finaux. Les LLM sont capables d'un haut niveau de précision grâce à une immense quantité de données, ayant servie lors de leur conception, qui leur donne la capacité à générer une multitude de contenus (rapport administratif, code informatique, mails, etc.). Les LLM semblent également capables d'une grande capacité d'analyse (de contexte, de jeu de données, de contenu visuel, de texte, etc.). Ceci qui facilite grandement leur intégration aux usages professionnels avec des interactions Humain-IA qui se font plus naturellement, de manière similaire à des interactions entre humains.

Au cours de nos échanges avec les collaborateurs de SIGMA Informatique, nous avons constaté qu'il reste difficile de choisir l'outil génératif adéquat pour une situation de travail. Cela est d'autant plus vrai que leur nombre est florissant. Il y a également besoin d'une prise de position claire de la part des décideurs quant à l'usage de ces outils pour en tirer le plein potentiel. En effet, il existe une certaine crainte des collaborateurs à utiliser ou à déclarer utiliser les outils génératifs. Ils supposent que leurs compétences pourraient ne pas être reconnues à leur juste valeur. Ils craignent qu'on leur reproche d'utiliser des outils pour réaliser une partie de leurs tâches ou encore qu'on leur retire le mérite de l'accomplissement de la tâche. Ceci nous oriente vers le fait que l'acceptabilité des outils génératifs sur les postes de travail peut nécessiter des initiatives soutenues par l'organisation. Des actions telles que la formation et la sensibilisation, tout en tenant compte de l'objectif de leur mise en œuvre (cas d'usage), sont essentielles.

L'implication de l'organisation est également nécessaire pour minimiser les risques de sécurité et de confidentialité liés à l'usage de ces outils. Les outils génératifs sont souvent importés par les employés, mais il est du rôle de l'organisation de réguler leur usage en orientant vers des solutions satisfaisantes mais aussi sécuritaires. De plus, il est crucial de considérer l'impact écologique des LLM en favorisant l'utilisation de modèles frugaux.

Bien que ces modèles puissent être légèrement moins performants sur certaines tâches par rapport à des modèles volumineux comme GPT-4 de OpenAI, ils offrent une alternative plus respectueuse de l'environnement et parfois même préférée par les utilisateurs. En effet, les modèles plus petits consomment moins d'énergie pour leur entraînement et leur utilisation, réduisant ainsi l'empreinte carbone globale de l'entreprise. Adopter de tels modèles contribue, sur le long terme, à une approche plus durable et responsable de l'intelligence artificielle.

Conclusion générale

Apports de la thèse

Pour conclure ce document, nous allons revenir sur chacune des questions de recherche présentées en introduction et développer comment nos travaux permettent d'apporter des éléments de réponse à chacune de ces questions. Cette conclusion propose ainsi un autre niveau de lecture du manuscrit en se concentrant uniquement sur les chapitres qui répondent aux questions posées.

Quelle est la perception de l'IA en contexte professionnel ?

Lorsque nous nous sommes intéressés à la perception des solutions IA en contexte professionnel (figure 10.7), nous avons réalisé que l'IA est un enjeu majeur pour les organisations. Cette discipline promet aux entreprises de gagner en productivité, de réduire les erreurs humaines par l'automatisation et de gagner en créativité. Mais tous les professionnels n'ont pas la même perception de ces technologies comme nous le montrent les travaux de Yann Ferguson [59] qui relatent très bien la présence de représentations très disparates de ces technologies parmi les employés, tels que l'augmentation par l'IA ou encore le remplacement par l'IA. De ces constats, nous estimons qu'il est plus que nécessaire d'accroître la transparence de ces outils pour que les opérateurs humains aient une meilleure visibilité de ce que ces outils font et ne font pas correctement pour pouvoir positionner plus justement où l'IA peut intervenir par rapport à leur travail. C'est en comprenant l'outil qu'on devient plus susceptible de lui faire confiance lorsqu'il faut interagir avec. Faire confiance aux solutions IA est le premier pas vers une adoption consentie, car les employés sont plus susceptibles de les trouver acceptables, s'ils leur font confiance (voir chapitre 1).

Á partir de ces éclaircissements sur le sujet, nous sommes allés interroger l'écosystème IA (personnes en lien avec des projets IA) au sein de SIGMA Informatique et de certains de ses partenaires. Il y semble que la vision technocentrée de conception conduise à faiblement intégrer les utilisateurs finaux dans la mise en place/déploiement de ces solutions IA. De plus, la transparence de ces outils n'est pas un enjeu de conception malgré la forte demande des utilisateurs(voir chapitre 4). Face à un besoin exprimé par les utilisateurs qui a tendance à être ignoré, ou en tout cas peu pris en compte, nous avons expérimenté sur un facteur de transparence : l'indice de confiance que déclare un modèle prédictif. L'objectif de notre expérimentation était d'identifier l'effet que cet indicateur pourrait avoir sur la confiance accordée par l'humain. Les résultats obtenus nous orientent vers un

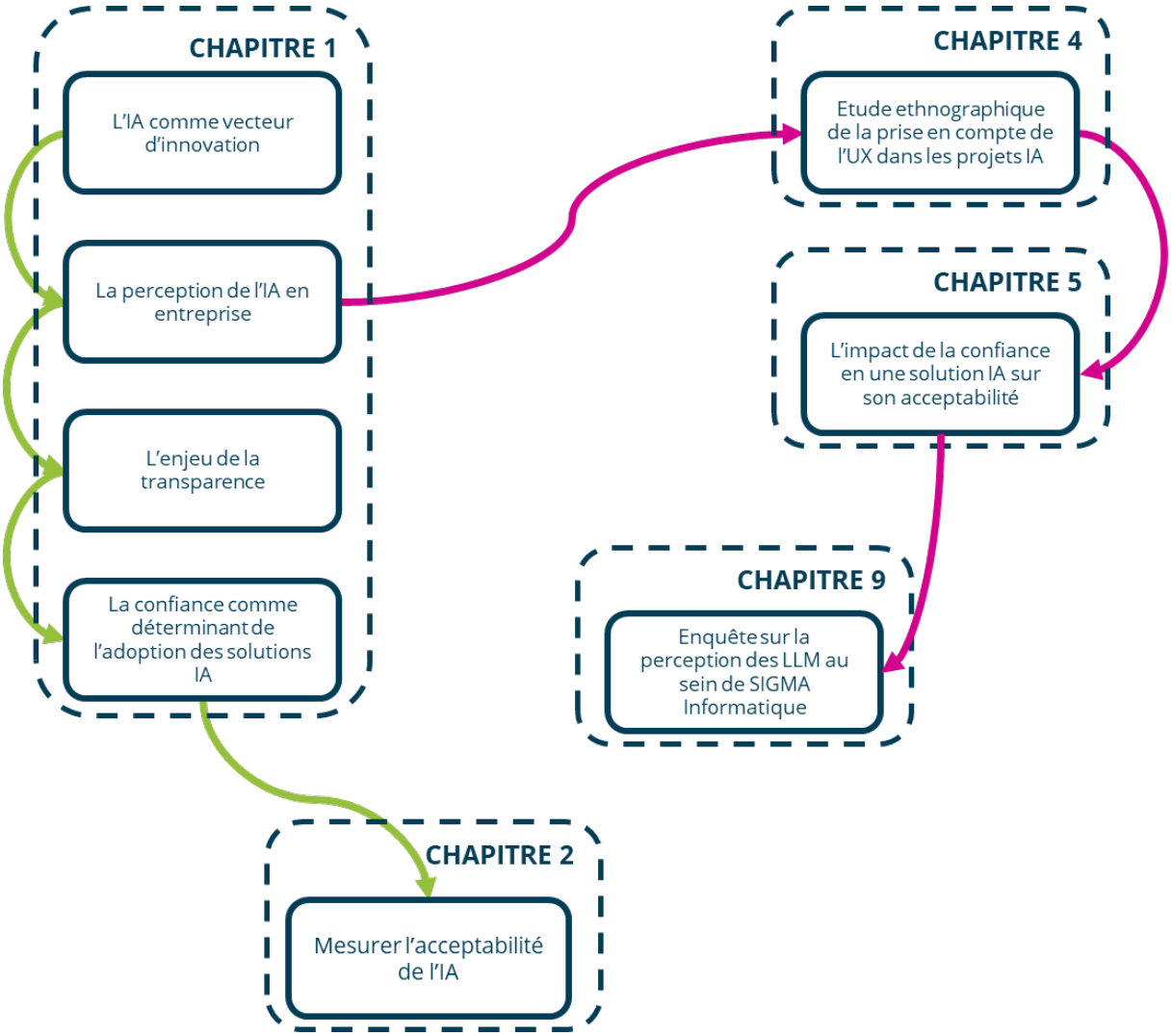


FIGURE 10.7 – Démarche d'étude de la perception de l'IA en contexte professionnel (Les flèches vertes représentent l'exploration de la littérature pour répondre à la question de recherche et les flèches violettes représentent les actions entreprises sur les terrains de SIGMA Informatique ou en laboratoire)

meilleur taux de confiance de la part de l'opérateur humain quand le modèle prédictif donne plus d'informations pour justifier sa réponse (voir chapitre 5).

Face à ces découvertes, nous nous sommes penchés sur les outils génératifs pour comprendre si la manière dont ils étaient perçus différait des solutions IA plus traditionnelles telles que celles explorées dans les parties I et II. C'est en tout cas ce que semble indiquer nos recherches, notamment au regard que les outils génératifs sont des outils de travail

portés par les employés eux-mêmes plutôt que par l'organisation (voir chapitre 8). Une des répercussions à ces tendances est qu'il existe une certaine gêne chez les employés à avouer s'en servir, le plus souvent par crainte d'un manque de reconnaissance. De plus, l'organisation arbore un regard assez différent envers les outils génératifs, notamment de crainte, ayant conscience de risques de sécurité plus élevés en cas de mauvais usage des employés.

Quelles sont les méthodes pour mesurer la confiance en des solutions IA ?

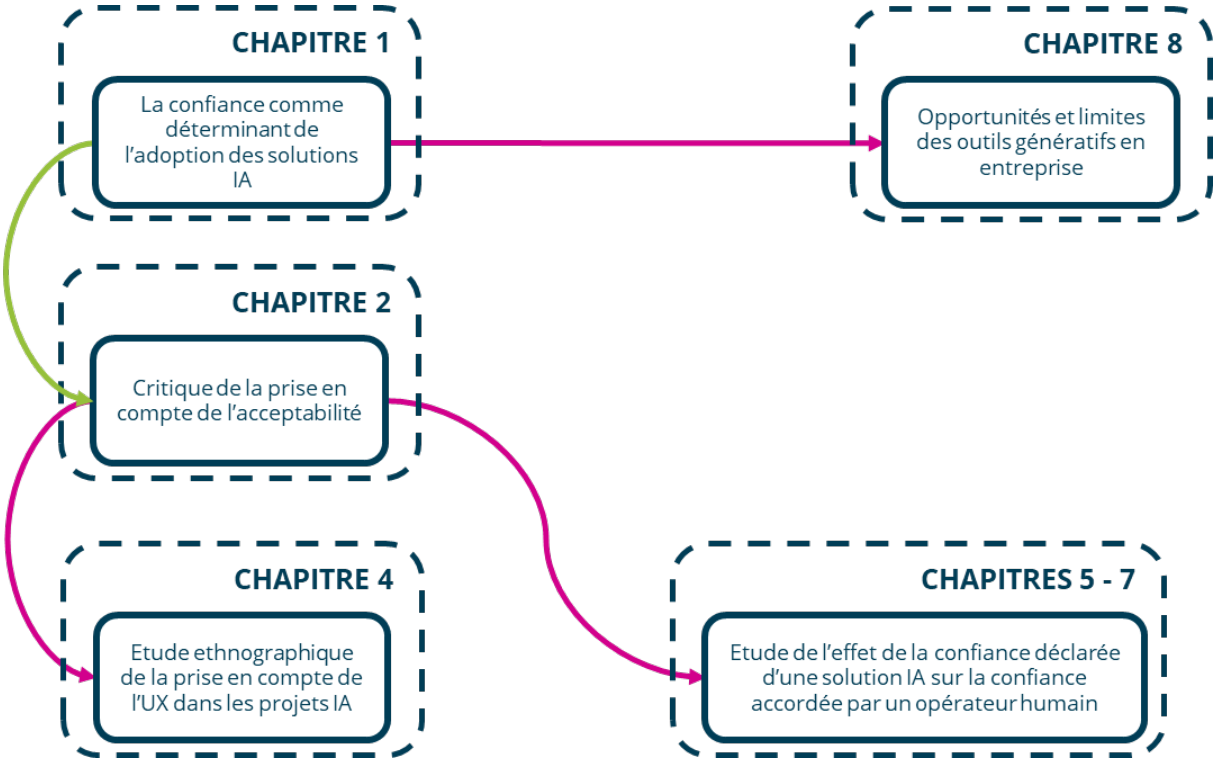


FIGURE 10.8 – Démarche d'étude des mesures de la confiance en des solutions IA (Légende : les flèches vertes représentent l'exploration de la littérature pour répondre à la question de recherche et les flèches violettes représentent les actions entreprises sur les terrains de SIGMA Informatique ou en laboratoire)

Ayant constaté que la confiance en une solution IA est un déterminant de son acceptabilité, nous avons réalisé que la plupart des moyens actuels pour mesurer l'acceptabilité sont insuffisants pour s'intéresser à ce concept en lien avec les solutions IA (voir chapitre 1). Bien que des questionnaires d'enquête, tels que ceux issus du TAM [43] nous montrent une bonne capacité prédictive, ces derniers sont focalisés sur l'interaction entre l'opérateur humain et la solution IA. Cependant l'aspect relationnel entre les deux est souvent ignoré alors qu'il est prépondérant au vu de la place qu'occupe la confiance pour vouloir utiliser ces outils (figure 10.8) (voir chapitre 2). C'est notamment ce que nous avons constaté en interrogeant les utilisateurs finaux des solutions IA conçues au sein de SIGMA Informatique et de ses partenaires (voir chapitre 4).

En questionnant la place de la confiance dans la conception de solutions IA, nous nous sommes aperçus qu'actuellement peu de moyens sont mis en place pour explorer ce concept dans le contexte de déploiement de solutions IA en entreprise. Nous avons donc mis en place une méthodologie d'évaluation de la confiance dans un contexte expérimental à l'aide de méthodes utilisées dans le cadre de l'évaluation subjective de la qualité d'expérience (QoE). Dans notre contexte, nous parlons d'accord Humain-IA lorsque la décision finale prise par l'opérateur humain est en adéquation avec la recommandation d'un modèle prédictif. Nous considérons qu'il y a un acte de confiance de l'humain envers la machine si les deux produisent le même résultat (voir chapitre 5, chapitre 6 et chapitre 7).

Avec l'arrivée des outils génératifs, il apparaît que le grand public est plus enclin à leur faire confiance qu'aux solutions IA précédentes. Ceci est notamment dû à leurs capacités qui les rendent très performants dans une multitude de domaines. De plus, en entreprise, l'employé crée lui-même les cas d'usage pour les outils génératifs au travail en étant décisionnaire des tâches dans lesquelles il va les intégrer (voir chapitre 8).

Quel est l'impact de la transparence sur la collaboration Humain-IA ?

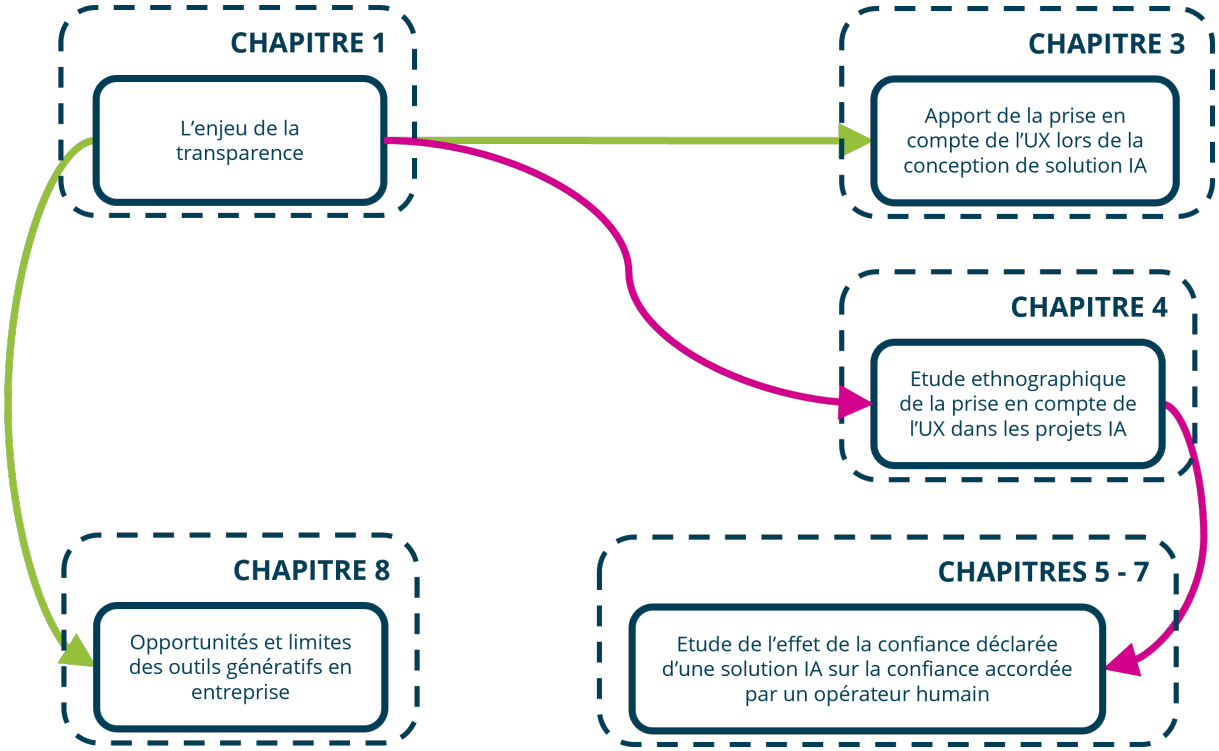


FIGURE 10.9 – Démarche d'étude de l'impact de la transparence sur la collaboration Humain-IA
(Les flèches vertes représentent l'exploration de la littérature pour répondre à la question de recherche et les flèches violettes représentent les actions entreprises sur les terrains de SIGMA Informatique ou en laboratoire)

Comme abordé précédemment, nous avons pu découvrir que la transparence des solutions IA représentait un enjeu déterminant dans leur acceptabilité (figure 10.9). Notamment parce que c'était un besoin exprimé de manière récurrente par les utilisateurs finaux des différentes solutions IA explorées (voir chapitre 1). Souvent ignoré, ce besoin nous a fait nous interroger sur la place de la prise en compte de l'expérience utilisateur (UX) dans la conception de ce type de technologies. Alors que cette action est considérée comme élémentaire dans une démarche de conception centrée utilisateurs, offrant des retours pertinents sur ce qui fonctionne, sur ce qui est apprécié ou encore sur ce qui est à améliorer sur les solutions conçues (voir chapitre 3 et chapitre 4).

Faiblement mise en avant par les concepteurs, la transparence est pourtant déterminante pour comprendre le fonctionnement des solutions IA, que ce soit leur comportement, les données utilisées pour les entraîner ou encore comment attribuer la responsabilité de la décision prise avec ou par la solution IA [21]. Et face à une demande récurrente de transparence de la part des utilisateurs finaux de projets comme l'assistance à la conception de marchandises, les concepteurs n'estiment pas toujours que cette information est nécessaire pour les utilisateurs, d'autant qu'elle engendrait une charge de travail supplémentaire pour l'équipe en charge de la conception du modèle. Cependant, avec ce besoin très clairement exprimé par les utilisateurs, nous avons fait l'hypothèse que cette information pourrait avoir un impact direct sur la confiance accordée par les utilisateurs aux prédictions d'un modèle IA. Cette hypothèse nous a conduit à réaliser une expérimentation sur l'effet de la certitude déclarée d'un modèle prédictif en ses propres prédictions sur la confiance qu'un opérateur humain accorde à ce modèle pour une tâche générique (estimation d'âge à partir de portraits photo dans notre cas de figure). Cette expérimentation a permis de mettre en avant que les opérateurs humains ont davantage confiance en la solution IA (mesurée par les accords Humain-IA) lorsque cette dernière communique d'avantage d'informations, notamment sur le degré de certitude en ses propres prédictions. Les opérateurs humains sont ainsi plus susceptible de travailler avec la solution IA. Mais pour cela, ils ont besoin de garder le contrôle sur la décision finale au cas où ils ne seraient pas d'accord avec le modèle (voir chapitre 5, chapitre 6 et chapitre 7).

Par la suite, nous avons commencé à nous interroger sur l'impact de la transparence, mais avec les outils génératifs qui sont connus pour leur complexité et opacité. Il semble que la transparence joue un rôle bien moins significatif dans l'acceptabilité de ces outils génératifs, mais il subsiste tout de même des interrogations de la part des utilisateurs professionnels notamment sur la source des informations mises à disposition par ceux-ci (voir chapitre 8).

Quelles typologies de collaboration Humain-IA en contexte professionnel ?

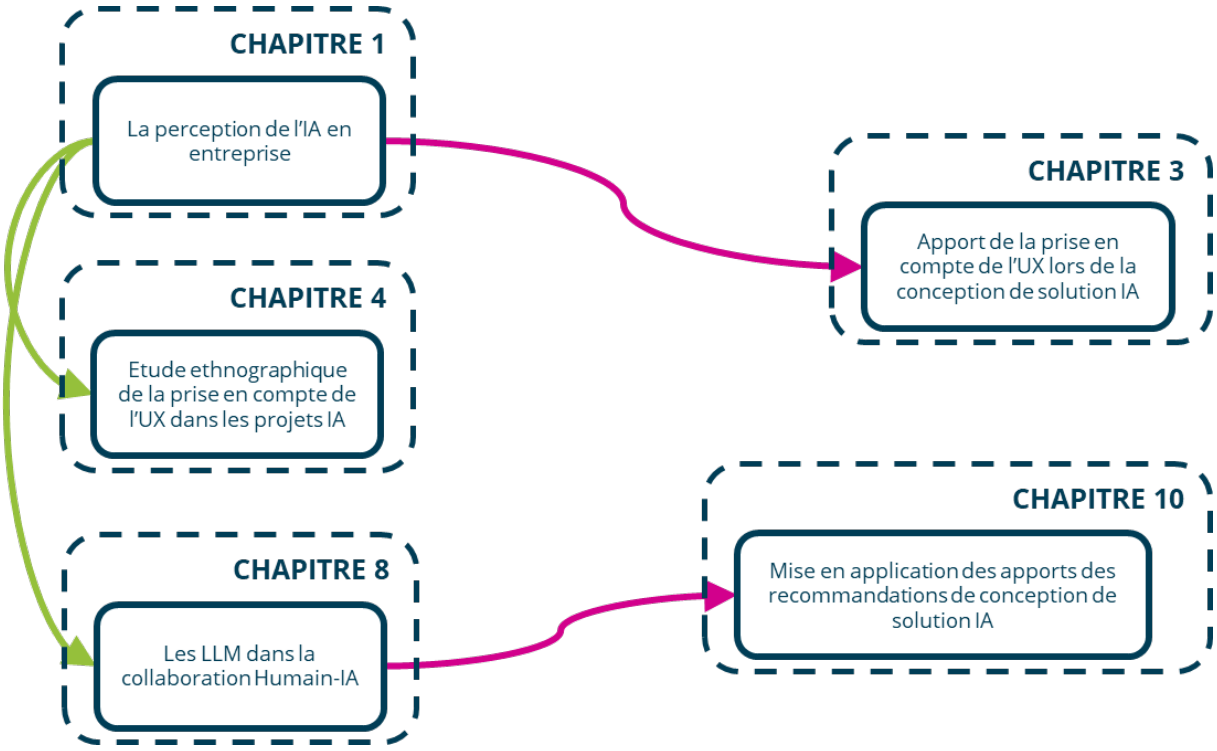


FIGURE 10.10 – Démarche d'étude des typologies de collaboration Humain-IA en contexte professionnel
(Les flèches vertes représentent l'exploration de la littérature pour répondre à la question de recherche et les flèches violettes représentent les actions entreprises sur les terrains de SIGMA Informatique ou en laboratoire)

Il apparaît que les solutions IA peuvent être de puissants partenaires professionnels pour aider les employés à réaliser leurs tâches, à condition d'être sur un mode de collaboration appropriée (voir chapitre 1). En effet, la solution IA peut être :

- Autonome : la solution IA est indépendante vis-à-vis de l'intervention humaine ;
- Assistée : la solution IA réalise des tâches spécifiques, mais nécessite une intervention humaine, par exemple pour valider la décision ;
- Conseillère : l'opérateur humain réalise la tâche, mais il est conseillé par la solution IA ;
- Collaboratrice : la solution IA et l'opérateur humain partagent les responsabilités pour accomplir une tâche commune. Elle peut, par exemple, réaliser des micro-

tâches facilement automatisables pour que l'opérateur humain se focalise sur celles à plus forte plus-value.

Lors de notre étude ethnographique sur la place des utilisateurs finaux dans la conception de solutions IA au sein de SIGMA Informatique, nous avons constaté que les solutions étaient principalement conseillères. Mais il subsiste une volonté des organisations d'accroître la performance de ces outils pour les rendre plus autonomes. Pour s'assurer que la solution IA conseillère est le mode de collaboration le plus adapté, il aurait été intéressant de mobiliser davantage d'effort sur la prise en compte de l'UX dans la conception. Les utilisateurs et leur besoin pourraient nécessiter d'interagir différemment selon le contexte (voir chapitre 3 et chapitre 4).

Dans notre projet de sélection de modèles, dans un contexte d'aide au développement, nous réalisons que l'interaction avec les outils génératifs est susceptible de bouleverser la manière de travailler (voir chapitre 9 et chapitre 10). Ces outils sont souvent sollicités le plus souvent utilisés :

— En tant que conseiller, l'utilisateur demande des informations à l'outil génératif, puis accomplit la tâche ;
— En tant qu'assistant, l'opérateur humain lui dit quoi faire avec des prompts, l'outil réalise la tâche qui lui est confiée et l'opérateur humain valide ou transpose les résultats obtenus par le LLM.

Perspectives

Notre exploration des déterminants de l'acceptabilité des solutions IA en contexte professionnel nous montre que la confiance est un facteur déterminant de leur intention d'usage. Pour autant, nous estimons qu'il peut être intéressant d'approfondir l'étude de facteurs complémentaires comme l'anthropomorphisme ou encore le type de collaboration. En se rapprochant des interactions Humain-Humain, l'utilisation de solution IA montre un besoin de considérer cette interaction au-delà de la capacité de l'outil à atteindre le but prescrit.

Au niveau méthodologique, utiliser des méthodes de comparaison des accords Humain-IA en fonction des modalités du modèle IA a mis en lumière un impact réel sur la confiance qu'accordent les opérateurs humains. Á partir de ces résultats, nous estimons qu'il serait souhaitable de réitérer cette expérimentation avec d'autres composantes de la transparence et pour une tâche plus représentative des situations de travail que nous avons rencontrées. Par exemple, dans le contexte de SIGMA Informatique, cette nouvelle version de l'expérimentation pourrait démontrer que la hausse de transparence des outils prédictifs aurait un effet sur l'adhésion des collaborateurs à l'utilisation de solutions IA portées par l'organisation.

Par ailleurs, les méthodes initialement employées pour de l'évaluation subjective de la QoE, combinées aux approches que nous avons présentées, montrent leur efficacité dans la mesure de la confiance et dans la conception de solutions IA plus acceptables. Cette approche à bas coût que nous avons utilisé pour sélectionner un LLM, comme socle technique d'un assistant de développement, est à répliquer dans des situations plus représentatives et significatives pour SIGMA Informatique. La majorité de l'activité de la BU DIGITAL nécessite des compétences en développement dans les langages Java et PHP. Notre dernière expérimentation concerne le langage Python, notamment par souci de cohérence avec les jeux de données d'évaluation de référence. Il faudrait la réitérer avec des langages de programmation plus cohérents avec l'activité professionnelle de l'organisation au sein de laquelle va se faire le déploiement. En complément, nous constatons que de plus en plus de LLM apparaissent sur le marché, dont certains de plus en plus frugaux et spécialisés dans des micro-tâches. Nous souhaitons donc réitérer notre dernière expérimentation, mais avec des LLM plus récents et avec une plus grande variabilité de taille, de complexité

et de frugalité. L'objectif serait d'identifier si par exemple, pour une même tâche, les développeurs préféreraient un LLM "généraliste" plus grand et moins frugal ou un LLM spécifique à la tâche, qui peut être bien plus petit, mais spécialisé dans beaucoup moins de domaines.

Nos travaux ont finalement apporté des éclaircissements quant aux facteurs d'acceptabilité de solutions IA en contexte professionnel, mais aussi des solutions méthodologiques à la conception d'outils IA sur-mesure qui s'ajustent au contexte réglementaire des sociétés numériques. La mise en place des LLM dans les environnements de travail vise à améliorer la performance des collaborateurs tout en réduisant les coûts et les erreurs humaines. Dans ce contexte, nous proposons un cadre sécuritaire pour concevoir des solutions plus acceptables pour les utilisateurs finaux. Les retombées attendues de cette méthodologie seraient 1) de réduire les coûts de conception associés en se basant sur les préférences des utilisateurs (sélection, configuration, de personnalisation et de formation), et 2) d'optimiser la performance des collaborateurs en libérant du temps pour des tâches à plus forte valeur ajoutée. En se basant sur les préférences des collaborateurs, les LLM choisis seraient plus susceptibles de répondre aux attentes et d'améliorer la satisfaction des collaborateurs avec une méthodologie qui standardise l'évaluation et la sélection de LLM, garantissant une approche cohérente et rigoureuse au sein d'organisation comme des entreprises du numérique.

BIBLIOGRAPHIE

[1] V. Agarwal, N. Thureja, M. K. Garg, S. Dharmavaram, D. Kumar et al., « " Which LLM should I use ?" : Evaluating LLMs for tasks performed by Undergraduate Computer Science Students in India », *arXiv preprint arXiv :2402.01687*, 2024.

[2] A. Agossah, F. Krupa, M. Perreira Da Silva, G. Deconde et P. Le Callet, « Déploiement de l'IA en situation de travail : une trop faible considération de l'expérience des employé · es ? », *Sciences du Design*, *2*, p. 68-85, 2022.

[3] A. Agossah, F. Krupa, M. Perreira Da Silva et P. Le Callet, « LLM-based Interaction for Content Generation : A Case Study on the Perception of Employees in an IT department », in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, 2023, p. 237-241.

[4] M. AI, *Mistral AI | Frontier AI in your hands — mistral.ai*, `https://mistral.ai/fr/`, [Accessed 22-05-2024].

[5] I. Ajzen, « The theory of planned behavior », *Organizational behavior and human decision processes*, t. 50, *2*, p. 179-211, 1991.

[6] I. Ajzen et M. Fishbein, « A Bayesian analysis of attribution processes. », *Psychological bulletin*, t. 82, *2*, p. 261, 1975.

[7] I. Ajzen et M. Fishbein, « Attitude-behavior relations : A theoretical analysis and review of empirical research. », *Psychological bulletin*, t. 84, *5*, p. 888, 1977.

[8] A. Ak, M. Abid, M. P. D. Silva et P. L. Callet, « On Spammer Detection In Crowdsourcing Pairwise Comparison Tasks : Case Study On Two Multimedia Qoe Assessment Scenarios », en, in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Shenzhen, China : IEEE, juill. 2021, p. 1-6, isbn : 978-1-66544-989-2. doi : `10.1109/ICMEW53276.2021.9455992`. adresse : `https://ieeexplore.ieee.org/document/9455992/` (visité le 11/05/2023).

[9] A. Ak, A. Goswami, W. Hauser, P. Le Callet et F. Dufaux, « Rv-tmo : Large-scale dataset for subjective quality assessment of tone mapped images », *IEEE Transactions on Multimedia*, 2022.

[10] L. Alben, « Quality of experience : defining the criteria for effective interaction design », *interactions*, t. 3, *3*, p. 11-15, 1996.

[11] B. Alexandre, F. Osiurak, J. Navarro et E. Reynaud, « Tool acceptance and acceptability : insights from a real tool use activity », *Cognitive Processing*, t. 22, *4*, p. 627-639, 2021.

[12] S. Amershi, D. Weld, M. Vorvoreanu et al., « Guidelines for Human-AI Interaction », issn : 9781450359702.

[13] K. Amoako-Gyampah et A. F. Salam, « An extension of the technology acceptance model in an ERP implementation environment », *Information & management*, t. 41, *6*, p. 731-745, 2004.

[14] S. Atarodi, A.-M. Berardi et A. M. Toniolo, « The technology acceptance model since 1986 : 30 years of development », *Psychologie du Travail et des Organisations*, t. 25, *3*, p. 191-207, 2019. doi : 10.1016/j.pto.2018.08.001.

[15] J. Austin, A. Odena, M. Nye et al., « Program synthesis with large language models », *arXiv preprint arXiv :2108.07732*, 2021.

[16] L. Bainbridge, « Ironies of Automation », *Automatica*, t. 19, *6*, p. 775-779, 1983, issn : 978-0-08-029348-6. doi : 10.1016/b978-0-08-029348-6.50026-9. adresse : http://www.bainbrdg.demon.co.uk/Papers/Ironies.html%0Ahttps://www.sciencedirect.com/science/article/pii/B9780080293486500269.

[17] J. Barcenilla et J.-M.-C. Bastien, « L'acceptabilité des nouvelles technologies : quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur ? », *Le travail humain*, t. 72, *4*, p. 311-331, 2009.

[18] J. Barcenilla et J. M. C. Bastien, « L'acceptabilité des nouvelles technologies : quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur ? : » fr, *Le travail humain*, t. Vol. 72, *4*, p. 311-331, mars 2010, issn : 0041-1868. doi : 10.3917/th.724.0311. adresse : https://www.cairn.info/revue-le-travail-humain-2009-4-page-311.htm?ref=doi (visité le 11/05/2023).

[19] I. Baroni, G. Re Calegari, D. Scandolari et I. Celino, « AI-TAM : a model to investigate user acceptance and collaborative intention inhuman-in-the-loop AI applications », en, *Human Computation*, t. 9, *1*, p. 1-21, mai 2022, issn : 2330-8001. doi : 10.15346/hc.v9i1.134. adresse : https://hcjournal.org/index.php/jhc/article/view/134 (visité le 11/05/2023).

[20] A. Barredo Arrieta, « TESIS_ALEJANDRO_BARREDO_ARRIETA.pdf », en, thèse de doct., UNIVERSITY OF THE BASQUE COUNTRY, 2022.

[21] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser et al., « Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI », *Information Fusion*, t. 58, *December 2019*, p. 82-115, 2020, Publisher : Elsevier B.V. doi : 10.1016/j.inffus.2019.12.012. adresse : https://doi.org/10.1016/j.inffus.2019.12.012.

[22] C. Bauchet, B. Hubert et J. Dinet, « Entre acceptabilité et appropriation des outils numériques intégrés dans le système éducatif : Le modèle des 4A », fr, 2020.

[23] L. Ben Ytzhak, *Petit détour par la vallée de l'étrange*, 2016. adresse : https://lejournal.cnrs.fr/articles/petit-detour-par-la-vallee-de-letrange.

[24] Y. Bisk, R. Zellers, J. Gao, Y. Choi et al., « Piqa : Reasoning about physical commonsense in natural language », in *Proceedings of the AAAI conference on artificial intelligence*, t. 34, 2020, p. 7432-7439.

[25] M. Bobillier-Chaumon et M. Dubois, « L'adoption des technologies en situation professionnelle : Quelles articulations possibles entre acceptabilité et acceptation ? », *Travail Humain*, t. 72, *4*, p. 355-382, 2009, issn : 9782130573258. doi : 10.3917/th.724.0355.

[26] S. Borel, Y. Ferguson et J. Condé, « Etude des impacts de l'IA sur le travail : Rapport d'enquête LaborIA Explorer », LaborIA, 2024. adresse : https://www.laboria.ai/laboria-explorer-synthese-generale/.

[27] N. Boujemaa, *Créer une Intelligence Artificielle de confiance est l'affaire de tous | Nozha Boujemaa | TEDxSaclay — ted.com*, https://www.ted.com/talks/nozha_boujemaa_creer_une_intelligence_artificielle_de_confiance_est_l_affaire_de_tous, [Accessed 22-02-2024].

[28] K. Brunnström, S. A. Beker, K. De et al., « Qualinet White Paper on Definitions of Quality of Experience Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad », *European Network on Quality of Experience in in Multimedia Systems and Services (COST Action IC 1003)*, *March*, p. 26-26, 2013.

[29] A. Cambon, B. Hecht, B. Edelman et al., « Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity », Technical Report MSR-TR-2023-43. Microsoft. https ://www. microsoft. com/en ..., rapp. tech., 2023.

[30] F. Cassano, J. Gouwar, D. Nguyen et al., « MultiPL-E : a scalable and polyglot approach to benchmarking neural code generation », *IEEE Transactions on Software Engineering*, 2023.

[31] F. Cassano, L. Li, A. Sethi et al., « Can It Edit ? Evaluating the Ability of Large Language Models to Follow Code Editing Instructions », *arXiv preprint arXiv :2312.12450*, 2023.

[32] A. Chander, R. Srinivasan, S. Chelian, J. Wang et K. Uchino, « Working with Beliefs : AI Transparency in the Enterprise », 2018.

[33] D. T. Chang, « Concept-Oriented Deep Learning with Large Language Models », *arXiv preprint arXiv :2306.17089*, 2023.

[34] Y. Chang, X. Wang, J. Wang et al., « A survey on evaluation of large language models », *ACM Transactions on Intelligent Systems and Technology*, t. 15, *3*, p. 1-45, 2024.

[35] P. Y. Chau et P. J. Hu, « Examining a model of information technology acceptance by individual professionals : An exploratory study », *Journal of management information systems*, t. 18, *4*, p. 191-229, 2002.

[36] H Chaudet, F Anceaux, M. C. Beuscart, S Pelayo et L Pellegrin, *Facteurs humains et ergonomie en informatique médicale*, Springer-V. Paris, 2013, Pages : 500 Publication Title : Informatique médicale, e-santé—fondements et applications., isbn : 978-2-8178-0337-1.

[37] M. Chen, J. Tworek, H. Jun et al., « Evaluating large language models trained on code », *arXiv preprint arXiv :2107.03374*, 2021.

[38] W. Cherif, « Adaptation de contexte basée sur la Qualité d'Expérience dans les réseaux Internet du Futur », thèse de doct., Université Rennes 1, 2013.

[39] E. Choi, H. He, M. Iyyer et al., « QuAC : Question Answering in Context », in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, p. 2174-2184.

[40] K. Cobbe, V. Kosaraju, M. Bavarian et al., « Training verifiers to solve math word problems », *arXiv preprint arXiv :2110.14168*, 2021.

[41] M. Colbert, « User experience of communication before and during rendezvous : interim results », *Personal and Ubiquitous Computing*, t. 9, p. 134-141, 2005.

[42] A. M. Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais et Z. M. J. Jiang, « Github copilot ai pair programmer : Asset or liability ? », *Journal of Systems and Software*, t. 203, p. 111 734, 2023.

[43] F. Davis, « Perceived Usefulness, Perceived Ease Of Use, And User Acceptance of Information Technology », *MIS Quarterly*, t. 13, *3*, p. 319-340, 1989, issn : 02767783. doi : 10.2307/249008.

[44] F. D. Davis, « A technology acceptance model for empirically testing new end-user information systems : Theory and results », thèse de doct., Massachusetts Institute of Technology, 1985.

[45] F. D. Davis, R. P. Bagozzi et P. R. Warshaw, « Extrinsic and intrinsic motivation to use computers in the workplace 1 », *Journal of applied social psychology*, t. 22, *14*, p. 1111-1132, 1992.

[46] F. D. Davis et V. Venkatesh, « A critical assessment of potential measurement biases in the technology acceptance model : Three experiments », *International Journal of Human Computer Studies*, t. 45, *1*, p. 19-45, 1996. doi : 10.1006/ijhc.1996.0040.

[47] G. Deconde, « Etude itérative des liens entre utilisabilité et acceptabilité d'un dispositif de saisie et de reconnaissance de l'écriture manuscrite », thèse de doct., 2011. adresse : http://www.sudoc.abes.fr/DB=2.1/SRCH?IKT=12&TRM=151303649.

[48] H. Delerue et C. Bérard, « Les dynamiques de la confiance dans les relations interorganisationnelles », *Revue Francaise de Gestion*, t. 175, *6*, p. 125-138, 2007, issn : 9782746219076. doi : 10.3166/RFG.175.125-138.

[49] E. R. DeLong, D. M. DeLong et D. L. Clarke-Pearson, « Comparing the areas under two or more correlated receiver operating characteristic curves : a nonparametric approach », *Biometrics*, p. 837-845, 1988.

[50] J. Deng et Y. Lin, « The benefits and challenges of ChatGPT : An overview », *Frontiers in Computing and Intelligent Systems*, t. 2, *2*, p. 81-83, 2022.

[51] S. Dick, « Artificial Intelligence », en, *Harvard Data Science Review*, juin 2019. doi : 10.1162/99608f92.92fe150c. adresse : https://hdsr.mitpress.mit.edu/pub/0aytgrau (visité le 11/05/2023).

[52] A. Dillon et M. G. Morris, « User acceptance of new information technology : theories and models », 1996.

[53] H. Dumez, « Les trois risques épistémologiques de la recherche qualitative », *Le libellio d'AEGIS*, t. 8, *4*, p. 29-33, 2012.

[54] J. Earthy, « Usability maturity model : Human centredness scale », *INUSE Project deliverable D*, t. 5, p. 1-34, 1998.

[55] K. D. Eason, « Towards the experimental study of usability », *Behaviour & Information Technology*, t. 3, *2*, p. 133-143, 1984.

[56] M. Eiband, D. Buschek, A. Kremer et H. Hussmann, « The impact of placebic explanations on trust in intelligent systems », *Conference on Human Factors in Computing Systems - Proceedings*, 2019, issn : 9781450359719. doi : 10.1145/3290607.3312787.

[57] R. Falotico et P. Quatto, « Fleiss' kappa statistic without paradoxes », *Quality & Quantity*, t. 49, p. 463-470, 2015.

[58] R. Feijo, « Planning your UX strategy », *Retrieved January*, t. 11, p. 2017, 2010.

[59] Y. Ferguson, « Ce que l'intelligence artificielle fait de l'homme au travail. Visite sociologique d'une entreprise », *Les mutations du travail*, 2019, issn : 9782348037498. doi : 10.3917/dec.dubet.2019.01.0023.

[60] Y. Ferguson, *Comprendre les enjeux sociétaux de l'IA*, AI ACT DAY 2023 organisé par Impact AI, Disponible sur YouTube, datacraft, 2023. adresse : https://www.youtube.com/watch?v=N9KeSJon5ho.

[61] F. Février, « Vers un modèle intégrateur" expérience-acceptation" : rôle des affects et de caractéristiques personnelles et contextuelles dans la détermination des intentions d'usage d'un environnement numérique de travail », thèse de doct., Université Rennes 2 ; Université Européenne de Bretagne, 2011.

[62] J. L. Fleiss, J. C. Nee et J. R. Landis, « Large sample variance of kappa in the case of different sets of raters. », *Psychological bulletin*, t. 86, *5*, p. 974, 1979.

[63] L. Floridi, J. Cowls, M. Beltrametti et R. Chatila, « AI4People — An Ethical Framework for a Good AI Society : Opportunities, Risks, Principles, and Recommendations », *Minds and Machines*, t. 28, *4*, p. 689-707, 2018, Publisher : Springer Netherlands, ISSN : 0123456789. DOI : 10.1007/s11023-018-9482-5. adresse : https://doi.org/10.1007/s11023-018-9482-5.

[64] J. Fraser et S. Plewes, « Applications of a UX Maturity Model to Influencing HF Best Practices in Technology Centric Companies – Lessons from Edison », *Procedia Manufacturing*, t. 3, *Ahfe*, p. 626-631, 2015, Publisher : The Authors. DOI : 10.1016/j.promfg.2015.07.285. adresse : http://dx.doi.org/10.1016/j.promfg.2015.07.285.

[65] T. Gamkrelidze, F. Barcellini et Z. Moustafa, « Intelligence Artificielle dans les activités professionnelles : quelles visions des acteurs concernés ? », *February*, 2021.

[66] T. Gamkrelidze, F. Barcellini et M. Zouinar, « Intelligence Artificielle : quelles conséquences sur les activités et l'organisation du travail ? », p. 1-7, 2020.

[67] O. Gillath, T. Ai, M. Branicky, S. Keshmiri, R. Davison et R. Spaulding, « Attachment and trust in artificial intelligence », *Computers in Human Behavior*, t. 115, *April 2020*, p. 106 607-106 607, 2021, Publisher : Elsevier Ltd. DOI : 10.1016/j.chb.2020.106607. adresse : https://doi.org/10.1016/j.chb.2020.106607.

[68] *GitHub - Netflix/sureal : Subjective quality scores recovery from noisy measurements.* — *github.com*, https://github.com/Netflix/sureal, [Accessed 09-05-2024].

[69] D. Glukhov, I. Shumailov, Y. Gal, N. Papernot et V. Papyan, « Llm censorship : A machine learning challenge or a computer security problem ? », *arXiv preprint arXiv :2307.10719*, 2023.

[70] G. GRONIER, « L'évaluation ergonomique d'un système expert relève-t-elle de l'expérience utilisateur ou de l'utilisabilité ? », 2018.

[71] J. A. Hanley et B. J. McNeil, « A method of comparing the areas under receiver operating characteristic curves derived from the same cases. », *Radiology*, t. 148, *3*, p. 839-843, 1983.

[72] M. Hassenzahl, « User experience (UX) towards an experiential perspective on product quality », in *Proceedings of the 20th Conference on l'Interaction Homme-Machine*, 2008, p. 11-15.

[73] M. Hassenzahl, S. Diefenbach et A. Göritz, « Needs, affect, and interactive products–Facets of user experience », *Interacting with computers*, t. 22, *5*, p. 353-362, 2010.

[74] M. Hassenzahl et N. Tractinsky, « User experience-a research agenda », *Behaviour & information technology*, t. 25, *2*, p. 91-97, 2006.

[75] J. Hatzius, J. Briggs, D. Kodnami et G. Pierdomenico, *The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani)*, https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html, [Accessed 09-05-2024], 2023.

[76] A. F. Hayes et K. Krippendorff, « Answering the call for a standard reliability measure for coding data », *Communication methods and measures*, t. 1, *1*, p. 77-89, 2007.

[77] P. Hekkert, « Design aesthetics : principles of pleasure in design », *Psychology science*, t. 48, *2*, p. 157, 2006.

[78] D. Hendrycks, C. Burns, S. Basart et al., « Measuring Massive Multitask Language Understanding », in *International Conference on Learning Representations*.

[79] D. Hendrycks, C. Burns, S. Kadavath et al., « Measuring Mathematical Problem Solving With the MATH Dataset », in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[80] O. Hengxuan Chi, S. Jia, Y. Li et D. Gursoy, « Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery », *Computers in Human Behavior*, t. 118, *May 2020*, p. 106 700-106 700, 2021, Publisher : Elsevier Ltd. doi : 10.1016/j.chb.2021.106700. adresse : https://doi.org/10.1016/j.chb.2021.106700.

[81] J.-C. Heudin, « Emotion selection in a multi-personality conversational agent », in *International Conference on Agents and Artificial Intelligence*, SCITEPRESS, t. 2, 2017, p. 34-41.

[82] J. M. Hoc, « La relation homme-machine en situation dynamique », *Revue d'Intelligence Artificielle*, t. 14, *1-2*, p. 55-71, 2000.

[83] C. P. Holland, « The importance of trust and business relationships in the formation of virtual organisations », *Organizational virtualness*, t. 3, p. 53-54, 1998. adresse : http://intranet.iwi.unisg.ch/org/iwi/iwi_pub.nsf/wwwPublAuthorGer/7B6FDD226B098CDDC1256DF900395DB0/$file/proc-98.pdf#page=54.

[84] L. M. Hsu et R. Field, « Interrater agreement measures : Comments on Kappan, Cohen's Kappa, Scott's π, and Aickin's α », *Understanding Statistics*, t. 2, *3*, p. 205-219, 2003.

[85] S. Imai, « Is GitHub Copilot a Substitute for Human Pair-programming ? An Empirical Study », en, in *2022 IEEE/ACM 44th International Conference on Software Engineering : Companion Proceedings (ICSE-Companion)*, Pittsburgh, PA, USA : IEEE, mai 2022, p. 319-321, isbn : 978-1-66549-598-1. doi : 10.1109/ICSE-Companion55297.2022.9793778. adresse : https://ieeexplore.ieee.org/document/9793778/ (visité le 11/05/2023).

[86] Impact IA (organization), « IA digne de confiance : construire une gouvernance adaptée à chaque entreprise », 2020.

[87] J. Isaak et M. J. Hanna, « User Data Privacy : Facebook, Cambridge Analytica, and Privacy Protection », *Computer*, t. 51, *8*, p. 56-59, 2018. doi : 10.1109/MC.2018.3191268.

[88] ISO - ISO 9241-210, « Human-centred design for interactive systems », rapp. tech., 2019, Publication Title : ISO/TC 159/SC 4 Ergonomics of human-system interaction, p. 33-33. adresse : https://www.iso.org/standard/77520.html.

[89] D. A. Jackson, K. M. Somers et H. H. Harvey, « Similarity coefficients : measures of co-occurrence and association or simply measures of occurrence ? », *The American Naturalist*, t. 133, *3*, p. 436-453, 1989.

[90] A. Jacovi, A. Marasović et T. Miller, « Formalizing Trust in Artificial Intelligence : Prerequisites , Causes and Goals of Human Trust in AI », *Section 2*, 2021, issn : 9781450383097.

[91] P. Jen-Hwa Hu, C. Lin et H. Chen, « User acceptance of intelligence and security informatics technology : A study of COPLINK », *Journal of the American Society for Information Science and Technology*, t. 56, *3*, p. 235-244, 2005.

[92] A. Q. Jiang, A. Sablayrolles, A. Mensch et al., *Mistral 7B*, 2023. arXiv : 2310.06825 [cs.CL].

[93] E. Jones, T. Oliphant, P. Peterson et al., *SciPy : Open source scientific tools for Python*, 2001–. adresse : http://www.scipy.org.

[94] M. Joshi, E. Choi, D. S. Weld et L. Zettlemoyer, « TriviaQA : A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension », in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 2017, p. 1601-1611.

[95] G. Jouis, « Explicabilité des modèles profonds et méthodologie pour son évaluation : application aux données textuelles de Pôle emploi », Thèse de doctorat dirigée par Mouchère, Harold et Picarougne, Fabien Informatique Nantes Université 2023, thèse de doct., 2023. adresse : http://www.theses.fr/2023NANU4007.

[96] H. C. Kelman, « Compliance, identification, and internalization three processes of attitude change », *Journal of conflict resolution*, t. 2, *1*, p. 51-60, 1958.

[97] H. C. Kim, « Acceptability engineering : The study of user acceptance of innovative technologies », *Journal of Applied Research and Technology*, t. 13, *2*, p. 230-237, 2015, Publisher : Universidad Nacional Autónoma de México, Centro de Ciencias Aplicadas y Desarrollo Tecnológico. doi : 10.1016/j.jart.2015.06.001. adresse : http://dx.doi.org/10.1016/j.jart.2015.06.001.

[98] J. K. Kim, M. Chua, M. Rickard et A. Lorenzo, « ChatGPT and large language model (LLM) chatbots : The current state of acceptability and a proposal for guidelines on utilization in academic medicine », *Journal of Pediatric Urology*, 2023.

[99] L. Krasula, K. Fliegel, P. Le Callet et M. Klima, « On the accuracy of objective image and video quality models : New methodology for performance evaluation », en, in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal : IEEE, juin 2016, p. 1-6, isbn : 978-1-5090-0354-9. doi : 10.1109/QoMEX.2016.7498936. adresse : http://ieeexplore.ieee.org/document/7498936/ (visité le 11/05/2023).

[100] L. Krasula, P. Le Callet, K. Fliegel et M. Klima, « Quality Assessment of Sharpened Images : Challenges, Methodology, and Objective Metrics », en, *IEEE Transactions on Image Processing*, t. 26, *3*, p. 1496-1508, mars 2017, issn : 1057-

7149, 1941-0042. DOI : 10.1109/TIP.2017.2651374. adresse : http://ieeexplore.ieee.org/document/7812797/ (visité le 11/05/2023).

[101] K. Krippendorff, *Computing Krippendorff's alpha-reliability*, 2011.

[102] F. Krupa, A. Leroi, M. Gioani, M. Rischewski et E. Durand, *UX Design : Le livret des méthodes*. Digital Design Lab de L'école de design Nantes Atlantique, 2020.

[103] T. Kwiatkowski, J. Palomaki, O. Redfield et al., « Natural questions : a benchmark for question answering research », *Transactions of the Association for Computational Linguistics*, t. 7, p. 453-466, 2019.

[104] M. Körber, « Theoretical considerations and development of a questionnaire to measure trust in automation », 2018.

[105] J. R. Landis et G. G. Koch, « The measurement of observer agreement for categorical data », *biometrics*, p. 159-174, 1977.

[106] W. H. Lee, C. W. Lin et K. H. Shih, « A technology acceptance model for the perception of restaurant service robots for trust, interactivity, and output quality », *International Journal of Mobile Communications*, t. 16, *4*, p. 361-376, 2018. DOI : 10.1504/IJMC.2018.092666.

[107] B. Lelong, F. Thomas et C. Ziemlicki, « Des technologies inégalitaires ? L'intégration de l'internet dans l'univers domestique et les pratiques relationnelles », *Réseaux*, *5-6*, p. 141-180, 2004.

[108] F.-M. Lesaffre et M. Poujol, *IA de confiance : quels enjeux ?*, 2020. adresse : https://app.livestorm.co/decode-media-sas/ia-and-confiance-quels-enjeux (visité le 23/05/2023).

[109] R. Li, L. B. Allal, Y. Zi et al., « StarCoder : may the source be with you ! », *Transactions on machine learning research*, 2023.

[110] T. Li, M. Vorvoreanu, D. DeBellis et S. Amershi, « Assessing human-ai interaction early through factorial surveys : A study on the guidelines for human-ai interaction », *ACM Transactions on Computer-Human Interaction*, t. 30, *5*, p. 1-45, 2023.

[111] Z. Li et C. G. Bampis, « Recover Subjective Quality Scores from Noisy Measurements », in *2017 Data Compression Conference (DCC)*, 2017, p. 52-61. DOI : 10.1109/DCC.2017.26.

[112] Z. Li et C. G. Bampis, « Recover subjective quality scores from noisy measurements », in *2017 Data compression conference (DCC)*, IEEE, 2017, p. 52-61.

[113] X. Liu, H. Yu, H. Zhang et al., « AgentBench : Evaluating LLMs as Agents », in *The Twelfth International Conference on Learning Representations*.

[114] Z. Luo, C. Xu, P. Zhao et al., « WizardCoder : Empowering Code Large Language Models with Evol-Instruct », in *The Twelfth International Conference on Learning Representations*.

[115] S. Mahlke et M. Thüring, « Studying antecedents of emotional experiences in interactive contexts », in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, p. 915-918.

[116] A. Mäkelä et J Fulton Suri, « Supporting users' creativity : Design to induce pleasurable experiences », p. 387-394, 2001.

[117] P. Maniatis et D. Tarlow, « Large sequence models for software development activities », *Google Research Blog–https ://ai. googleblog. com/2023/05/large-sequence-models-for-software. html. Accessed*, p. 06-04, 2023.

[118] K. Mathieson, E. Peacock et W. W. Chin, « Extending the technology acceptance model : the influence of perceived user resources », *ACM SIGMIS Database : the DATABASE for Advances in Information Systems*, t. 32, *3*, p. 86-112, 2001.

[119] R. McGill, J. W. Tukey et W. A. Larsen, « Variations of box plots », *The american statistician*, t. 32, *1*, p. 12-16, 1978.

[120] N. McNamara et J. Kirakowski, « Functionality, usability, and user experience : Three areas of concern », *interactions*, t. 13, *6*, p. 26-28, 2006.

[121] K. Merritt et S. Zhao, « An innovative reflection based on critically applying ux design principles », *Journal of Open Innovation : Technology, Market, and Complexity*, t. 7, *2*, 2021. doi : 10.3390/joitmc7020129.

[122] *Microsoft Copilot : Your everyday AI companion — copilot.microsoft.com*, https://copilot.microsoft.com/, [Accessed 22-05-2024].

[123] M. Midena, *Découvrez En Exclusivité L'Étude Qui Prouve Que L'IA Augmente de 80Forbes France — forbes.fr*, https://www.forbes.fr/technologie/decouvrez-en-avant-premiere-letude-qui-prouve-que-lia-augmente-de-80-la-croissance-des-entreprises/, [Accessed 22-02-2024].

[124] K. W. Miller, J. Voas et G. F. Hurlburt, « BYOD : Security and privacy considerations », *It Professional*, t. 14, *5*, p. 53-55, 2012.

[125] T. Miller, « Explanation in artificial intelligence : Insights from the social sciences », *Artificial Intelligence*, t. 267, p. 1-38, 2019, Publisher : Elsevier B.V. doi : 10.1016/j.artint.2018.07.007. adresse : https://doi.org/10.1016/j.artint.2018.07.007.

[126] H. Naveed, A. U. Khan, S. Qiu et al., « A comprehensive overview of large language models », *arXiv preprint arXiv :2307.06435*, 2023.

[127] N. Nguyen et S. Nadi, « An empirical evaluation of GitHub copilot's code suggestions », en, in *Proceedings of the 19th International Conference on Mining Software Repositories*, Pittsburgh Pennsylvania : ACM, mai 2022, p. 1-5, isbn : 978-1-4503-9303-4. doi : 10.1145/3524842.3528470. adresse : https://dl.acm.org/doi/10.1145/3524842.3528470 (visité le 11/05/2023).

[128] J Nielsen, « Corporate ux maturity », *Nielsen Norman Group, URL : http ://www. nngroup. com/articles/usability-maturitystages-1-4*, 2006.

[129] J. Nielsen, *Usability engeeniring*, Academic P. Boston, 1993.

[130] J.-F. Nogier, T. Bouillot et J. Leclerc, *Ergonomie des interfaces-5e éd : Guide pratique pour la conception des applications web, logicielles, mobiles et tactiles*. Dunod, 2011.

[131] D. Norman, J. Miller et A. Henderson, « What you see, some of what's in the future, and how we go about doing it : HI at Apple Computer », in *Conference companion on Human factors in computing systems*, 1995, p. 155.

[132] D. A. Norman, *The Design of everything day*. Basic books, 1988, isbn : 978-0-465-06710-7.

[133] D. A. Norman et J. Nielsen, *The Definition of User Experience (UX)*. adresse : https://www.nngroup.com/articles/definition-user-experience/ (visité le 06/06/2023).

[134] M. Nourani, S. Kabir, S. Mohseni et E. D. Ragan, « The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems », *The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19)*, p. 97-105, 2019, issn : 9781450359719. doi : 10.1145/3290607.3312787.

[135] OCDE, *Perspectives de l'Emploi de l'OCDE 2019* (Perspectives de l'emploi de l'OCDE). OECD, sept. 2019, ISBN : 978-92-64-42974-1. DOI : 10.1787/b7e9e205-fr. adresse : https://www.oecd-ilibrary.org/employment/perspectives-de-l-emploi-de-l-ocde-2019_b7e9e205-fr.

[136] OpenAI, *ChatGPT*, https://openai.com/chatgpt, [Accessed 06-05-2024].

[137] H. M. L. Pasquier, « Définir l'acceptabilité sociale dans les modèles d'usage : Vers l'introduction de la valeur sociale dans la prédiction du comportement d'utilisation », Pages : 316, thèse de doct., 2012.

[138] K. Pernice, S. Gibbons, K. Moran et W. Kathryn, *The 6 Levels of UX Maturity*, 2021. adresse : https://www.nngroup.com/articles/ux-maturity-model/.

[139] R. Pillai et B. Sivathanu, « Adoption of AI-based chatbots for hospitality and tourism », *International Journal of Contemporary Hospitality Management*, t. 32, *10*, p. 3199-3226, 2020. DOI : 10.1108/IJCHM-04-2020-0259.

[140] *QU'EST-CE QUE L'UX, L'EXPÉRIENCE UTILISATEUR ?*, août 2019. adresse : https://www.usabilis.com/definition-ux-experience-utilisateur-user-experience/ (visité le 06/06/2023).

[141] P. Rajpurkar, J. Zhang, K. Lopyrev et P. Liang, « SQuAD : 100,000+ Questions for Machine Comprehension of Text », in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, p. 2383-2392.

[142] S. Rasnayaka, G. Wang, R. Shariffdeen et G. N. Iyer, « An empirical study on usage and perceptions of llms in a software engineering project », *arXiv preprint arXiv :2401.16186*, 2024.

[143] V. Rawte, P. Priya, S. M. T. I. Tonmoy, S. M. M. Zaman, A. Sheth et A. Das, *Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances : Readability, Formality, and Concreteness*, 2023. arXiv : 2309.11064 [cs.AI].

[144] M. Reinert, « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », *Les cahiers de l'analyse des données*, t. 8, *2*, p. 187-198, 1983.

[145] D. J. Rogers et T. T. Tanimoto, « A Computer Program for Classifying Plants : The computer is programmed to simulate the taxonomic process of comparing each case with every other case. », *Science*, t. 132, *3434*, p. 1115-1118, 1960.

[146] R. Rothe, R. Timofte et L. Gool, « IMDB-WIKI–500k+ face images with age and gender labels », *Online] URL : https ://data. vision. ee. ethz. ch/cvl/rrothe/imdb-wiki*, t. 4, 2015.

[147] B. Rozière, J. Gehring, F. Gloeckle et al., *Code Llama : Open Foundation Models for Code*, 2024. arXiv : 2308.12950 [cs.CL].

[148] C. Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », *Nature Machine Intelligence*, t. 1, *5*, p. 206-215, 2019, Publisher : Springer US, issn : 4225601900. doi : 10.1038/s42256-019-0048-x. adresse : http://dx.doi.org/10.1038/s42256-019-0048-x.

[149] A. Salles, K. Evers et M. Farisco, « Anthropomorphism in AI », *AJOB Neuroscience*, t. 11, *2*, p. 88-95, 2020. doi : 10.1080/21507740.2020.1740350.

[150] A. L. Samuel, « Some studies in machine learning using the game of checkers », *IBM Journal of research and development*, t. 3, *3*, p. 210-229, 1959.

[151] M. Sap, H. Rashkin, D. Chen, R. Le Bras et Y. Choi, « Social IQa : Commonsense Reasoning about Social Interactions », in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, p. 4463-4473.

[152] D. L. Scapin et C. J. M. Bastien, « Inspection d'interfaces et critères ergonomiques », thèse de doct., INRIA, 1996.

[153] J. Schade et B. Schlag, « Acceptability of urban transport pricing strategies », en, *Transportation Research Part F : Traffic Psychology and Behaviour*, t. 6, *1*, p. 45-61, mars 2003, issn : 13698478. doi : 10.1016/S1369-8478(02)00046-3. adresse : https://linkinghub.elsevier.com/retrieve/pii/S1369847802000463 (visité le 11/05/2023).

[154] B. Shackel et S. Richardson, *Human factors for informatics usability*. Cambridge University Press, 1991.

[155] E. Shaffer, *Institutionalization of usability*, 2004.

[156] J. H. Sharp, « Development, extension, and application : a review of the technology acceptance model », *Director*, t. 7, *9*, p. 3-11, 2006.

[157] D. Shin, « The effects of explainability and causability on perception , trust , and acceptance : Implications for explainable AI », *International Journal of Human - Computer Studies*, t. 146, *April 2020*, p. 102 551-102 551, 2021, Publisher : Elsevier Ltd. doi : 10.1016/j.ijhcs.2020.102551. adresse : https://doi.org/10.1016/j.ijhcs.2020.102551.

[158] V. K. Singh, E. Andre, S. Boll, M. Hildebrandt et D. A. Shamma, « Legal and Ethical Challenges in Multimedia Research », *IEEE Multimedia*, t. 27, *2*, p. 46-54, 2020. doi : 10.1109/MMUL.2020.2994823.

[159] *Site web du LaborIA — laboria.ai*, https://www.laboria.ai/, [Accessed 21-05-2024].

[160] M. Suzgun, N. Scales, N. Schärli et al., « Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them », in *Findings of the Association for Computational Linguistics : ACL 2023*, 2023, p. 13 003-13 051.

[161] D. Sward et G. Macarthur, « Making user experience a business strategy », in *E. Law et al.(eds.), Proceedings of the Workshop on Towards a UX Manifesto*, Citeseer, t. 3, 2007, p. 35-40.

[162] A. Talmor, J. Herzig, N. Lourie et J. Berant, « CommonsenseQA : A Question Answering Challenge Targeting Commonsense Knowledge », in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, p. 4149-4158.

[163] A.-L. Thieullent, A. Yardi, F. Schladitz et al., « The AI-powered enterprise : Unlocking the potential of AI at scale », Capgemini Research Institute, 2020. adresse : https://www.capgemini.com/fr-fr/perspectives/publications/entreprise-intelligence-artificielle/.

[164] H. Touvron, L. Martin, K. Stone et al., *Llama 2 : Open Foundation and Fine-Tuned Chat Models*, 2023. arXiv : 2307.09288 [cs.CL].

[165] A. Tricot, F. Plégat-Soutjis, J.-F. Camps, A. Amiel, G. Lutz et A. Morcillo, « Utilité, utilisabilité, acceptabilité : interpréter les relations entre trois dimensions de l'évaluation des EIAH », *Environnements Informatiques pour l'Apprentissage Humain*, p. 391-402, 2003, Place : Paris, ISSN : 9788578110796. DOI : 10.1017/CBO9781107415324.004.

[166] Venkatesh, Morris, Davis et Davis, « User Acceptance of Information Technology : Toward a Unified View », en, *MIS Quarterly*, t. 27, *3*, p. 425, 2003, ISSN : 02767783. DOI : 10.2307/30036540. adresse : https://www.jstor.org/stable/10.2307/30036540 (visité le 12/05/2023).

[167] V. Venkatesh et H. Bala, « Technology Acceptance Model 3 and a Research Agenda on Interventions », *Decision Sciences Institute*, t. 39, *2*, p. 273-315, 2008.

[168] V. Venkatesh et F. D. Davis, « A Theoretical Extension of the Technology Acceptance Model : Four Longitudinal Field Studies », *Management Science*, t. 46, *2*, p. 186-204, 2000, ISSN : 0025-1909, 0025-1909. DOI : 10.1287/mnsc.46.2.186.11926. adresse : http://pubsonline.informs.org/doi/abs/10.1287/mnsc.46.2.186.11926.

[169] V. Venkatesh, M. G. Morris, G. B. Davis et F. D. Davis, « User acceptance of information technology : Toward a unified view », *MIS Quarterly : Management Information Systems*, t. 27, *3*, p. 425-478, 2003. DOI : 10.2307/30036540.

[170] V. Venkatesh, J. Y. L. Thong et X. Xu, « Consumer Acceptance and Use of Information Technology : Extending the Unified Theory of Acceptance and Use of Technology », *MIS Quarterly*, t. 36, *1*, p. 157-178, 2012, ISSN : 9781479982752. DOI : 10.1109/MWSYM.2015.7167037.

[171] V. Venkatesh, J. Y. Thong et X. Xu, « Unified theory of acceptance and use of technology : A synthesis and the road ahead », *Journal of the Association for Information Systems*, t. 17, *5*, p. 328-376, 2016. DOI : 10.17705/1jais.00428.

[172] O. Vereschak, G. Bailly et B. Caramiaux, « How to Evaluate Trust in AI-Assisted Decision Making ? A Survey of Empirical Methodologies », *Proceedings of the ACM on Human-Computer Interaction*, t. 5, *CSCW2*, 2021. DOI : 10.1145/3476068.

[173] S. M. Vieira, U. Kaymak et J. M. Sousa, « Cohen's kappa coefficient as a performance measure for feature selection », in *International conference on fuzzy systems*, IEEE, 2010, p. 1-8.

[174] J. Wang et A. Moulden, « AI Trust Score : A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features », en, in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan : ACM, mai 2021, p. 1-7, isbn : 978-1-4503-8095-9. doi : 10.1145/3411763.3443452. adresse : https://dl.acm.org/doi/10.1145/3411763.3443452 (visité le 11/05/2023).

[175] J. Wang, Z. Yang, Z. Yao et H. Yu, « JMLR : Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability », *arXiv preprint arXiv :2402.17887*, 2024.

[176] P. Wang, L. Li, L. Chen et al., « Large language models are not fair evaluators », *arXiv preprint arXiv :2305.17926*, 2023.

[177] Q. Wang, M. Madaio, S. Kane, S. Kapania, M. Terry et L. Wilcox, « Designing responsible ai : Adaptations of ux practice to meet responsible ai challenges », in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, p. 1-16.

[178] Y. Wang, W. Zhong, L. Li et al., « Aligning large language models with human : A survey », *arXiv preprint arXiv :2307.12966*, 2023.

[179] Z. Wang, L. Zhang, C. Cao, N. Luo et P. Liu, *A Case Study of Large Language Models (ChatGPT and CodeBERT) for Security-Oriented Code Analysis*, 2024. arXiv : 2307.12488 [cs.CR].

[180] M. J. Warrens, « Five ways to look at Cohen's kappa », *Journal of Psychology & Psychotherapy*, t. 5, 2015.

[181] S. Woodill et Y. Akiyama, « Integrating User-Centred Design Approaches for a Course Design Framework for Interdisciplinary Studies in Teaching and Learning. », *Journal of Teaching and Learning*, t. 14, *1*, p. 93-107, 2020.

[182] X. J. Yang, V. V. Unhelkar, K. Li et J. A. Shah, « Evaluating Effects of User Experience and System Transparency on Trust in Automation », *ACM/IEEE International Conference on Human-Robot Interaction*, t. Part F1271, p. 408-416, 2017, issn : 9781450343367. doi : 10.1145/2909824.3020230.

[183] B. Yetiştiren, I. Özsoy, M. Ayerdem et E. Tüzün, « Evaluating the code quality of ai-assisted code generation tools : An empirical study on github copilot, amazon codewhisperer, and chatgpt », *arXiv preprint arXiv :2304.10778*, 2023.

[184] B. Yetistiren, I. Ozsoy et E. Tuzun, « Assessing the quality of GitHub copilot's code generation », in *Proceedings of the 18th international conference on predictive models and data analytics in software engineering*, 2022, p. 62-71.

[185] M. Yin, J. W. Vaughan et H. Wallach, « Understanding the effect of accuracy on trust in machine learning models », *Conference on Human Factors in Computing Systems - Proceedings*, p. 1-12, 2019, issn : 9781450359702. doi : 10.1145/3290605.3300509.

[186] M. Yin, J. Wortman Vaughan et H. Wallach, « Understanding the effect of accuracy on trust in machine learning models », in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, p. 1-12.

[187] C. Zeitoun, *Peut-on faire confiance à l'intelligence artificielle ?*, Publication Title : CNRS Le Journal, 2018. adresse : https://lejournal.cnrs.fr/articles/peut-faire-confiance-a-lintelligence-artificielle.

[188] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi et Y. Choi, « HellaSwag : Can a Machine Really Finish Your Sentence ? », in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 4791-4800.

[189] Y. Zhang, Q. Vera Liao et R. K. Bellamy, « Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making », *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 295-305, 2020, issn : 9781450369367. doi : 10.1145/3351095.3372852.

[190] Q. Zheng, X. Xia, X. Zou et al., *CodeGeeX : A Pre-Trained Model for Code Generation with Multilingual Evaluations on HumanEval-X*, 2023. arXiv : 2303.17568 [cs.LG].

[191] W. Zhong, R. Cui, Y. Guo et al., « AGIEval : A Human-Centric Benchmark for Evaluating Foundation Models », in *Findings of the Association for Computational Linguistics : NAACL 2024*, 2024, p. 2299-2314.

[192] A. Ziegler, E. Kalliamvakou, X. A. Li et al., « Measuring GitHub Copilot's Impact on Productivity », *Communications of the ACM*, t. 67, *3*, p. 54-63, 2024.

[193] *Gemini - Discutez pour donner vie à vos idées — gemini.google.com*, https://gemini.google.com/app?hl=fr, [Accessed 22-05-2024].

Liste des tableaux

Liste des figures

PRODUCTIONS SCIENTIFIQUES

Voici la liste des articles et communications réalisés au cours de la thèse. Seuls Les productions acceptées sont présentées dans cette partie.

Articles de revue

— Agossah, A., Krupa, F., Deconde, G., Perreira Da Silva, M., & Le Callet, P. (2022, November). Déploiement de l'IA en situation de travail : une trop faible considération des employés ? Sciences du Design, 15.

Communications avec acte en conférence internationale

— Agossah, A., Krupa, F., Deconde, G., Perreira da Silva, M., & Le Callet, P. (2022, April). Déploiement de l'IA en situation de travail : une trop faible considération des employés ? In 33ème Conférence Internationale Francophone sur l'Interaction Humain-Machine (IHM'22).

— Agossah, A., Lévêque, L., Perreira da Silva, M., Le Callet, P., Krupa, F., & Deconde, G. (2022, April). Liens entre confiance et acceptabilité dans un dispositif IA. In 33ème Conférence Internationale Francophone sur l'Interaction Humain-Machine (IHM'22).

— Agossah, A., Krupa, F., Perreira da Silva, M., & Le Callet, P. (2023, June). LLM-based interaction for content generation : a case study on the perceptions of employees in an IT department. In ACM International Conference on Interactive Media Experiences (IMX '23).

Posters

— Agossah, A. (2022, April). ACCEPTABILITÉ DES SOLUTIONS IA EN CONTEXTE PROFESSIONNEL. In Intelligences, différentes par nature-L'IA nantaise s'invite au muséum.

— Agossah, A. (2022, June). Mesurer la confiance accordée aux outils d'aide à la prise de décision. In JDoc, Journée des doctorants de l'Ecole Doctorale MathSTIC

Autres communications

— Agossah A. (2022). Etude de la transparence des outils prédictifs : quel impact de l'indice de confiance ? In Table ronde - Confiance numérique : Sciences humaines et sociales et conception avionique, organisé par l'Université Bordeaux Montaigne

— Agossah A. (2022, December). Quel est le rôle de la confiance dans l'acceptabilité de l'IA en situation de travail ? In Technoférence : Les dernières tendances de l'IA face aux enjeux du monde réel, organisée par Images & Réseaux

— Agossah A. (2023, September). L'acceptabilité de l'IA pour les métiers créatifs. In Table Ronde - IA & Acceptabilité, organisée par Nantes Digital Week

ANNEXES

Détails des indicateurs statistiques utilisés

Alpha de Krippendorff

L'alpha de Krippendorff est un indice statistique utilisé comme un indicateur de fiabilité inter-juges. Il mesure l'accord entre deux répondants ou plus qui doivent classer des éléments dans des catégories exclusives [76] [101]. Dans notre cas, nous nous en servons sur les réponses (Oui/Non) des participants de chaque groupes. Voici le détail de la formule.

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_{c\in R}\sum_{k\in R}\delta(c,k)\sum_{u\in U} m_u \frac{n_{cku}}{P(m_u,2)}}{\frac{1}{P(n,2)}\sum_{c\in R}\sum_{k\in R}\delta(c,k)P_{ck}} \quad (1)$$

, où D_o correspond au désaccord observé et D_e correspond au désaccord supposément attribué au hasard.

Pour calculer le désaccord observé D_o, on utilise donc :

— $\delta(c,k)$ qui mesure le désaccord entre les réponses c et k. Dans notre cas de figure, on considère un ensemble de réponses R $(R = \{"oui", "non"\})$ où c et k représentent toutes les paires de réponses possibles dans l'ensemble R. Ainsi, $\delta(c,k) = 0$ si toutes les réponses des participants sont identiques et qu'il y a donc un accord complet. Et à l'inverse s'il y a un désaccord complet entre les participants, c'est-à-dire que $c \neq k$ alors $\delta(c,k) = 1$;

— n correspond au nombre total de paires d'éléments $\{"oui", "non"\}$ pouvant être appariés ($n = 20$ pour le nombre de portraits à évaluer dans notre cas) ;

— u correspond à l'ensemble des réponses données par tous les participants pour tous les portraits ($u = 20$ c'est le nombre total de paire d'éléments évaluée par les participants) ;

— m_u représente le nombre d'items dans l'unité u ($m_u = 10$, c'est le nombre d'évaluation pour chaque portrait u) ;

— n_{cku} indique le nombre de paires (c,k) dans l'unité u ($n_{cku} \in [0,10]$). Sa valeur

dépend du nombre de fois que les réponses spécifiques c et k apparaissent ensemble dans une unité u ;

— $P(m_u, 2)$ est le nombre de combinaisons de évaluations possibles parmi m_u éléments. Ce qui signifie que $P(m_u, 2) = P(10, 2) = \frac{10\times9}{2} = 45$.

Pour calculer D_e (le désaccord attendu par hasard) dans le cadre de l'alpha de Krippendorff avec 20 paires d'éléments et 10 participants, où chaque élément reçoit une réponse binaire ("oui" ou "non"), nous utilisons les composantes suivants :

— $P(n, 2)$ qui désigne le nombre total de manières de choisir deux évaluations parmi toutes les évaluations données ;

— également $\delta(c, k)$ expliqué précédemment ;

— $\sum_{c\in R}\sum_{k\in R}\delta(c, k)$, ici cette composante correspond à la somme des désaccords pour toutes les combinaisons possibles de réponses dans R, qui contient "oui" ou "non". $\delta(c, k) = 0$ si $c = k$ et 1 si $c \neq k$

— P_{ck} qui correspond au nombre de manières dont la paire (c, k) peut être formée. Dans notre cas, la réponse "oui" correspond à n_c et la réponse "non" correspond à n_k. Ainsi, $P_{ck} = n_c n_k$ si $c \neq k$ et $P_{ck} = n_c(n_c - 1)$ si $c = k$.

Note Z pour comparer des AUC de courbes ROC

Proposée par Hanley et McNeil en 1983 [71] en alternative à la méthode de Delong [49], la méthode de comparaison d'AUC par note Z permet de déterminer s'il y a une différence entre des AUC d'une même distribution en fonction des manières de l'objectiver (FOs) et ce, de manière robuste et fiable en évaluant la performance des modèles prédictifs. Dans le cadre de notre étude, nous employons cette méthode sur un seul prédicteur (les réponses des participants par rapport aux réponses du modèle) selon différentes manières d'objectiver notre situation de test (les FO). En comparant les AUC des courbes ROC, nous pouvons donc identifier les FO qui offrent la meilleure performance prédictive et ainsi orienter nos efforts pour améliorer la qualité de nos prédictions d'accord entre l'humain et le modèle, et leur applicabilité dans des contextes réels. Voici la formule de la note Z :

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{SE^2_{AUC_1} + SE^2_{AUC_2} - 2rSE_{AUC_1}SE_{AUC_2}}} \tag{2}$$

, où SE représente l'écart-type d'une AUC et calculé de la manière suivante :

$$SE = \sqrt{\frac{AUC(1-AUC) + (n_{g1}-1)(Q_1 - AUC^2) + (n_{g2}-1)(Q_2 - AUC^2)}{n_{g1}n_{g2}}} \quad (3)$$

Nous allons expliquer ce à quoi correspondent les différentes composantes de l'équation 2 et de l'équation 3 dans notre contexte expérimental :

- n_{g1} et n_{g2} correspondent aux nombres d'éléments de chaque groupe de l'analyse ROC (respectivement le nombre d'accords Humain-IA et le nombre de désaccords Humain-IA)
- $Q_1 = \frac{AUC}{2-AUC}$
- $Q_2 = \frac{2AUC^2}{1+AUC}$
- r désigne ici le coefficient de corrélation estimée entre les deux aires, donné par la table des coefficients de corrélation de Hanley et McNeil [71] à partir de :
  - la corrélation moyenne entre les votes, calculée $\frac{r_n + r_A}{2}$, où R_n est le coefficient de corrélation (Kendall Tau) des accords Humain-IA pour les deux FO comparées et r_A est le coefficient de corrélation (Kendall Tau) des désaccords Humain-IA pour les deux FO comparées ;
  - et l'aire moyenne des AUC, calculée $\frac{AUC_1 + AUC_2}{2}$.

Après avoir obtenu notre valeur Z, nous déterminons si la différence entre les deux AUC est statistiquement significative avec la fonction de distribution cumulative ($cdf(z)$) comme valeur p (voir équation 4).

$$p = cdf(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} exp(\frac{z^2}{2})dz \quad (4)$$

Nous avons utilisé la *cdf* sur nos notes Z, calculées via la librairie python SciPy [93].

Matériel d'enquête

Pour les différents questionnaires qui affichent des affirmations, les répondants doivent décrire à quel point ils sont d'accord ou pas avec la phrase. Leurs réponses se positionnent alors sur une échelle de Likert allant de 1-Pas du tout à 7-Totalement. Nous utilisons le questionnaire du tableau 2, relatif aux travaux de Davis (1989) [43] dans la plupart de nos études de l'acceptabilité des solutions IA. Ce questionnaire est lui-même adapté de celui utilisé par Deconde dans sa thèse (tableau 3) [47].

Dimension	**Item**
Utilité perçue	L'utilisation de ... améliore mon efficacité.
	L'utilisation de ... améliore la qualité de mes prédictions.
	Je trouve que ... est un outil utile pour ... .
	L'utilisation de ... augmentera ma productivité pour ... .
Facilité d'utilisation perçue	Comprendre les prédictions du modèle ne demande pas beaucoup d'effort mental.
	Il est facile de comprendre comment ... fonctionne.
	Le fonctionnement de ... est claire et compréhensible.
Intention d'usage	Pour ... , j'ai l'intention d'utiliser ... .
	Je vais utiliser ... pour de futures tâches de ce type.

TABLE 2 – Questionnaire TAM (*Technology Acceptance Model*) utiliser pour mesurer l'intention d'usage d'un dispositif

Dimension	Item
Normes subjectives	Les gens qui influencent ma conduite penseraient que je devrais utiliser ce dispositif de saisie de texte.
	Les gens qui sont importants pour moi penseraient que je devrais utiliser ce dispositif de saisie de texte.
Utilité perçue	Utiliser ce dispositif de saisie de texte pourrait améliorer mes performances au quotidien.
	Utiliser ce dispositif de saisie de texte est une . . . . . idée.
	Utiliser ce dispositif de saisie pourrait améliorer mon efficacité au quotidien.
	Utiliser ce dispositif de saisie de texte est une idée . . . . .
	Utiliser ce dispositif de saisie pourrait améliorer ma productivité au quotidien.
	Je trouverai ce dispositif de saisie utile au quotidien.
Facilité d'utilisation perçue	Apprendre à utiliser le dispositif de saisie de texte serait facile pour moi.
	Je trouve facile de faire faire au dispositif de saisie de texte ce que je veux qu'il fasse.
	Je trouverais facile d'utiliser ce dispositif de saisie de texte.
	Je serais capable d'utiliser ce dispositif de saisie de texte.
	J'ai les compétences, les connaissances, et la capacité à utiliser ce dispositif de saisie de texte.
Attitude	Il me serait facile de devenir compétent dans l'utilisation de ce dispositif de saisie de texte.
	Je . . . . . l'idée d'utiliser ce dispositif de saisie de texte.
	L'utilisation de ce dispositif de saisie est entièrement sous mon contrôle.
	Utiliser ce dispositif de saisie de texte serait . . . . .
Intention d'usage	J'ai l'intention d'utiliser ce dispositif de saisie de texte à l'avenir.
	J'ai l'intention d'utiliser ce dispositif de saisie de texte fréquemment à l'avenir. ;

TABLE 3 – Questionnaire d'acceptabilité de dispositif de saisie de texte de G. Deconde, basé sur le TAM

Dimension	**Item**
Fiabilité / Compétence de ...	... est capable de correctement ... .
	... fonctionne de manière fiable.
	Un dysfonctionnement de ... est probable.*
	... peut commettre des erreurs de temps en temps.*
	Je suis confiant dans les capacités de ... à ... .
Compréhensibilité / Prédictabilité des résultats	L'état de ... est toujours clair pour moi.
	... réagit de manière imprévisible.*
	J'ai pu comprendre pourquoi certaines choses se sont produites.
	Il est difficile d'identifier ce que ... fera ensuite.*
Familiarité avec ...	Je connais déjà des outils comme ... .
	J'ai déjà utilisé des outils comme ... .
Intention des développeurs	Les développeurs de ... sont dignes de confiance.
	Les développeurs de ... prennent en compte mon bien-être de manière rigoureuse.
Propension à faire confiance à ...	Il faut être prudent avec les outils automatisés non familiers comme ... .*
	Je préfère faire confiance à un outil comme ... que de m'en méfier.
	Les outils comme ... fonctionnent généralement bien.
Confiance en l'automatisation	Je fais confiance à ... .
	Je peux me fier à ... .

TABLE 4 – Questionnaire TiA (*Trust in Automation*) utiliser pour mesurer l'intention d'usage d'un outil d'automatisation (Pour l'analyse et l'interprétation des résultats, il faut prendre en compte que les items suivis de * sont des items inversés.)

Dimension	Item
Poste de travail	Quel poste occupez-vous aujourd'hui au sein de la BU DIGITAL ?
	Quelle est votre niveau d'expérience sur ce type de poste (en années) ?
	Dans le cadre de votre travail, est-ce que vous utilisez des outils qui ne sont pas mis à disposition par SIGMA ? Si oui, de quel type d'outils il s'agit ? Et pour quel type de tâches vous les utilisez ?
	Lorsque vous rencontrez des difficultés à effectuer vos tâches, vous préférez : 1) Faire des recherche en ligne 2) Demander de l'aide à un collègue de travail 3) Utiliser avec la documentation officielle 4) Autre, préciser :
Connaissances sur les outils génératifs	Connaissez-vous ChatGPT ? Si oui, l'avez-vous déjà utilisé ?
	Connaissez-vous Github Copilot ? Si oui, l'avez-vous déjà utilisé ?
	Connaissez-vous d'autres outils de génération de code, de texte ou d'image ? Si oui, lesquels ?
	Dans le cadre de votre travail, est-ce que les outils génératifs vous semble utile ?
	Pour quelles tâches liées à votre travail estimez-vous que des outils génératifs seraient utiles ?
Questionnaire TAM	
Influence sociale	Si j'utilise des outils génératifs pour mon travail, je le dis à mes collègues.
	Je recommande à mes amis et collègues d'utiliser des outils génératifs.
	L'avis de mon entourage influence mon utilisation des outils génératifs.
Motivation hédonique	Utiliser des outils génératifs me semble agréable.
	Utiliser des outils génératifs me semble amusant.
	Utiliser des outils génératifs me semble très divertissant.

TABLE 5 – Questionnaire sur la perception des outils génératifs diffusé au sein de la BU DIGITAL de SIGMA Informatique

Dimensions	Remarques
Perception de la tâche	
Perception de l'outil prédictif	
Ressenti face aux indices de confiance affiché	
Ressenti face aux prédictions du modèle - lorsque la prédiction semble éloigné de l'âge réel avec une confiance forte - lorsque la prédiction semble éloigné de l'âge réel avec une confiance faible	
Discours libre	

TABLE 6 – Grille d'entretien post-expérimentation pour l'expérimentation d'estimation d'âge de portraits photos

Résultats de l'évaluation heuristique de l'outil de centralisation de rapports diagnostiques pour les bailleurs sociaux

	Critère ergonomique impacté	Description du problème identifié	Impact estimé
1	Guidage – Groupement / distinction des items	Certaines listes déroulantes contiennent beaucoup d'éléments (ex. sélection de pièces/d'espaces) – Plus il y a de choix à sa disposition, plus l'utilisateur mettra de temps à faire son choix	Mineur
2	Charge de travail – Brièveté – Concision	Temps de génération de nouvelles lignes de plus en plus long au fur et à mesure qu'il y a du contenu ajouté (ex. pour 200 lignes demandées, l'ajout prend plusieurs minutes) – L'augmentation du temps de génération entraîne une augmentation de la frustration des utilisateurs et une diminution de leur productivité	Critique / modérée
3	Charge de travail – Brièveté – Actions minimales	Lorsque l'utilisateur a une difficulté à identifier un élément (ex. une adresse) dans le PDF à retranscrire dans la solution cible, il doit manipuler un fichier PDF externe pour retrouver l'information – L'ajout d'une tâche annexe peut être sensiblement disruptif pour l'utilisateur (perte d'attention, frustration, confusion, etc.)	Modéré
4	Charge de travail – Densité informationnelle	Les utilisateurs font des copier-coller des informations du PDF vers la solution cible par colonne parce qu'il y a des listes déroulantes sur chaque ligne (autre type d'interaction qui est disruptif) – paraît plus intuitif pour les utilisateurs de traiter par colonne que par ligne – Les utilisateurs se sont approprié l'usage pour éviter d'avoir à changer d'interactions trop souvent ce qui pourrait leur rajouter une charge mentale (texte -> liste déroulante -> texte)	Modéré / Mineur
5	Contrôle explicite – Contrôle utilisateur	Les utilisateurs font des copier-coller des informations du PDF vers la solution cible par colonne parce qu'il y a des listes déroulantes sur chaque ligne (autre type d'interaction qui est disruptif) – paraît plus intuitif pour les utilisateurs de traiter par colonne que par ligne – Les utilisateurs se sont approprié l'usage pour éviter d'avoir à changer d'interactions trop souvent ce qui pourrait leur rajouter une charge mentale (texte -> liste déroulante -> texte)	Modérée / Mineur
6	Adaptabilité - Flexibilité	Les raccourcis clavier ne semblent pas cohérents pour les utilisateurs – Sur des tableurs, la touche TAB fait aller à la cellule de droite et la touche ENTRÉE fait aller à la cellule en dessous	Mineur

TABLE 7 – Évaluation heuristique de la prise en compte de l'UX dans le de centralisation de rapports diagnostiques

Le Framework SPACE

Le framework SPACE sert à évaluer et à améliorer la productivité des équipes de développement logiciel en prenant en compte diverses dimensions qui vont au-delà des simples métriques de production. Son auteur, Nicole Forsgren, reconnaît que la productivité des développeurs ne peut pas être mesurée uniquement en termes de code produit, mais doit également considérer la qualité de leur travail, leur bien-être, et l'efficacité de leur collaboration au sein de l'équipe. Voici comment fonctionne le framework SPACE :

- *Satisfaction and well being* : Cette dimension reflète l'épanouissement des développeurs dans leur travail et l'utilisation des outils, ainsi que leur santé et leur bonheur au travail. Elle se mesure à l'aide d'enquêtes de satisfaction, du feedback des employés, d'indicateurs de santé mentale et de bonheur au travail.
- *Performance* : Cette dimension vise à quantifier les résultats obtenus plutôt que la quantité de travail effectué. Elle se mesure à travers la qualité et la fiabilité du code, le taux d'adoption par les clients ou encore la satisfaction des clients.
- *Activity* : Il s'agit du décompte les activités des développeurs. Elle se mesure au moyen d'analyse des données système, comptage des *pull requests*, de documentation rédigée et de spécifications de conception créées.
- *Communication and collaboration* : Cette dimension vise à capturer comment les équipes de développement communiquent et collaborent entre elles. Elles se mesure notamment via la facilité d'accès à la documentation, la rapidité de réponse aux questions, ou le processus d'intégration des nouveaux membres de l'équipe.
- *Effiency and flow* : Cette dimension reflète la capacité des développeurs à accomplir leurs tâches avec peu d'interruptions ou de retards. Pour cette dimension, il est possible de mesure à la fois les retards et interruptions peuvent être causés par des systèmes ou des humains, au moyen de mesures auto-déclarées et observées.

Résultats des tests binomiaux dans la sélection de LLM selon les préférences des développeurs

Comparaison par paires des solutions générées par les LLM GPT-4 et Code Llama 7B

Dans le cadre de notre étude de sélection de LLM avec la méthode de comparaison par paire, nous avons confronté nos participants à deux codes Python produits par deux LLM différents en réponse à un problème de programmation. Il y avait 20 problème de programmation en tout. Trois LLM devaient être comparés, mais les comparaisons se sont bien faites deux à deux. A partir des données obtenus, nous avons réalisés un test binomial pour chaque participant pour identifier s'il y avait une différence significative dans les choix des modèles pour chaque participant. Nous avons appliqué cette méthode pour les comparaisons GPT-4 vs Code Llama (tableau 8), GPT-4 vs Mistral 7b (tableau 9) et Code Llama vs Mistral 7b (tableau 10).

Participant	Modèle	Nombre de votes	Total	Proportion	p
participant_1	GPT	4	10	0.400	0.754
	LLAMA	6	10	0.600	0.754
participant_2	GPT	6	10	0.600	0.754
	LLAMA	4	10	0.400	0.754
participant_3	GPT	4	10	0.400	0.754
	LLAMA	6	10	0.600	0.754
participant_4	GPT	6	10	0.600	0.754
	LLAMA	4	10	0.400	0.754
participant_5	GPT	3	10	0.300	0.344
	LLAMA	7	10	0.700	0.344
participant_6	GPT	4	10	0.400	0.754
	LLAMA	6	10	0.600	0.754
participant_7	GPT	5	10	0.500	1.000
	LLAMA	5	10	0.500	1.000
participant_8	GPT	1	10	0.100	0.021
	LLAMA	9	10	0.900	0.021
participant_9	GPT	5	10	0.500	1.000
	LLAMA	5	10	0.500	1.000
participant_10	GPT	3	10	0.300	0.344
	LLAMA	7	10	0.700	0.344
participant_11	GPT	1	10	0.100	0.021
	LLAMA	9	10	0.900	0.021
participant_12	GPT	3	10	0.300	0.344
	LLAMA	7	10	0.700	0.344
participant_13	GPT	2	10	0.200	0.109
	LLAMA	8	10	0.800	0.109
participant_14	GPT	3	10	0.300	0.344
	LLAMA	7	10	0.700	0.344
participant_15	GPT	6	10	0.600	0.754
	LLAMA	4	10	0.400	0.754
participant_16	GPT	3	10	0.300	0.344
	LLAMA	7	10	0.700	0.344
participant_17	GPT	1	10	0.100	0.021
	LLAMA	9	10	0.900	0.021

TABLE 8 – Effectif des préférences entre GPT-4 et Code Llama 7B (une valeur p inférieure à .05 signifie qu'au risque de 5% d'erreur, il y a une différence significative de préférences d'un modèle par rapport à l'autre pour un participant donné)

Comparaison par paires des solutions générées par les LLM GPT-4 et Mistral 7B

Participant	Modèle	Nombre de votes	Total	Proportion	p
participant_1	GPT	5	10	0.500	1.000
	MISTRAL	5	10	0.500	1.000
participant_2	GPT	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754
participant_3	GPT	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754
participant_4	GPT	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754
participant_5	GPT	2	10	0.200	0.109
	MISTRAL	8	10	0.800	0.109
participant_6	GPT	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754
participant_7	GPT	3	10	0.300	0.344
	MISTRAL	7	10	0.700	0.344
participant_8	MISTRAL	10	10	1.000	0.002
participant_9	GPT	6	10	0.600	0.754
	MISTRAL	4	10	0.400	0.754
participant_10	GPT	3	10	0.300	0.344
	MISTRAL	7	10	0.700	0.344
participant_11	GPT	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754
participant_12	GPT	3	10	0.300	0.344
	MISTRAL	7	10	0.700	0.344
participant_13	GPT	6	10	0.600	0.754
	MISTRAL	4	10	0.400	0.754
participant_14	GPT	2	10	0.200	0.109
	MISTRAL	8	10	0.800	0.109
participant_15	GPT	7	10	0.700	0.344
	MISTRAL	3	10	0.300	0.344
participant_16	GPT	6	10	0.600	0.754
	MISTRAL	4	10	0.400	0.754
participant_17	GPT	2	10	0.200	0.109
	MISTRAL	8	10	0.800	0.109

TABLE 9 – Effectif des préférences entre GPT-4 et Mistral 7B (une valeur p inférieure à .05 signifie qu'au risque de 5% d'erreur, il y a une différence significative de préférences d'un modèle par rapport à l'autre pour un participant donné)

Comparaison par paires des solutions générées par les LLM Code Llama 7B et Mistral 7B

Participant	Modèle	Nombre de votes	Total	Proportion	p
participant_1	LLAMA	7	10	0.700	0.344
	MISTRAL	3	10	0.300	0.344
participant_2	LLAMA	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754
participant_3	LLAMA	7	10	0.700	0.344
	MISTRAL	3	10	0.300	0.344
participant_4	LLAMA	5	10	0.500	1.000
	MISTRAL	5	10	0.500	1.000
participant_5	LLAMA	6	10	0.600	0.754
	MISTRAL	4	10	0.400	0.754
participant_6	LLAMA	7	10	0.700	0.344
	MISTRAL	3	10	0.300	0.344
participant_7	LLAMA	6	10	0.600	0.754
	MISTRAL	4	10	0.400	0.754
participant_8	LLAMA	8	10	0.800	0.109
	MISTRAL	2	10	0.200	0.109
participant_9	LLAMA	8	10	0.800	0.109
	MISTRAL	2	10	0.200	0.109
participant_10	LLAMA	6	10	0.600	0.754
	MISTRAL	4	10	0.400	0.754
participant_11	LLAMA	3	10	0.300	0.344
	MISTRAL	7	10	0.700	0.344
participant_12	LLAMA	5	10	0.500	1.000
	MISTRAL	5	10	0.500	1.000
participant_13	LLAMA	7	10	0.700	0.344
	MISTRAL	3	10	0.300	0.344
participant_14	LLAMA	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754
participant_15	LLAMA	9	10	0.900	0.021
	MISTRAL	1	10	0.100	0.021
participant_16	LLAMA	8	10	0.800	0.109
	MISTRAL	2	10	0.200	0.109
participant_17	LLAMA	4	10	0.400	0.754
	MISTRAL	6	10	0.600	0.754

TABLE 10 – Effectif des préférences entre Code Llama 7B et Mistral 7B (une valeur p inférieure à .05 signifie qu'au risque de 5% d'erreur, il y a une différence significative de préférences d'un modèle par rapport à l'autre pour un participant donné)

Dimensions de l'évaluation automatisée de LLM

Nous présentons ici en détail les six dimensions les plus utilisées pour refléter les capacités des LLM.

Compréhension du monde La compréhension du monde désigne la capacité d'un LLM à comprendre et à interpréter des informations factuelles et contextuelles. Des datasets comme NaturalQuestions [103] et Trivia QA [94] sont utilisés pour tester cette dimension. Ces outils comprennent des items qui sont des combinaisons de questions, réponses et preuves de réponse permettant de donner un score global. L'objectif est d'interroger les LLM sur des faits généraux et spécifiques, et leur capacité à fournir des réponses précises.

Résultats populaires agrégés Cette dimension évalue la performance des LLM en matière d'agrégation et de synthèse d'informations provenant de différentes sources. On parle également de précision multi-tâches, puisqu'il s'agit de couvrir un haut niveau de connaissance approfondie dans une multitude de domaine. Des outils tels que BBH (pour *BIG-Bench Hard*) [160] et AGI Eval [191] sont utilisés pour tester la capacité des LLM à fournir des réponses agrégées basées sur des sources diverses. Ces sources peuvent être des connaissances en matières élémentaires (histoire, géographie, biologie, etc.), le plus souvent en employant des examens réels initialement adressés à l'humain (ex. examens universitaires, tests d'admission). L'objectif de ce type d'évaluation est d'estimer correctement la capacité de production et de synthèse d'informations.

Mathématiques La compétence mathématique des LLM est testée à l'aide d'outils comme MATH [79] et GSM8K (pour *Grade School Math 8K*) [40]. Cette dimension est cruciale et nécessaire à évaluer car les compétences de résolution de problèmes mathématiques sont requises dans de nombreuses activités.

Code La dimension de codage évalue la capacité des LLM à comprendre et à générer du code informatique. HumanEval [37] et MBPP (pour *Mostly Basic Python Problems*) [15] sont les outils le plus couramment utilisés pour tester cette dimension. La capacité à analyser, générer et optimiser du code est particulièrement pertinente dans les environnements professionnels où la programmation et le développement logiciel sont primordiaux. Cette dimension est approfondie dans la section 8.3.1.

Compréhension de la lecture La compréhension de la lecture évalue la capacité des LLM à lire et interpréter des textes. SQuAD (pour *Stanford Question Answering Dataset*) [141] et QUAC (pour *Question Answering in Context*) [39] sont des outils clés pour cette évaluation. SQuAD est un dataset composé de questions issues du site Wikipedia, et qui explore les types de raisonnement requis pour y répondre. Et QUAC est un dataset composé de dialogues type questions-réponses, pour comprendre les sens cachés de dialogues.

Raisonnement / Bon sens Le raisonnement et le bon sens évaluent la logique intuitive des LLM et leur capacité à faire des déductions qui sont considérés comme du bon sens pour un être humain. Les outils comme PIQA (pour *Physical Interaction : Question Answering*) [24], SIQA (pour *Social Interaction QA*) [151], HellaSwag [188] et CommonSwagQA [162] sont utilisés pour tester cette dimension. La performance dans cette dimension est particulièrement pertinente pour évaluer la capacité des LLM à prendre des décisions éclairées et à fournir des recommandations sensées dans des scénarios professionnels.

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : **Acceptabilité de l'Intelligence Artificielle en contexte professionnel : facteurs d'influence et méthodologies d'évaluation**

Mot clés : Acceptabilité, Conception centrée-utilisateur, Confiance, Contexte professionnel, Intelligence artificielle

Résumé : La conception de solutions informatiques intégrant de l'Intelligence Artificielle, sur les postes de travail connaît une croissance rapide. Bien qu'elles améliorent la performance des tâches, elles rencontrent souvent une faible acceptabilité de la part des employés, principalement en raison de la peur du remplacement et de la méfiance envers les décisions automatisées. Dans ce contexte, nos travaux s'intéressent à comment une solution IA peut être acceptable par les employés, utilisateurs finaux, et les entreprises, décideuses de son introduction sur les postes de travail. Après un état de l'art sur la perception des solutions IA en entreprise et de la notion d'acceptabilité, nous nous intéressons à la mobilisation de méthodes de prise en compte de l'expérience utilisateur. Ensuite nous présentons les stratégies de conception de solutions IA qui sont actuellement mobilisées par SIGMA Informatique, et nous poursuivons avec une étude de l'effet de l'accroissement de la transparence de ces technologies sur la confiance que leur accordent les opérateurs humains. En complément, nous présentons également une méthodologie de mesure de la confiance des humains envers les prédictions d'une solution IA pour une tâche assistée par un algorithme prédictif. Pour évaluer cette confiance, nous utilisons notamment des méthodes utilisées pour faire de l'évaluation subjective de la qualité d'expérience (QoE). Les résultats révèlent que plus une solution IA communique d'informations pour accompagner sa prédiction, plus l'opérateur humain lui fera confiance. Et enfin, nous examinons l'impact de l'arrivée des outils génératifs, qui bouleversent les théories traditionnelles de la confiance accordée aux solutions IA. Ces nouvelles technologies, souvent introduites par les employés eux-mêmes, changent cette dynamique, nécessitant une adaptation des entreprises. Nous proposons donc une approche de conception, basée sur le recueil des préférences des utilisateurs finaux, pour sélectionner des socles techniques, visant à trouver un compromis entre performance, acceptabilité par les utilisateurs et contrainte de l'organisation. Ces travaux proposent donc un éclairage sur la place de facteurs, tels que la confiance, dans l'acceptabilité des solutions IA en contexte professionnel, mais aussi des socles méthodologiques d'évaluation et de conception de l'acceptabilité de ces solutions.

Title: **Acceptability of Artificial Intelligence in a Professional Context: influencing factors and evaluation methodologies**

Keywords: Acceptability, User-centered design, Trust, Professional context, Artificial intelligence

Abstract: The design of Artificial Intelligence-based computer solutions in the workplace is experiencing rapid growth. Although they improve task performance, they often face low acceptability from employees, mainly due to fear of replacement and distrust of automated decisions. In this context, our work focuses on how an AI solution can be acceptable to employees, the end-users, and to companies, the decision-makers of its introduction at workstations. After a state-of-the-art review on the perception of AI solutions in business and the notion of acceptability, we focus on the use of methods to consider user experience. We then present the design strategies of AI solutions currently used by SIGMA Informatique and continue with a study on the effect of increasing the transparency of these technologies on the trust they gain from human operators. Additionally, we present a methodology for measuring human trust in the predictions of an AI solution for a task assisted by a predictive algorithm. To evaluate this trust, we use methods typically employed for subjective quality of experience (QoE) assessment. The results reveal that the more an AI solution communicates information to support its prediction, the more the human operator will trust it. Finally, we examine the impact of the emergence of generative tools, which disrupt traditional theories of trust in AI solutions. These new technologies, often introduced by the employees themselves, change this dynamic, requiring companies to adapt. We therefore propose a design approach based on collecting end-user preferences to select technical foundations aimed at finding a compromise between performance, user acceptability, and organizational constraints. Thus, this work sheds light on the role of factors such as trust in the acceptability of AI solutions in a professional context, as well as methodological foundations for evaluating and designing the acceptability of these solutions.