HEC MONTRÉAL

Impact des visualisations d'explication de systèmes d'intelligence artificielle sur la charge cognitive et la certitude des utilisateurs

par

Antoine Hudon

Pierre-Majorique Léger et Sylvain Sénécal

HEC Montréal

Co-Directeurs de recherche

Sciences de la gestion
(Spécialisation Expérience utilisateur en contexte d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Juin 2021

Résumé

La recherche en intelligence artificielle (IA) interprétable (Explainable Artificial Intelligence [XAI]) vise à répondre au manque de transparence caractérisant les systèmes d'IA en développant des méthodes et techniques permettant de transformer ces systèmes en système interprétables pour l'utilisateur. Il a été démontré que cette augmentation en transparence et en interprétabilité a un impact positif sur la confiance et la certitude de l'utilisateur envers ces systèmes, facilitant le développement d'une meilleure collaboration entre l'utilisateur et les systèmes et entrainant une diminution des diverses craintes relatives à l'IA.

Plusieurs techniques d'interprétabilité ont été développées pour les systèmes de reconnaissance d'image. Ces techniques permettent de visualiser, sous forme de représentation visuelle, l'explication de la classification d'une image par le système. Ces visualisations de l'explication illustrent les zones de l'image analysée qui ont influencé le système lors de sa décision.

Ce mémoire par articles étudie donc l'impact de différentes visualisations de l'explication d'un système de reconnaissance d'image sur la charge cognitive de l'utilisateur ainsi que sur la certitude de l'utilisateur envers le système. Six types de visualisations sont évalués, dont chacun varie au niveau de l'adjacence entre l'image et l'explication ainsi qu'au niveau de la technique d'interprétabilité utilisée. Pour étudier ce phénomène, deux études ont été menées. La première, conduite en laboratoire auprès de 19 participants, étudie l'impact de chaque type de visualisations sur la charge cognitive et la certitude de l'utilisateur. L'expérience a eu recours à des mesures implicites (pupillométrie) et explicites (questionnaires). Les résultats obtenus suggèrent que l'adjacence et la technique d'interprétabilité utilisée ont un effet sur la charge cognitive de l'utilisateur et qu'il y a une corrélation négative entre la charge cognitive et la certitude de l'utilisateur. La deuxième étude a été menée auprès de 350 participants dans le cadre d'une enquête en ligne évaluant l'effet de chacun des types de visualisations sur la certitude de l'utilisateur

envers le système d'IA. Ici aussi, l'adjacence et la technique d'interprétabilité ont un effet sur la certitude de l'utilisateur.

Ce mémoire contribue à la recherche en XAI en conservant l'humain au cœur de l'évaluation des techniques d'interprétabilité. En effet, la recherche en XAI se concentre davantage sur la création de modèles mathématiques interprétables négligeant l'expérience de l'utilisateur final. Notre approche apporte des explications centrées utilisateur favorisant la collaboration entre l'humain et les systèmes d'IA.

Mots clés : Certitude, Confiance, Charge cognitive, Explications, Visualisations, Intelligence artificielle, Système de reconnaissance d'image, Cognitive fit, Pupillométrie

Méthodes de recherche : Enquête, Expérimentation, Intelligence artificielle et heuristique, Recherche quantitative

Table des matières

Liste des tableaux et des figures

La numérotation des figures et tableaux a été modifiée dans les articles pour faciliter la lecture de ce mémoire.

Liste des tableaux

Liste des figures

Liste des abréviations

AI : Artificial intelligence

CF : Cognitive fit

CNN : Convolutional Neural Networks

CP : Cloud of points

DKN : Do not know the object's name

DNO : Did not know the object

EV : Explanation visualizations

HM : Heatmap

IA : Intelligence artificielle

MC : Morphological clarity

MTurk : Amazon Mechanical Turk

ON : Outline

PcB : Percentage change from a baseline

PConf : Perceived Confidence

TOT : Tip of the tongue

VE : Visualisations de l’explication

XAI : Explainable Artificial Intelligence

Avant-propos

L'autorisation de rédiger ce mémoire sous forme de deux articles complémentaires a été obtenue auprès de la direction du programme M.Sc. en gestion de HEC Montréal. Tous les coauteurs de ces articles ont donné leur accord afin qu'ils soient présentés dans ce mémoire. De plus, le comité d'éthique en recherche (CER) de HEC Montréal a approuvé, en juin 2020, la collecte de données pour ce projet de recherche.

Le premier article évalue, dans le cadre d'une étude en laboratoire, l'effet des différents types de visualisation de l'explication sur la charge cognitive et la certitude de l'utilisateur envers le système d'IA. Cet article a été présenté à l'occasion de la retraite virtuelle NeuroIS en juin 2021 et sera publié dans les actes de la conférence[1].

Le second article s'intéresse plutôt à l'effet de chaque type de visualisation sur la certitude de l'utilisateur envers le système d'IA dans le cadre d'une enquête en ligne auprès de plusieurs centaines de participants. Cet article a été soumis à la conférence ICIS 2021[2] et est toujours en évaluation au moment du dépôt de ce mémoire.

[1] Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Explainable Artificial intelligence (XAI), How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. *NeuroIS 2021 Proceedings*, 1–10.

[2] Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Towards user-centric AI explanations: increasing confidence in AI through visualizations of an AI system explanation. Soumis à La Conférence ICIS 2021.

Remerciements

Pour débuter, j'aimerais grandement remercier mes codirecteurs Pierre-Majorique Léger et Sylvain Sénécal pour leur implication et leurs judicieux conseils tant sur le plan académique que sur le plan professionnel. C'était un honneur de faire partie de l'équipe du Tech3Lab et cette expérience acquise va certainement m'être utile durant ma carrière qui débute enfin.

J'aimerais également remercier Théophile Demazure et Alexander Karran qui ont tous deux agi à titre de mentor tout au long de la réalisation de ce projet de maitrise. La réalisation de ce mémoire n'a pas été une tâche facile, mais avec votre aide, vos conseils et votre temps j'ai pu y arriver et remettre un travail dont je suis fier, et j'en suis très reconnaissant.

Je remercie aussi le Conseil de recherche en sciences naturelles et génie (CRSNG), Prompt et HEC Montréal pour m'avoir soutenu financièrement durant ces deux années de maitrise.

Merci aussi à mes amis de maitrise et du Tech3Lab, Amélie et Hugo, qui ont rendu mes deux années à HEC Montréal exceptionnelles. J'ai grandement apprécié votre soutien durant la maitrise et la réalisation du projet de recherche en ce temps de pandémie.

Merci aux membres de l'équipe du Tech3Lab qui ont contribué de près ou de loin à la réalisation de ce projet, dont Julianne, Shang Lin, David et tous les assistants de recherche.

Merci à mes parents, Martine et François, et mon frère, Mathieu, pour leur support constant durant ce long périple universitaire qui tire enfin à sa fin.

Finalement, merci à Alexandre qui m'a encouragé et soutenu durant mon établissement à Montréal et durant la réalisation de cette maitrise. Tu as su trouver chaque fois les bons mots pour me remonter le moral et m'encourager à persévérer et je t'en remercie énormément.

Chapitre 1
Introduction

1.1 Mise en contexte

Les systèmes d'intelligence artificielle (IA) sont maintenant omniprésents dans notre société. Les assistants personnels intégrés à nos téléphones intelligents, les agents de recommandations bonifiant les sites d'achat en ligne ou bien les systèmes de reconnaissance d'images bénéficiant aux voitures semi-autonomes ou aux systèmes de surveillance sont des exemples de systèmes d'IA bien présents dans notre quotidien. La sophistication, l'efficacité et la complexité de ces systèmes ne cessent de croitre leur permettant d'accomplir diverses tâches autrefois réservées à l'humain. Cependant, cette complexité grandissante accentue l'effet de boite noire qui caractérise les systèmes d'IA (Adadi & Berrada, 2018; Rudin, 2019; Wanner et al., 2021) les rendant, ainsi que leurs décisions, difficilement interprétables et compréhensibles pour l'utilisateur final. Bien qu'un grand nombre de systèmes d'IA ne puissent avoir de répercussions graves sur la vie d'humains, certains systèmes, tels que ceux utilisés en médecine ou par les services de police, peuvent avoir au contraire un impact considérable sur certains individus. En effet, des biais ont été découverts dans certains systèmes qui faisaient en sorte que certaines personnes étaient défavorisées par rapport à d'autres simplement à cause de leur genre ou de leur origine éthique (Angwin et al., 2016; Dastin, 2018). Ces biais relevant d'un manque de compréhension du fonctionnement interne des systèmes d'IA soulèvent un besoin en recherche pour tenter de comprendre et d'expliquer les processus camouflés dans cette boite noire permettant ainsi d'éviter de telles situations dans le futur et d'encourager une meilleure utilisation de ces systèmes (Abdul et al., 2018).

Rendre ces systèmes interprétables devient donc une nécessité plus la place qu'ils occupent est importante et plus les tâches qu'ils effectuent peuvent avoir un impact considérable sur la vie d'individus. L'interprétabilité d'un système d'IA est définie comme l'habilité à expliquer ou présenter son fonctionnement et ses décisions dans des termes compréhensibles pour un humain (Doshi-Velez & Kim, 2017). Un système interprétable peut permettre d'identifier plus facilement des sources de biais potentiels

dans le système en plus de s'assurer qu'il fonctionne comme prévu (Gilpin et al., 2018). La recherche en interprétabilité de l'IA (Explainable Artificial Intelligence [XAI]) cherche à créer des modèles et des techniques d'interprétabilité dans l'optique de répondre à ce manque de transparence dans les systèmes d'IA. La recherche en XAI vise ainsi à rendre ces systèmes plus interprétables, c'est-à-dire de les rendre plus compréhensibles pour l'utilisateur et plus équitables socialement (Barredo Arrieta et al., 2020). De plus, il a été démontré qu'une transparence accrue, véhiculée sous forme d'explication du fonctionnement et du résonnement du système, a un impact positif sur la confiance des utilisateurs envers le système (Cofta, 2009; Eiband et al., 2019; Glikson & Woolley, 2020; Lee & See, 2004; Meske & Bunde, 2020).

La confiance d'un utilisateur envers une technologie se base sur plusieurs aspects, dont la structure de navigation, l'apparence, la facilité d'utilisation et les croyances de l'utilisateur (Vance et al., 2008). Plus encore, la confiance des utilisateurs envers les systèmes d'IA dépend, elle, de facteurs plus spécifiques, comme la fiabilité, la performance, la sécurité, la confidentialité et finalement la transparence du système (Cofta, 2009; Fairclough et al., 2015; Glikson & Woolley, 2020; Yin et al., 2019). Cette confiance est importante puisque des utilisateurs n'ayant pas confiance en un système quelconque risquent tout simplement de ne pas l'utiliser (Lee & See, 2004). Au contraire, un niveau trop élevé de confiance envers le système peut amener à une mauvaise utilisation de ce dernier. Dans une telle situation, l'utilisateur risque de faire confiance aveuglément aux décisions du système même si certaines d'entre elles sont erronées (Lee & See, 2004). Autant un rejet du système par manque de confiance qu'une mauvaise utilisation du système par excès de confiance peut avoir des effets négatifs variables au niveau de la sécurité et de la profitabilité en fonction de l'ampleur et de l'importance des tâches accomplies par le système (Lee & See, 2004; Parasuraman & Riley, 1997). Ainsi, le développement de techniques d'interprétabilité augmentant la transparence des systèmes et aidant ainsi à l'atteinte d'un niveau de confiance adéquat risque de favoriser une meilleure collaboration entre les humains et l'IA (Mcknight et al., 2011).

La confiance que ressent un utilisateur envers une technologie n'est cependant pas la même que celle ressentie envers un autre individu. En effet, une personne est portée à

perdre confiance plus rapidement en un système plutôt qu'en un autre humain, même si les deux ont accompli la même tâche et ont fait la même erreur (Dietvorst et al., 2015; Jakesch et al., 2019). De plus, retrouver cette confiance dans le système prend plus de temps que de retrouver cette confiance dans l'autre individu (Glikson & Woolley, 2020; Nourani et al., 2020). Certains lancent même qu'il est erroné de parler de confiance entre un humain et un système d'IA, donnant une valeur au système équivalente à celle d'un humain (DeCamp & Tilburt, 2019; Luhmann, 2001; Vance et al., 2008). De plus, certaines recherches proposent qu'il soit seulement approprié de parler de confiance lorsque l'individu qui fait confiance est placé dans une situation de vulnérabilité et d'incertitude par rapport à la personne ou le système de confiance (Luhmann, 2001; Vance et al., 2008). Dans le contexte de cette recherche et en se basant sur cette dernière définition de la confiance, il serait ainsi plus approprié de parler de la certitude de l'utilisateur envers le système d'IA et ses décisions. La certitude[3] envers le système d'IA est définie comme étant une mesure de risque quant au degré que les utilisateurs soient certains qu'ils ont reçu les bonnes suggestions du système d'IA et s'ils considèrent que le système est fiable, c.-à-d. que le système fonctionne toujours correctement, fonctionnel, c.-à-d. que le système fait ce qu'il est censé faire, et utile, c.-à-d. que le système fournit une aide adéquate aux utilisateurs (Lankton et al., 2015; Wanner et al., 2021).

La recherche actuelle en XAI vise à développer des modèles mathématiques interprétables qui, ironiquement, sont difficilement compréhensibles pour un humain. Ces méthodes peuvent demander à l'utilisateur un effort cognitif important, risquant de le décourager d'utiliser l'explication. La conception d'explications présentées de manières compréhensibles pour l'utilisateur permettrait donc à ce dernier d'avoir une meilleure compréhension du système qu'il utilise. La théorie du Cognitive Fit (Vessey, 1991) tente de comprendre l'effet de la représentation de l'information présentée à un individu sur sa performance dans la réalisation d'une tâche donnée. La théorie propose que lorsque le format utilisé pour présenter l'information est adapté au type de problème à résoudre, l'individu crée une représentation mentale optimisée du problème qui en résulte en de meilleures performances. Au contraire, lorsque le format de l'information n'est pas adapté

[3] Dans les articles des chapitres 2 et 3, il est fait mention de « trust » et de « confidence » que nous avons traduits respectivement ici par « confiance » et « certitude ».

à la tâche à accomplir, un effort cognitif supérieur est requis, entrainant une performance inférieure puisque l'individu doit mentalement convertir l'information dans un format adapté à la résolution de la tâche (Adipat et al., 2011; Nuamah et al., 2020; Vessey, 1991). Bien que la théorie ait été développée en évaluant des représentations graphiques et tabulaires de données numériques, elle peut tout de même rester valide lors de l'évaluation de nouvelles représentations de l'information présentant des données qui ne sont pas exclusivement numériques (Adipat et al., 2011; Chen, 2017; Gillespie et al., 2018; van der Land et al., 2013).

La charge cognitive est définie comme étant la demande imposée par une tâche sur les ressources limitées de la mémoire à court terme d'un utilisateur (Wickens, 2008). Ainsi, une tâche complexe demandant une grande charge cognitive à l'utilisateur augmente les risques d'erreurs par rapport à une tâche simple demandant peu de ressources mentales (Palinko et al., 2010). Par ailleurs, l'effort cognitif est décrit comme étant le nombre de ressources mentales requis pour traiter la charge cognitive imposée par une tâche (Nuamah et al., 2020). Ainsi, une tâche imposant une charge cognitive importante à l'individu va lui demander un effort cognitif supérieur en comparaison à tâche imposant une charge cognitive faible. Il faut cependant noter que la charge cognitive ne peut être mesurée directement (Vieira, 2016). Cependant, elle peut en revanche être estimée à l'aide de trois types de mesures : des mesures de performances (ex. : rapidité, taux de réussite et taux d'erreurs), des mesures subjectives (ex. : questionnaires et entrevues) et des mesures physiologiques ou objectives. Pour ce dernier type, des mesures de nature électroencéphalographique (Nuamah et al., 2020) ou pupillométrique (Beatty, 1982; Palinko et al., 2010) permettent entre autres d'estimer la charge cognitive de manière objective. Diverses recherches ont d'ailleurs utilisé le nombre de fixations, le temps moyen de fixation ou bien la dilatation de la pupille comme moyen d'estimer la charge cognitive (Isabella et al., 2019; Niezgoda et al., 2015; Palinko et al., 2010; Vieira, 2016). Dans cette étude, la mesure de la dilatation de la pupille sera utilisée pour estimer la charge cognitive des utilisateurs faces à une tâche donnée. Ce phénomène caractérisé par la dilatation de la pupille lorsque l'utilisateur est confronté à une tâche lui demandant un effort cognitif considérable est appelé *task-evoked pupillary response* (Beatty, 1982).

1.2 Objectifs et questions de recherche

La complexité des systèmes d'IA actuels rend la conception de modèles d'interprétabilité compréhensibles pour l'utilisateur très difficile à accomplir. Ces systèmes doivent à la fois permettre d'expliquer le processus de décision interne du système d'IA tout en rapportant un niveau adéquat de détails dans l'optique de rester fidèle à ce processus. La recherche en XAI tente donc depuis quelques années de créer de tels modèles d'interprétabilité. En effet, dans le domaine des systèmes de reconnaissance d'images, on y retrouve diverses méthodes d'interprétabilité, dont celles étudiées dans ce mémoire, soit les méthodes Grad-CAM (Selvaraju et al., 2020) et Integrated Gradients (Sundararajan et al., 2017). Ces méthodes tentent de mettre en évidence sous forme de visualisation les zones de l'image qui ont le plus influencé le système à prendre une décision. La qualité de ces visualisations influence grandement l'efficacité de l'explication produite par le modèle d'interprétabilité (Sundararajan et al., 2019), en permettant entre autres à l'utilisateur d'évaluer la fiabilité et les biais potentiels dans le système d'IA (Gilpin et al., 2018; Selvaraju et al., 2020). Cependant, l'effet de ces visualisations de l'explication (VE) sur le développement de la certitude de l'utilisateur envers le système d'IA est inconnu. Plus encore, la charge cognitive demandée aux utilisateurs de ces visualisations est de plus inconnue.

Ce mémoire a donc comme objectif premier d'étudier l'effet de différentes VE sur la certitude de l'utilisateur dans le système d'IA. Plus encore, nous voulons observer l'effet de ces mêmes visualisations sur la charge cognitive de l'utilisateur, en plus d'observer la relation de cette charge cognitive avec la certitude de l'utilisateur. Les représentations des VE seront manipulées selon deux facteurs, soit l'adjacence de la visualisation (adjacente ou non adjacente) et la clarté morphologique de la visualisation (basse, moyenne, élevée). Les VE adjacentes se différencient des non adjacentes par leur proximité entre l'explication et l'image originale fournie en entrée au système. En effet, les VE adjacentes présentent l'explication directement sur l'image originale alors que les VE non adjacentes présentent l'explication séparément de l'image originale. La clarté morphologique d'une visualisation représente le degré qu'une visualisation présente des caractéristiques clairement délimitées en ajustant l'apparence de certaines données ou bien en supprimant

certaines données (p. ex., le bruit) afin de rendre la délimitation plus claire pour l'utilisateur (Sundararajan et al., 2019). Une VE ayant une clarté morphologique élevée va donc présenter des formes bien définies en éliminant au maximum le bruit contenu dans l'explication, alors qu'au contraire une VE ayant une clarté morphologique basse va présenter un niveau élevé d'information, donc un niveau élevé de précision, mais en ayant des formes floues et non délimitées. Voir le tableau 2.1 pour des exemples de VE représentant ces diverses caractéristiques. Ensuite, la certitude perçue des utilisateurs sera mesurée à l'aide de mesure explicite (questionnaire) alors que la charge cognitive sera estimée à l'aide de la dilatation de la pupille des utilisateurs. Cette étude tentera donc de répondre aux questions de recherche suivantes :

Article 1 :

- *Dans quelle mesure l'adjacence et la clarté morphologique de la visualisation de l'explication influencent-elles la charge cognitive requise par l'utilisateur utilisant ces explications ?*

- *Y a-t-il une corrélation entre la charge cognitive requise par l'utilisateur utilisant ces explications et la certitude de l'utilisateur envers le système d'IA et ses décisions ?*

Article 2 :

- *Dans quelle mesure la visualisation d'une explication de la décision d'un système d'IA affectera-t-elle la certitude de l'utilisateur envers ce système et ses décisions?*

- *Dans quelle mesure l'adjacence et la clarté morphologique de la visualisation de l'explication favorisent-elles ou dégradent-elles la certitude de l'utilisateur envers le système d'IA et ses décisions ?*

1.3 Information sur les articles

Ces deux articles ont été écrits dans l'optique d'être éventuellement combiné, après le dépôt de ce mémoire, en un seul article y regroupant les résultats des deux études, mais

bonifié d'une analyse sociodémographique des participants ainsi que d'une analyse des impacts possibles de ces facteurs sociodémographiques sur la certitude et la charge cognitive des utilisateurs. Cela étant dit, la revue de la littérature des deux articles inclus dans ce mémoire se base sur les mêmes fondations. La banque de stimuli utilisés a aussi été développée dans le but d'être réutilisée pour les deux études.

La banque de stimuli utilisée est composée de 100 images (2 par catégories) sélectionnées à partir de la base de données ImageNet (Jia Deng et al., 2009). Cette dernière regroupe plus d'un million d'images catégorisées et annotées. Ces images sélectionnées serviront ensuite à être analysé par le système de reconnaissance d'image. Puisque nous ne voulions pas que la nature de l'image influence les résultats, seulement des images neutres (aucune représentation pouvant choquer ou déranger) et non ambiguës ont été sélectionnées. Une image non ambiguë est une image dont une grande majorité de personnes s'entend sur l'objet qui y est représenté (Snodgrass & Vanderwart, 1980). De plus, le choix des catégories d'images s'est fait selon le travail de Snodgrass & Vanderwart (1980) qui ont défini une liste standardisée de 260 images de différents concepts. Ces concepts ont été choisis par les auteurs, entre autres, parce qu'ils sont identifiables sans ambiguïté et qu'ils représentent des concepts au niveau de base de la catégorisation (ex. : chien, vélo, cuillère).

Pour s'assurer de la non-ambiguïté des images sélectionnées, un panel de 18 juges, divisé en deux rondes (ronde 1 = 9 juges, ronde 2 = 9 juges), a été recruté. Ces juges devaient répondre à un questionnaire en ligne, où on leur demandait d'écrire pour chacune des images une étiquette composée d'un à deux mots qui selon eux décrivaient le mieux l'image. Cette première ronde nous a permis d'écarter huit images qui n'ont pas atteint un niveau d'accord satisfaisant. Ce même processus a été répété lors de la 2^{e} ronde avec huit nouvelles images. Cette fois-ci, toutes les images ont atteint un niveau d'accord satisfaisant nous confirmant que notre banque d'images sélectionnées était à la fois neutre et non ambiguë.

Un programme intégrant l'algorithme de reconnaissance d'images Xception (Chollet, 2017) ainsi que les algorithmes permettant de générer les six types de visualisations

étudiées (Selvaraju et al., 2020; Sundararajan et al., 2019) a ensuite été développé. L'algorithme Xception prend en entrée une image et retourne la classification qui lui semble la plus probable. Les 100 images sélectionnées à l'étape précédente ont ensuite pu être injectées dans le programme préalablement développé. Le produit résultant constitue le jeu de données complet servant de stimuli lors des deux études. Ainsi, ce jeu de données est composé de 50 paires d'images, où chaque pair appartient à une catégorie distincte, d'une classification du système d'IA par image et de six visualisations de l'explication du système d'IA par image.

1.3.1 Article 1

Le premier article a été soumis et présenté dans le cadre de la conférence virtuelle NeuroIS 2021 à Vienne (Hudon et al., 2021b). Ce papier présente les résultats préliminaires d'une étude évaluant l'effet des différents types de visualisations de l'explication sur la charge cognitive requise par l'utilisateur ainsi que sur la certitude de l'utilisateur envers le système. De plus, nous voulions observer s'il y a une corrélation entre la charge cognitive et la certitude de l'utilisateur.

Une expérience en laboratoire intra-sujet a été conduite en août 2020 auprès de 19 participants où on y mesurait la dilatation de leur pupille à l'aide d'un pupillomètre. La charge cognitive étant impossible à mesurer directement, ces mesures de dilatation de la pupille nous permettent d'estimer la charge cognitive requise lors de la réalisation d'une tâche. Chaque participant devait réaliser la même tâche. Cette dernière était décomposée en plusieurs itérations, chacune présentant des stimuli différents. Lors de chaque itération, une image, la classification de l'image par le système et la VE leur étaient présentées de manière successive. Les participants devaient ensuite évaluer s'ils ont la certitude que le système va bien classifier une image similaire à celle qui vient d'être présentée. Finalement, chaque itération était suivie d'une image neutre composée d'une croix noire sur un fond gris qu'on appelle référence. La charge cognitive a ensuite été calculée en calculant le pourcentage de variation entre le diamètre de la pupille durant l'affichage des stimuli et l'affiche de la référence.

Les résultats montrent que les visualisations adjacentes demandent une charge cognitive inférieure aux visualisations non adjacentes et qu'une clarté morphologique de niveau moyen demande une charge cognitive inférieure aux deux autres niveaux. Nous apportons d'ailleurs dans l'article une hypothèse expliquant ce dernier résultat surprenant. Nous observons aussi une corrélation négative entre la charge cognitive et le niveau de certitude des utilisateurs envers le système d'IA. Ainsi, une explication demandant une charge cognitive élevée à un utilisateur risque de faire diminuer son niveau de certitude envers le système. Cette étude présente des résultats préliminaires qui serviront à l'écriture éventuelle d'un article complet (non inclus dans ce mémoire).

1.3.2 Article 2

Le second article du mémoire a été soumis à la conférence ICIS 2021 (Hudon et al., 2021a). L'objectif de cette recherche intra-sujet, menée durant l'été 2020, est premièrement de vérifier si les VE ont un effet sur la certitude des utilisateurs envers les le système IA utilisé. Plus encore, nous voulons observer l'étendue de l'effet sur la certitude de l'adjacence de la visualisation ainsi que du niveau de clarté morphologique utilisé.

Pour étudier ce phénomène, une étude en ligne a été menée sur la plateforme Mechanical Turk (MTurk) d'Amazon auprès de 350 participants dont 206 nous ont fourni des réponses utilisables. Cette plateforme donne accès à un vaste bassin de participants prêt à répondre à des études en ligne en échange d'une compensation financière. Le questionnaire développé pour cette étude se composait de quelques questions sociodémographiques et d'une tâche principale. Cette dernière, décomposée en 50 itérations, présentait sur la même page une image provenant de notre jeu de données, la visualisation de l'explication, la classification du système et finalement l'image restante associée à la première image présentée. Les participants devaient ensuite évaluer s'ils ont la certitude que le système va bien classifier la deuxième image en se basant sur les informations qui lui sont présentées à l'écran.

Les résultats obtenus suggèrent que le type de visualisation a bel et bien un effet sur le niveau de certitude de l'utilisateur. En effet, les visualisations adjacentes ont un effet

positif sur la certitude de l'utilisateur supérieur à celui des visualisations non adjacentes. Plus encore, les visualisations ayant une clarté morphologique basse ont provoqué une certitude supérieure aux deux autres niveaux de clarté morphologique. Par contre, en analysant davantage ces résultats, on observe que les visualisations ayant une clarté morphologique basse ont un effet plus important sur la certitude de l'utilisateur, seulement lorsqu'elles sont non adjacentes. Cette étude contribue à la littérature en XAI en apportant des recommandations pour la représentation de techniques d'interprétabilité. En considérant l'utilisateur final lors de la conception d'interface IA et de techniques d'interprétabilité, la certitude de l'utilisateur envers ces systèmes risque de s'améliorer, favorisant ainsi une meilleure collaboration entre ces deux partis.

1.4 Contributions et responsabilités personnelles

Les deux études présentées dans ce travail ont été réalisées avec le Tech3lab. Ma contribution pour chacune des étapes nécessaires à la réalisation de ces projets est présentée sous forme de pourcentage dans le **Tableau 1.1** ci-dessous.

Tableau 1.1. Contributions et responsabilités personnelles pour chaque étape du processus.

Étape du processus	Contributions et responsabilités
Revue de la littérature	Rédaction de la revue de la littérature comprenant les principales recherches et construits — **100 %** • Les coauteurs des articles m'ont cependant beaucoup conseillé à cette étape, par exemple en m'invitant à lire certains textes. Choix des différents algorithmes utilisés — **75 %** • Soutenu dans ce processus par l'équipe de recherche

Design expérimental	Rédaction de la demande au CER et des demandes de changement — **90 %** • Je me suis chargé de remplir les demandes au CER. L'équipe du Tech3Lab s'est assuré que les demandes étaient remplies en bonne et due forme. Création du jeu de données comportant les images, leur classification et leurs explications respectives — **100 %** • J'ai sélectionné chacune des images. • J'ai créé le programme intégrant l'algorithme de reconnaissance d'image et les algorithmes générant les visualisations. • J'ai créé et distribué le questionnaire en ligne permettant aux juges de valider chacune des images. Conception du questionnaire en ligne utilisé pour la 2e étude — **100 %**
Recrutement des participants	Recrutement des participants — **50 %** • Je me suis chargé du recrutement des juges servant à la validation des stimuli

	• Pour la première étude, le recrutement s'est fait à partir du panel de HEC Montréal sous ma supervision. • Pour la seconde étude, j'ai collaboré avec l'équipe du tech3lab pour le recrutement directement à partir de la plateforme MTurk.
Extraction et transformation des données	Extraction des données des questionnaires et tests — **100 %** Tri des données utilisables de celles devant être rejetées — **100 %** Transformation des données dans un format permettant l'analyse — **100 %**
Analyse des résultats	Analyses statistiques — **90 %** • Utilisation des logiciels SAS et JASP pour l'élaboration de modèle statistique permettant l'analyse des résultats • Statistiques descriptives • Vérification de la conformité des résultats • Génération de graphiques et figures illustrant les résultats obtenus.

	• J’ai été soutenu par l’équipe de recherche pour analyser les résultats de la validation des stimuli par les juges. Interprétation des résultats — **100 %**
Rédaction	Écriture des articles du mémoire — **100 %** • Les articles ont été améliorés et bonifiés à la suite de l’aide et des commentaires des coauteurs. Écriture du mémoire — **100 %**

Chapitre 2
Explainable Artificial intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence[4]

Antoine Hudon, Théophile Demazure, Alexander Karran, Pierre-Majorique Léger, Sylvain Sénécal

Tech3Lab, HEC Montréal, Montréal, Canada

Abstract

Explainable Artificial Intelligence (XAI) aims to bring transparency to AI systems by translating, simplifying, and visualizing its decisions. While society remains skeptical about AI systems, studies show that transparent and explainable AI systems result in improved confidence between humans and AI. We present preliminary results from a study designed to assess two presentation-order methods and three AI decision visualization attribution models to determine each visualization's impact upon a user's cognitive load and confidence in the system by asking participants to complete a visual decision-making task. The results show that both the presentation order and the morphological clarity impact cognitive load. Furthermore, a negative correlation was revealed between cognitive load and confidence in the AI system. Our findings have implications for future AI systems design, which may facilitate better collaboration between humans and AI.

Keywords: Explainable Artificial Intelligence · XAI · Cognitive load · Confidence in AI · Explanation · Visualization · Cognitive Fit Theory

---

[4] Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Explainable Artificial intelligence (XAI), How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. *NeuroIS 2021 Proceedings*, 1–10.

2.1 Introduction

Artificial intelligence (AI) algorithms are growing in complexity, sophistication, and accuracy, aided by increasing data volumes. However, as these algorithms grow in complexity and performance, they become less interpretable and opaquer, making the decisions these machine learning models form hard to comprehend and explain (Rudin, 2019).

Research in explainable AI (XAI) seeks to address the problems associated with a lack of transparency in AI systems. XAI methods attempt to decipher which components of the AI model or system are perturbed to create decisions through visualizations or descriptions of discriminative mechanisms. These methods aim at giving the user access to the AI's decision chain and discriminative variables. Implementing XAI methods within AI systems will help create more transparent and fair systems, thereby helping users become more aware of a system's behavior and support a richer collaboration between humans and AI (Gilpin et al., 2018). Unfortunately, XAI currently focuses more on creating mathematically interpretable models than on information presentation, neglecting the systems' final users (Abdul et al., 2018), that is, the ones that will benefit from the explanations of those models. Therefore, these interpretability methods may impose a higher cognitive load, making them hard to understand and apply to real-world problems. This study attempts to bridge this gap in XAI by assessing different visualizations of an AI system's explanation, with the aim of providing human-centered explanations without compromising the faithfulness of the AI system's visualizations.

Using the Cognitive Fit theory, we propose to investigate if the visualization of explanations of an AI decision system can affect a user's cognitive load and confidence in the AI system. Specifically, we present in this paper the preliminary results from a study designed to assess two presentation-order methods and three AI decision visualization attribution models, referred to as explanation visualizations (EV), to determine each visualization's impact upon a user's cognitive load and confidence in the system by asking participants to complete a visual decision-making task. For this research, we focus only on identifying the effects of different EVs on a user's cognitive load and confidence in an AI context.

Cognitive Fit theory was initially developed to assess the impact on user's performance when viewing numerical data presented in tabular versus graph format in a symbolic and spatial task (Vessey, 1991). This theory has been applied in several studies within other research domains, adapting it for new information presentation formats (e.g., online rating systems (Chen, 2017), and database structure representations (Bizarro, 2015)) using numerical, textual (Adipat et al., 2011), as well as visual data (Brunelle, 2009). The data at the core of explanations produced by interpretable models are numerical. However, these explanations are complex for a human to process, requiring translation using shapes and colors to represent their values. This study evaluates which EV type, presentation-order method, and AI decision visualization attribution models result in the best cognitive fit paired with a spatial task. The EV type that results in a better cognitive fit will allow us to make recommendations for future AI systems design, contributing a step on the road towards explainable AI that is more accessible to human mental capacities.

2.2 Hypothesis Development and Research Model

The research model **(Figure 2.1)** posits that the effect of the EV's adjacency and morphological clarity will affect a user's perceived confidence in the system's judgments. We hypothesize that this effect is mediated by the cognitive load imposed by the visualization on the user's working memory. The rationale behind these relationships is described below.

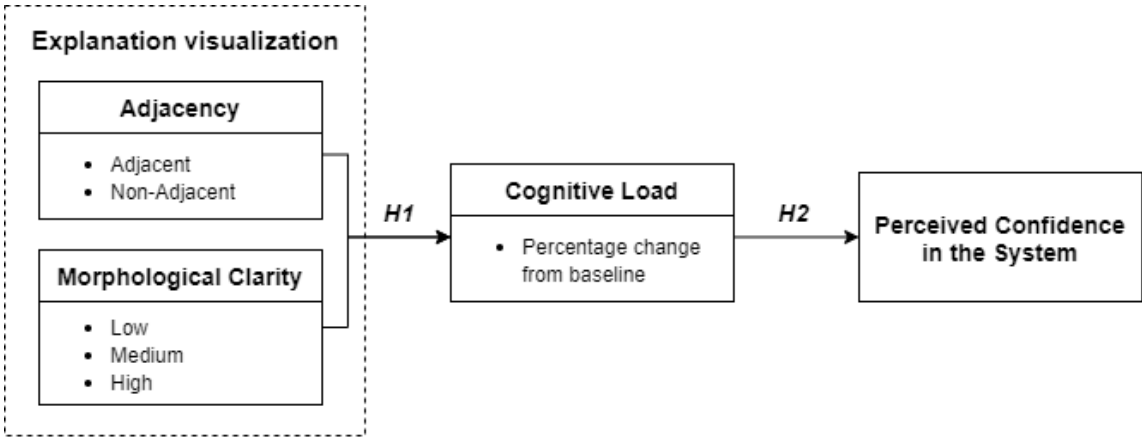


Figure 2.1. Research model

Cognitive Fit theory proposes that the level of congruence between task and information presentation mediates task performance, such that, while solving a problem, an individual creates a mental representation of the problem based on the information presented (Vessey & Galletta, 1991). This mental representation's complexity and usefulness are dependent on the user's working memory capacity (Adipat et al., 2011; Goodhue & Thompson, 1995; Vessey, 1991). Thus, additional cognitive effort is required in the presence of incongruence between task and the format of the information presented to help complete the task (Nuamah et al., 2020), requiring the individual to mentally transform that information into a format suitable for accomplishing the task, resulting in reduced performance (Adipat et al., 2011).

We posit that specific EVs will result in a better cognitive fit, reducing cognitive load. Cognitive load is defined as the demand imposed by a task on the user's working memory (Wickens, 2008). Therefore, a task requiring significant mental resources is more likely to prompt more user errors than a task requiring less cognitive resources, resulting in less perceived effort and greater cognitive fit (Nuamah et al., 2020).

H1a: Adjacent visualizations will result in a lower cognitive load.

An adjacent EV displays the explanation directly upon the original image by coloring the areas in different colors to indicate their data values (Dennis & Carte, 1998). In comparison, a non-adjacent EV is presented with the same explanation data but separated from the image. However, with this latter method, there is a loss in correspondence between the explanation and the original image, requiring significantly greater cognitive effort from the user (Sundararajan et al., 2019).

Furthermore, we posit that morphological clarity (MC) will play a role in mediating the effect of adjacency upon cognitive workload. Morphological Clarity (MC) represents the degree to which a visualization displays clearly delimited features by adjusting the appearance or removing specific data (e.g., noise) to help make the delimitation clearer for the user. For this research, High MC EVs are faithful to the MC definition by having clear delimited features without noise, providing greater clarity for the user but at the cost of faithfully depicting the model's behavior (Sundararajan et al., 2019). On the other hand,

Low MC EVs are more faithful to the model's behavior as they precisely illustrate the relevant image's areas, at the cost of being more cluttered. Moreover, Low MC EVs may cause humans to ignore the explanation by giving them too much information to process (Sundararajan et al., 2019).

H1b: High MC EVs will result in lower cognitive load.

H1c: Adjacent-High MC EVs will result in lower cognitive load.

Providing explanations of system behaviors as a form of transparency positively impacts the development of trust (Bigras et al., 2019) and confidence in new technology (Cofta, 2009; Eiband et al., 2019; Glikson & Woolley, 2020; Lee & See, 2004; Meske & Bunde, 2020). However, according to DeCamp & Tilburt (2019), it would be wrong to talk about trust in AI since this implies a loss of human agency. More appropriate, perhaps, is to speak of developing confidence in AI. Confidence in AI has been defined as a measure of risk or surety that AI systems provide correct suggestions and if users consider the system to be reliable (Wanner et al., 2021). More formally, confidence in this context is the goal of reducing the epistemic uncertainty associated with AI decisions with regard to accuracy, data provenance, and temporal qualities. To investigate this aspect, we formed the following hypothesis:

H2: EVs imposing a lower cognitive load on the user will result in higher perceived confidence in the system.

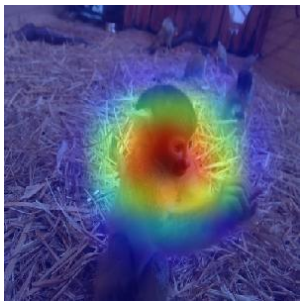
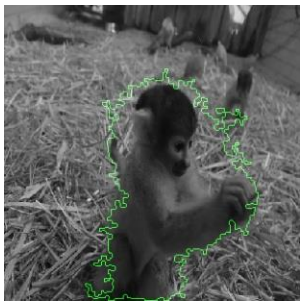
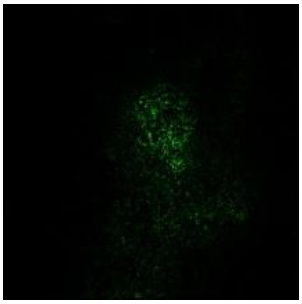
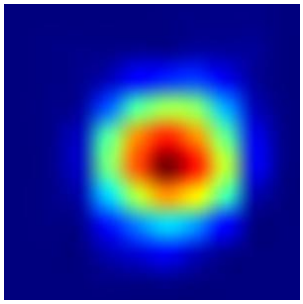
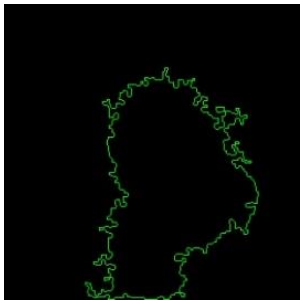
2.3 Method

2.3.1 Experimental Design

We designed a 2x3 within-subject factorial design to investigate the effects of adjacency and MC of AI-EVs on the user's cognitive load and perceived confidence. The first factor considers the representation's adjacency with two levels: EV with (adjacent) or without (non-adjacent) image background. The second factor considers the MC of the EV with three levels: low-cloud of points (CP), medium-heatmap (HM), and high-outline (ON)**.** CP faithfully depicts the model's attributions by highlighting all the image's pixels that positively impact the model's classification toward a result. HM is less precise than CP,

showing only the stimuli image's prime focus and does not have precisely delimited features. ON visualization draws only the most essential zones of the image used in the classification but at the cost of pixel-level precision. CP and ON-EVs were both implemented using the Integrated Gradients method (Sundararajan et al., 2017, 2019), and the HM-EV using the Grad-CAM class activation function (Selvaraju et al., 2020). See **Table 2.1** for examples of each type of EV used in this study.

Table 2.1. Examples of EV for the classification "Monkey."

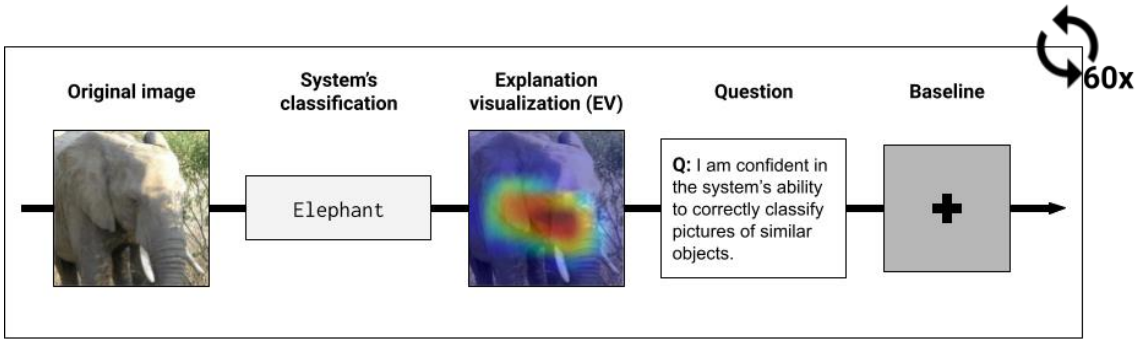
		Morphological Clarity (MC)		
		Low (Cloud of points)	Medium (Heatmap)	High (Outline)
Adjacency	Adjacent			
	Non-adjacent			

Note: Image selected from the ImageNet dataset (Jia Deng et al., 2009).

2.3.2 Experimental Procedure

19 participants (24.9 ± 8.3 years old, 10 males) took part in the study, all signed consent following the HEC ERB ethical approval. The task **(Figure 2.2)** consisted of a spatial task repeated over 60 randomized trials. Each trial involves a series of elements displayed on screen in the following order: (1) original image (e.g., image of an elephant), (2)

classification of the image given by the AI system[5] (e.g., "Elephant"), (3) the AI EV (e.g., an overlay IA explanation onto the original image) and (4) a perceived confidence question (e.g., confidence in the system). Participants were asked to rate their agreement with the following statement using a 7-point Likert scale ranging from "Strongly disagree" to "Strongly agree" and "I am confident in the system's ability to classify pictures of similar objects correctly." A baseline image was finally shown at the end of each trial for 1s. Each participant saw all six types of visualizations ten times. To measure pupil dilation, we used the Tobii x60 eye tracker.



Note: The EV used in this figure is Adjacent and has a medium MC (Heatmap). Images selected from the ImageNet dataset (Jia Deng et al., 2009).

Figure 2.2. Task's design.

2.3.3 Stimuli Images and Visualizations

Stimuli image categories were chosen based on Snodgrass & Vanderwart (1980), who defined a standardized set of 260 illustrations of different concepts. From these, 60 concepts were selected as image stimuli. We selected one image for each category from the ImageNet dataset (Jia Deng et al., 2009), using neutrality (no shocking or disturbing depiction), simplicity, and unambiguity as criteria. Also, no human subjects are present in any of the selected images. These selection criteria allowed us to control the impact of affective mediators on the users-AI confidence. For two rounds of image verification, a panel of 18 judges labeled each image. Tests for inter-rater reliability produced $\kappa = 0.838$ and $\kappa = 0.879$ for the 1st and 2nd verification rounds, respectively, showing a high degree

[5] The Xception (extreme inception) (Chollet, 2017) algorithm which comes with pre-trained weights on the ImageNet dataset was used to classify the images.

of agreement. Eight images were removed, and new images were added between the two rounds due to a lack of agreement.

2.3.4 Calculating Cognitive Effort

Change in Pupil dilation, when a user is faced with a task requiring a high cognitive effort, is referred to as the Task-evoked pupillary response (Beatty, 1982). We used pupil diameter to estimate the user's cognitive effort required to process each EV. We computed the average percentage change from a baseline (PcB) in this preliminary analysis for each participant and each EV. We used the percentage change of pupil diameter rather than the raw pupil size variation due to inter-participant variance (Attard-Johnson et al., 2019).

2.4 Results

We performed a repeated measures ANOVA for the dependent variable PcB, with both Adjacency and MC as factors. The results show a statistically significant main effect of Adjacency ($F(1, 189) = 20.99$, $p < .001$, $\eta^2 = 0.02$), MC ($F(2, 378) = 28.03$, $p < .001$, $\eta^2 = 0.05$) as well as the interaction between both factors ($F(2, 378) = 22.32$, $p < .001$, $\eta^2 = 0.04$). Post hoc comparisons (Bonferroni corrected) reported that Adjacent EV ($M = 1.27$, $SD = 6.36$) results in lower PcB than non-adjacent EV ($M = 3.05$, $SD = 6.52$), providing support for H1a. Concerning MC, post hoc comparisons showed that Medium MC EV ($M = 0.20$, $SD = 6.25$) results in lower PcB than both Low ($M = 2.90$, $SD = 6.25$), and High ($M = 3.40$, $SD = 6.55$), providing no support for H1b. The results **(Figure 2.3)** indicate no significant difference between the three adjacent visualizations. Therefore, H1c is also not supported. Unexpectedly, non-adjacent-Medium MC EV ($M = -0.35$, $SD = 5.63$) resulted in a negative PcB as well as the lowest PcB of all visualizations, which may explain the result of H1b.

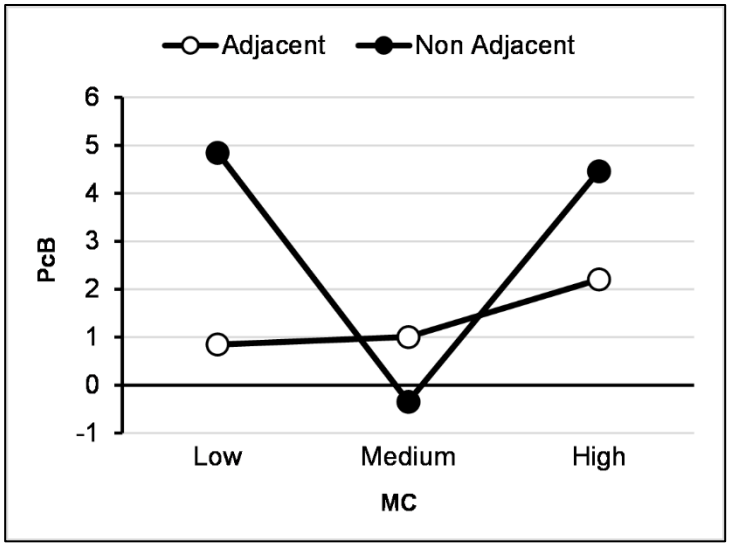


Figure 2.3. PcB for each EV type

With regard to the relationship between PcB and perceived confidence in the system, the results indicate a significant negative correlation between the two variables, $r(1136) = -0.12$, $p < .001$, providing support for H2. However, this requires further analysis. Results of perceived confidence for each EV type can be seen in **Figure 2.4.**

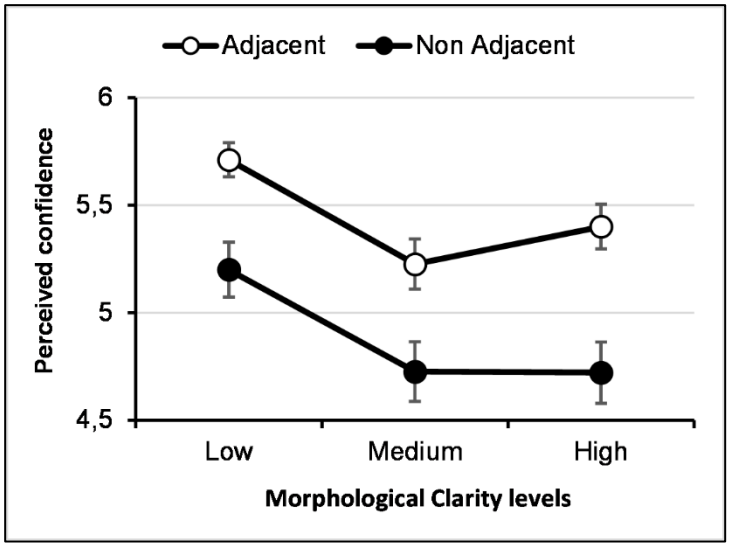


Figure 2.4. Perceived Confidence for each EV type

2.5 Discussion and Conclusion

The preliminary results indicate that adjacent EVs result in lower PcB. Potentially, this implies that the cognitive effort required to process and understand the AI's EV through

adjacent presentation is significantly lower than non-adjacent EV, given a reduced need to mentally associate the EV to the original image as the association is already made implicit within the EV itself. Indeed, this effect was posited by Dennis & Carte (1998), who stated that combining adjacent presentation-order with a spatial task may lead to faster and more accurate decision making, resulting in a better cognitive fit. Regarding the low PcB reported for non-adjacent-medium MC EV, this type of EV is very abstract, potentially making it difficult to process the original image's target object. We posit that this proved challenging for participants to identify precise forms and associate them with the original image, allowing for a snap judgment and lower cognitive workload through a disengagement effect. This may account for why this EV resulted in one of the lowest scores of perceived confidence.

Moreover, it would appear that a significant negative correlation between cognitive load and perceived confidence does not necessarily result in low confidence when mediated by adjacency. For example, the non-adjacent-low MC EV, which reported the highest PcB, resulted in the highest perceived confidence for non-adjacent visualizations. We propose that the high density of information presented in this EV potentially helped users identify the target object forms by reducing epistemic uncertainty. In that, extraneous but useful information allowed the user to swap from cognitive processing to perceptual processing to understand the model's behavior. Indeed, there is a small body of researchers in human factors investigating this new concept of epistemic uncertainty and how it affects man-machine teaming (Tomsett et al., 2020). Additionally, adjacent-Low MC EV also resulted in the highest perceived confidence, contradicting (Sundararajan et al., 2019), who stated that users had shown a high over low MC preference. Regarding the negative correlation reported between PcB (inferred cognitive load) and perceived confidence, future work will aim to explain this effect in more detail using further pupillometry measures.

This study investigated the relationships between various types of EVs used to explain an AI system's output and user cognitive load and their effects on a user's confidence in the system. The results indicate that design choices related to EVs can positively impact a user's confidence in AI systems by reducing epistemic uncertainty. In this study, users manifested a preference for visualizations providing precision rather than simplicity,

where they can easily associate the target with the explanation. Overall, our results suggest that the careful consideration of cognitive fit theory, information presentation adjacency methods, and explanation visualizations containing low morphological clarity applied to AI interface and task design may help accelerate confidence between a user and an AI decision support system. However, more work is required to tease apart each factor's role to determine how best to increase confidence in users of AI decision support systems.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, *2018-April*, 1–18. https://doi.org/10.1145/3173574.3174156

Adipat, Zhang, & Zhou. (2011). The Effects of Tree-View Based Presentation Adaptation on Mobile Web Browsing. *MIS Quarterly*, *35*(1), 99. https://doi.org/10.2307/23043491

Attard-Johnson, J., Ó Ciardha, C., & Bindemann, M. (2019). Comparing methods for the analysis of pupillary response. *Behavior Research Methods*, *51*(1), 83–95. https://doi.org/10.3758/s13428-018-1108-6

Bigras, É., Léger, P.-M., & Sénécal, S. (2019). Recommendation Agent Adoption: How Recommendation Presentation Influences Employees' Perceptions, Behaviors, and Decision Quality. *Applied Sciences*, *9*(20). https://doi.org/10.3390/app9204244

Bizarro, P. A. (2015). Effect of different database structure representations, query languages, and task characteristics on information retrieval. *Journal of Management Information and Decision Science*, *18*(1), 27–52.

Brunelle, E. (2009). The moderating role of cognitive fit in consumer channel preference. *Journal of Electronic Commerce Research*, *10*(3).

Chen, C.-W. (2017). Five-star or thumbs-up? The influence of rating system types on users' perceptions of information quality, cognitive effort, enjoyment and continuance intention. *Internet Research*, *27*(3), 478–494. https://doi.org/10.1108/IntR-08-2016-0243

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *2017-Janua*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Cofta, P. (2009). Designing for Trust. In *Handbook of Research on Socio-Technical Design and Social Networking Systems* (Vol. 731, Issue 9985433, pp. 388–401). IGI Global. https://doi.org/10.4018/978-1-60566-264-0.ch026

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *The Lancet Digital Health*, *1*(8), e390. https://doi.org/10.1016/S2589-7500(19)30197-9

Dennis, A. R., & Carte, T. A. (1998). Using Geographical Information Systems for Decision Making: Extending Cognitive Fit Theory to Map-Based Presentations. *Information Systems Research*, *9*(2), 194–203. https://doi.org/10.1287/isre.9.2.194

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goodhue, D. L., & Thompson, R. L. (1995). Task-Technology Fit and Individual Performance. *MIS Quarterly*, *19*(2), 213. https://doi.org/10.2307/249689

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPRW.2009.5206848

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Meske, C., & Bunde, E. (2020). Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12217 LNCS* (pp. 54–69). Springer International Publishing. https://doi.org/10.1007/978-3-030-50334-5_4

Nuamah, J. K., Seong, Y., Jiang, S., Park, E., & Mountjoy, D. (2020). Evaluating effectiveness of information visualizations using cognitive fit theory: A neuroergonomics approach. *Applied Ergonomics*, *88*(June 2019), 103173. https://doi.org/10.1016/j.apergo.2020.103173

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10, April 2017*, 141. https://doi.org/10.1145/1743666.1743701

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *34th International Conference on Machine Learning, ICML 2017*, *7*, 5109–5118. http://arxiv.org/abs/1703.01365

Sundararajan, M., Xu, S., Taly, A., Sayres, R., & Najmi, A. (2019). Exploring principled visualizations for deep network attributions. *CEUR Workshop Proceedings*, *2327*.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, *1*(4), 100049. https://doi.org/https://doi.org/10.1016/j.patter.2020.100049

Vessey, I. (1991). Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature. *Decision Sciences*, *22*(2), 219–240. https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., & Zschech, P. (2021). White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems. *Nternational Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global*, 0–9.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Chapitre 3
Towards user-centric AI explanations: increasing confidence in AI through visualizations of an AI system explanation[6]

Antoine Hudon, Théophile Demazure, Alexander Karran, Pierre-Majorique Léger, Sylvain Sénécal

Tech3Lab, HEC Montréal, Montréal, Canada

Abstract

AI systems are consistently growing in sophistication and complexity, at the cost of interpretability for the user. Explainable Artificial Intelligence (XAI) aims to bring transparency to AI systems' internal processes and decisions, helping to improve humans' confidence in AI. This study assesses three AI decision visualization attribution models manipulating the morphological clarity and two presentation-order methods to determine each visualization's impact upon the Human-AI confidence relationship. We utilized information presentation methods and visualizations delivered through an online experiment to explore confidence in AI by asking participants to complete a visual decision-making task. Results show that the adjacency of the explanation visualization further influences user confidence in AI and that low morphological clarity has a positive impact on Human-AI confidence only in situations of non-adjacency of the explanation. Our findings contribute to XAI research by bringing more human-centered AI decision explanations visualizations, facilitating human-AI collaboration.

Keywords: Explainable Artificial Intelligence · XAI · Confidence in AI · Explanation · Visualization · Cognitive Fit Theory

---

[6] Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Towards user-centric AI explanations: increasing confidence in AI through visualizations of an AI system explanation. *Soumis à La Conférence ICIS 2021*.

3.1 Introduction

Artificial intelligence and machine learning algorithms are growing in sophistication and accuracy to automate an ever-increasing array of tasks. However, these algorithms' rapid growth in complexity and performance makes them more opaque and less interpretable. Thus, the decisions as well as the process taken by these AI systems to make said decisions are more challenging to understand and explain for its final users (Rudin, 2019).

Within the field of artificial intelligence (AI), interpretability can be defined as "the ability to explain or to present in understandable terms to a human" (Doshi-Velez & Kim, 2017). It represents the degree to which a human can understand the basis of a decision or the degree to which a human can predict with a high degree of consistency a machine learning model's result. The more interpretable the model, the easier it is to understand why certain decisions or predictions have been made. This relationship also acts to engender confidence between the user of a decision support engine and the machine learning algorithm that forms and then provides those decisions.

In this work, we report on a study that aims to assess the effect of interpretability techniques and visualizations on user confidence in the AI system providing decisions. Current research in the field focuses more on creating mathematically interpretable models, neglecting the human who uses these explanations (Abdul et al., 2018). Taking a user-centric approach to interpretability and confidence-building may better facilitate collaboration between humans and AI decision systems and assuage societal concerns around the use of AI. This study attempts to bridge this gap in interpretability by providing human-centered explanations using various techniques without compromising the faithfulness of the AI visualization.

Research in explainable AI (XAI) seeks to create interpretability models and methods in the hope of addressing the problems associated with a lack of transparency in AI systems, therefore making them more explainable and fairer (Barredo Arrieta et al., 2020). Some methods developed for image recognition systems attempt to highlight which components of the AI model or system are perturbed to create decisions. Visualizations or descriptions of discriminative mechanisms are then used to represent perturbed components of the

system for users. These methods are essential to identify sources of potential bias in the training data and to ensure that algorithms perform as expected (Gilpin et al., 2018). Providing explanations of system behaviors as a form of transparency has been shown to have a considerable positive impact on developing confidence in new technology (Cofta, 2009; Eiband et al., 2019; Glikson & Woolley, 2020; Lee & See, 2004; Meske & Bunde, 2020). Implementing XAI methods within HCI design for systems that include AI as decision support will help users become more aware of a system's behavior and support a richer collaboration between humans and AI (Gilpin et al., 2018). Similarly, XAI research is also critical in domains such as transportation, finance, security, legal, and medicine, where AI decisions have the potential to impact human lives.

Improving congruence between algorithmic decisions and human perceptions concerning those decisions is a serious challenge within advanced and critical machine learning applications (Doshi-Velez & Kim, 2017; Rudin, 2019). However, the complexity, recursivity, and high degree of nonlinearity of current machine learning systems make it arduous to dissect the decision process and, thus, provide straightforward and understandable explanations for a human to process. To tackle this challenge, current approaches in XAI create post-hoc algorithmic and mathematical methods to help explain initially opaque models. Moreover, researchers in the field have recently begun to investigate how to visually represent the AI decision process to produce explanations that are intelligible for humans (Sundararajan et al., 2019), highlighting the importance of visualization during the interpretation process.

According to a recent call for research in the area made by Abdul et al. (2018), societal need is driving the rise in research interest investigating XAI, stating that there is a perception of bias in AI decision-making systems which now affect both users of those systems and more generally those for whom decisions are made. To address this need for research, we present a study investigating the effects of design choices of an AI model's explanation visualization (EV) on the level of confidence between a human and an AI system. As a basis for the study presented here, we formulated the following research questions: "To what extent will the EV of an AI system's decision affect the confidence of a user?" Furthermore, to highlight if aspects of information presentation further affect

user confidence, “To what extent does presentation order and visualization technique promote or degrade user confidence in the AI system and its decisions?"

Using the Cognitive Fit theory, we present a study designed to answer these questions by utilizing various EVs of AI decisions output to manipulate morphological clarity and two presentation-order methods to investigate adjacency. The impact of each is assessed through an online within-subject experiment where each participant was required to complete a simple decision-making task using an AI system's outputs. We investigate how users' confidence in future predictions of the system is affected by each type of EV. The results show that the visualization's adjacency and a low morphological clarity in specific circumstances has a positive impact on human-AI perceived confidence.

3.2 Previous Work

3.2.1 Trust and Confidence in AI

Trust in a specific technology, such as an AI system, affects the value-added of using the technology after its adoption (Mcknight et al., 2011). Moreover, a user who trusts a specific technology is more likely to explore and use its features (Mcknight et al., 2011). Therefore, it has become essential to consider the trust relationship between user and technology when developing and implementing an AI system in order to smooth its acceptance within the workplace. This relationship has a few proven antecedents, such as navigational structure and visual appeal, ease of use, and the national culture of the user (Vance et al., 2008). Furthermore, trust in an AI system is also dependent on more specific factors such as privacy, security, reliability, stated and perceived accuracy, and transparency (Cofta, 2009; Fairclough et al., 2015; Glikson & Woolley, 2020; Yin et al., 2019). As described by Lee & See (2004), trust within a human-AI partnership is the attitude that an agent, such as an AI system, will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. Insufficient trust placed in a system may result in distrust and disuse of the system, whereas too much trust may result in over-reliance in the system, whereby the AI takes away from human agency and the ability to make decisions. Depending on the scale, the importance, and the impact of the tasks and decisions taken by the AI system, misuses or disuses of the system by its users

can lead to safety and profitability problems (Lee & See, 2004; Parasuraman & Riley, 1997). While trust mediates how much reliance humans are willing to place on AI systems, trust also mediates how much humans rely on each other (Lankton et al., 2015). However, humans lose confidence in AI systems more rapidly than in other humans, even if both parts make the same mistake (Dietvorst et al., 2015). Thus, it can be stated that trust in AI systems tends to decrease rapidly as a function of the number of errors over the duration of interaction and that the restoration of trust in AI systems or tools requires an undetermined but greater amount of time (Glikson & Woolley, 2020; Nourani et al., 2020). Moreover, studies have shown that humans display greater trust towards other human agents than in AI agents, even though both perform the same tasks and make identical mistakes (Glikson & Woolley, 2020; Jakesch et al., 2019).

However, according to DeCamp & Tilburt (2019), it is an error to speak of a user-AI trust relationship. Referring to trust in AI as a concrete psycho-affective relationship would imply that the system belongs to the same category of human agents that can be trusted, such as a physician. Currently, for an AI system, the thoughts, motives, and actions, which comprise the human psycho-affective framework involved in developing and allocating trust, go beyond its technical and mechanical capacities. Furthermore, moving forward, the development of AI will have moral impacts in the long term as the performance of AI improves, given that human capacities, in terms of technical accuracy, speed, and judgment, may well prove inferior to that of AI systems resulting in a blanket of distrust irregardless of evidence proving their accuracy and worth.

Additionally, some researchers suggest a game theoretic approach where it is only possible to talk about trust when the truster is in a situation of vulnerability and uncertainty with the trustee and when the consequences of a betrayal of trust are more significant than the benefits sought (Luhmann, 2001; Vance et al., 2008). In the context of the current research study, it would therefore be more appropriate to speak of developing confidence rather than trust in AI systems. Confidence in AI is defined as a measure of risk as to how sure users are that they received the correct suggestions by the AI system and if they consider the system to be reliable, i.e., the system consistently operates properly,

functional, i.e., the system does what it is supposed to do, and helpful, i.e., the system provides adequate help for the users (Lankton et al., 2015; Wanner et al., 2021).

3.2.2 Cognitive Fit Theory

Cognitive Fit (CF) theory (Vessey, 1991) offers a means to understand how design choices of AI decision-making visualizations affect human cognitive performance and potentially confidence in the system. This theory proposes that a congruence between the task and the problem representation format in an individual mental representation of the problem results in reduced cognitive effort and superior task performance. The complexity and usefulness of this mental representation are dependent on the user's working memory capacity (Adipat et al., 2011; Goodhue & Thompson, 1995; Vessey & Galletta, 1991). Therefore, additional cognitive effort is required when there is a "misfit" between the task at hand and the information format used to complete the task (Vessey, 1991). The individual must mentally transform that information into a format that enables them to accomplish the task, resulting in reduced performance (Adipat et al., 2011). Nuamah et al. (2020) also proved this relation between cognitive fit and cognitive effort using electroencephalography metrics.

CF theory was first proposed to assess the effect on performance of numerical data presented in a tabular versus graph format paired with a symbolic and spatial task. This theory has been applied in several studies over the years, within other research domains, and using multiple other information presentation formats (Adipat et al., 2011; Chen, 2017; Gillespie et al., 2018; van der Land et al., 2013). CF theory can also be used to support the theoretical basis of this study, where numerical data are translated into visualizations using colors and shapes for human usage. In this respect, the visualization of the AI decision process can potentially help provide a stronger cognitive fit between the task and the information required to complete that task in human-AI collaboration contexts, where understanding the AI decision process is of importance.

The result from designing tasks using this process is greater efficiency and effectiveness, manifested as increased accuracy and speed in problem-solving. Therefore, applying this process to the presentation and visualization of AI decision reasoning may significantly

impact the level of confidence engendered between a human and AI system. Moreover, a visualization method that has a stronger cognitive fit would help users to see the reasoning behind the AI system's prediction, consequently helping to avoid situations of over-confidence or lack of confidence resulting in a misuse or disuse of the AI system (Lee & See, 2004). Furthermore, forthcoming results obtained from a similar study, using the same stimuli and AI algorithm (Hudon et al., 2021a), reported a negative correlation between cognitive load and the level of confidence in the AI system resulting from the use of an EV. These results further support that improving cognitive fit between the task and the EV type may positively impact a user's confidence in the system.

3.2.3 Progress in XAI

Recent advances in XAI have made it possible to produce explainable models and learning methods, especially in image recognition systems where multiple techniques have been developed (Barredo Arrieta et al., 2020). Techniques such as Grad-CAM (Selvaraju et al., 2020) and integrated gradient (Sundararajan et al., 2017) aim to show relationships between the inputs and outputs, focusing on the processing of information in the models. Studies have investigated the potential effect of interpretability and explanations on humans. Poursabzi-Sangdeh et al. (2018) showed that transparency in a model helped users simulate the model's prediction. Others showed that visualizations have a large influence on the effectiveness of the explanations (Sundararajan et al., 2019), that they can help discriminate between classes more accurately, help reveal a model's trustworthiness, and help identify biases in datasets or models (Gilpin et al., 2018; Selvaraju et al., 2020). However, it is unknown what impact these visualization techniques have on developing confidence between human and AI systems to our knowledge.

3.2.4 Hypothesis Development and Research Model

The research model **(Figure 3.1)** posits that an EV's design choices will affect a user's confidence in the system. More precisely, we hypothesize that an EV's adjacency and morphological clarity as well as the interaction between both will have an effect on said confidence. The rationale behind these relationships will be developed below.

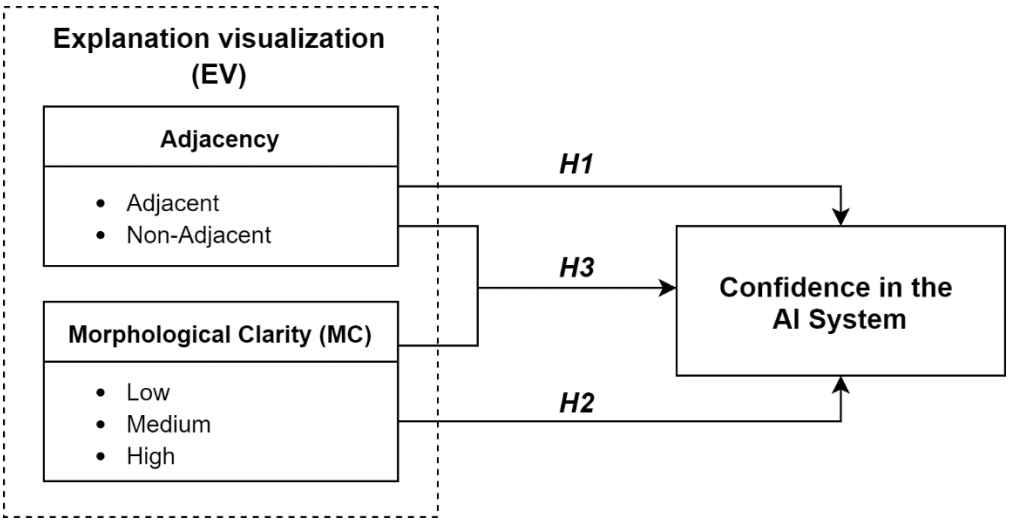


Figure 3.1. Research model

Adjacency is a form of presentation order in which an adjacent EV is presented with its explanation data displayed directly upon the original image by coloring the areas in different colors to indicate their data values, whereas a non-adjacent EV is presented with the same explanation data but separated from the image (Dennis & Carte, 1998). However, non-adjacent EV requires more cognitive effort to process since there is a loss in correspondence between the explanation and the visualization (Sundararajan et al., 2019). For spatial tasks, similar to the one used in the study presented here, (Dennis & Carte, 1998) showed that adjacent representations led to faster and more accurate decisions. They state that adjacent visualizations should help a decision-maker associate areas of a visualization image with its data, simplifying the complex relationship between them, leading to faster, more accurate decisions. We posit that this adjacency relationship will hold true for newer spatial tasks in which the visualization and data presented represent decisions already made by an AI and where the user of the AI is asked if he has confidence in future AI classification abilities.

H1: Adjacent EVs will result in a higher confidence in the AI system.

Furthermore, we posit that Morphological Clarity will also have an effect upon user confidence in the AI system. Morphological Clarity (MC) represents the degree to which a visualization displays clearly delimited features by adjusting the appearance or removing specific data (e.g., noise) to help make the delimitation clearer for the user

(Sundararajan et al., 2019). This study uses three levels of MC (i.e., low, medium, and high). First, low MC EVs are faithful to the model's behavior as they allow the user to precisely identify at a pixel level what areas of the image are relevant. However, this precision makes the EV more cluttered, preventing the user from having a clear overarching view of the image. Alternatively, High MC EVs are more faithful to the MC definition by having clear form features while reducing the noise due to excess information. High MC EV provides greater clarity for the user but at the cost of faithfully depicting the model's behavior. On the other hand, low MC EV might cause humans to ignore or misuse the explanation altogether by giving them too much information to process (Müller et al., 2021; Sundararajan et al., 2019). In contrast, high MC EVs tend to be easier to comprehend by reducing the user's cognitive load. Furthermore, this concept of information overload (i.e., too much available information or too much high-quality information) has been shown to decrease the user's decision effectiveness and can even make the user less confident about their decision (Keller & Staelin, 1987; Müller et al., 2021).

H2: High MC EVs will result in higher confidence in the AI system.

H3: Adjacent-High MC EVs will result in higher confidence in the AI system.

3.3 Method

3.3.1 Experimental Design

We designed an online experiment to test our hypotheses. Specifically, we used a 2x3 within-subject factorial design to examine the effects of adjacency and MC of AI EV on confidence between the user and the AI system. The first factor considers the adjacency of the representation with two levels: EV with (adjacent) or without (non-adjacent) image background, the second factor considers the MC of the EV with three levels: low-cloud of points (CP), medium-heatmap (HM), and high-outline (ON). We considered CP EV to be low MC since it faithfully depicts the model's attributions by highlighting the image's pixels that positively impact the model's classification. However, CP also displays much information that can be useless to the user. We considered HM EV to be a low MC since it is less precise than CP, showing only the model's focal point on the stimuli image. HM

also does not delimit its shapes enough to be considered a High MC visualization. Finally, ON EV draws only the most essential zones of the image used in the classification, therefore providing a high MC, but at the cost of pixel-level precision. See **Table 3.1** for EV examples and the Generation of Explanations section for detailed information about these factors' implementation.

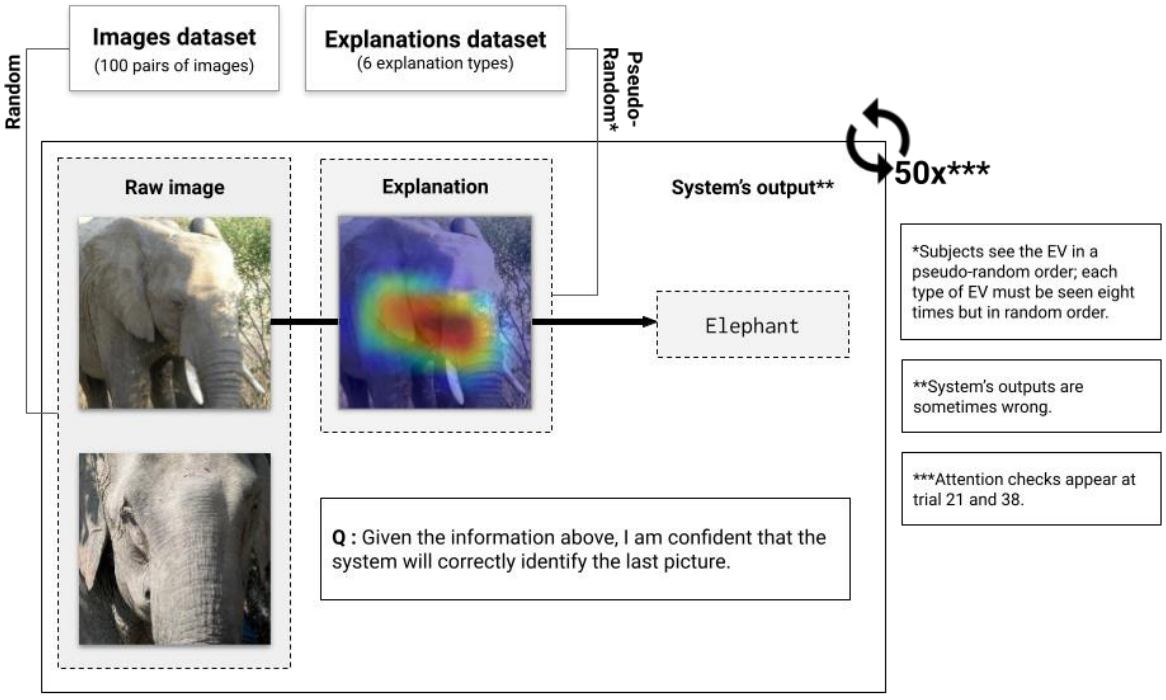
3.3.2 Subjects

Amazon Mechanical Turk (MTurk) was used to recruit participants. Based on recommendations by Jia et al. (2017), the participant pool was screened for North American residency, to prevent linguistic difficulty. In the study, we restricted ourselves to local explanations of the model as they are deemed to be accessible for novices. This selection criterion avoids confounding factors due to the potential machine learning expertise of participants. A total of 350 North American participants took part in the survey, of which 206 (Avg. age = 37.87 ± 10.51, Male = 123) provided usable data. 61 participants considered themselves to be an AI or Data expert based on Mohseni et al. (2018) criteria. To improve data quality, data were removed for participants who failed any of the two attention checks, did the survey multiple times, or did not correctly submit their survey. The attention checks verified participant attention to the stimuli by displaying black squares instead of the images. The study was approved by the ethics committee of our institution, and each participant provided informed consent.

3.3.3 Experimental Procedure

Once recruited, participants were directed to an online survey on Qualtrics. Upon beginning the study, participants were provided instructions and an example of a trial task. For each trial, several elements were presented on screen **(Figure 3.2)**: The image analyzed by the system; the system's EV; the system's output; and a related but unclassified image. In this case, the system's output is the classification of the image predicted by the system; classifications were presented irrespective of correctness. Classification mistakes were evenly distributed among all EV types. For our 206 participants, we found on average per EV type 350 classification mistakes (STD = 10.61) compared to 1259 correct classification (STD = 14,15). Participants were then asked to

rate their agreement with the following statement using a 7-point Likert scale ranging from “Strongly disagree” to “Strongly agree”: “Given the information above, I am confident that the system will correctly identify the next picture.” The decision-making task consisted of 50 trials per participant, with two trials consisting of attention checks. Stimulus material was presented randomly, with one label associated with a pair of similar images. In this context, similar images belong to the same class, therefore having the same label (e.g., two images of a car). A participant could not see the same pair of images more than one time. EV types were also presented in a random order for each trial of the task. The combination label-explanation type was also randomized amongst participants to avoid confounding effects due to the image label, object, or prediction. Therefore, a participant could see the image of an elephant paired with an adjacent-medium MC EV, and another could see the same image paired with a non-adjacent-low MC EV. After 50 trials, all six different EV types were seen and judged eight times by the participant. After completing the main task, participants were asked to answer demographic questions regarding their age, gender, and education level.



Note: The EV used in this figure is Adjacent and has a medium MC (Heatmap). Images selected from the ImageNet dataset (Jia Deng et al. 2009)

Figure 3.2. Task’s design.

3.3.4 Selection of Stimuli

Subsets of stimulus images were selected from the ImageNet dataset (Jia Deng et al., 2009), a publicly released image dataset with 1.2 million quality-controlled categorized images and associated human annotations. When building the stimulus image dataset neutrality and unambiguity were used as the criterion. With these criteria in mind, straightforward categories of image and annotation pairs such as “Dog” instead of “Golden Retriever” or “Elephant” instead of “African elephant” were chosen to reduce confusion and present images as “platonic” classes. Image categories were chosen based on the work of Snodgrass and Vanderwart (1980), who defined a standardized set of 260 illustrations of different concepts. These concepts were chosen based on three criteria: [1] They are unambiguously picturable, [2] they include exemplars from the widely used category norms of Battig and Montague (1969), and [3] they represent concepts at the basic level of categorization. Unambiguity, as described by Snodgrass and Vanderwart, is the degree to which subjects will show consensus about the name to give the object. From these 260 concepts, 100 were selected to represent the categories of the images used within the stimulus dataset.

ImageNet organizes images according to the WordNet hierarchy (Miller, 1995), where each concept is described by one or multiple names called a "synonym set" or "synset". The following criteria were established to select the stimulus images: [1] The image’s synset must contain the name of one of the 100 previously selected concepts, [2] No human subject is present in the image, and [3] the image must be neutral (no shocking or disturbing depiction). In total, 200 images were selected, giving two images per concept. Image ambiguity, familiarity, and complexity were measured and used as control variables in the final analysis.

3.3.5 Validation of Stimuli

To ensure that the selected images are an unambiguous representation of the concepts outlined by Snodgrass and Vanderwart, we used a panel of 18 judges to validate the images used for stimulus presentation. The validation was split into two rounds, and judges were split into six groups of three such that in Round one 9 (3x3) judges and Round

two 9 (3x3) judges were utilized to ensure three judges independently coded each stimulus image. To evaluate the stimulus images, judges were asked to complete an online survey that displayed images to a screen and then write a label composed of one or two words to describe the image. Each group of three judges would assess between 66 and 68 images. This ensured that all images were labeled by 3 judges. Furthermore, judges were also prompted to write a spontaneous description to minimize subjective bias. For each picture, the judges could also specify if: (1) They did not know the object (DNO), (2) if they knew the object but did not know its name (DKN), (3) and if they knew the name of the object but it was momentarily irretrievable (tip of the tongue [TOT]) (Snodgrass & Vanderwart, 1980).

In order to assess if a judge description (label) can be accepted or rejected, we specified the following criteria: 1) It is the same as the concept name (expected: bee, label: bee), 2) it is a synonym of the concept name (expected: airplane, label: plane), 3) it is a more precise term than the concept's name (expected: bear, label: polar bear), and 4) it is listed in the nondominant list of name of the concept (Snodgrass & Vanderwart, 1980) (expected: alligator, label: crocodile). An image is replaced if it matches one of these criteria: 1) At least two out of three judges put a rejected label or checked DKO, DKN, or TOT, or 2) at least two out of three judges use a nondominant name to describe the image.

The first round of stimuli verification composed of 3 groups of 3 judges produced an average Cohen's Kappa of 0.838, which is considered a strong agreement rate (Viera et al., 2005). However, eight images were replaced that did not meet the agreement criteria. We conducted a second round of stimuli validation to include the new images and nine new judges, which produced an average Cohen's Kappa of 0.879. We did not replace any images after this second round of verification, given the high rate of inter-rater agreement.

3.3.6 Choice of the Algorithm

To classify the stimulus images and act as the automated system in the study, we chose the Xception (extreme inception; Chollet 2017) algorithm which comes with pre-trained weights trained on the ImageNet dataset. Developed by a team at Google, Xception is a deep learning algorithm that relies heavily on prior effort done in the areas of

Convolutional Neural Networks (CNN), Xception has been proven to be a very effective, compact (88MB), and accurate (0.790) algorithm for computer vision problems (Chollet, 2017). We applied the trained Xception algorithm to the validated stimulus image. Output classifications, in the format of a sysnet-ID and label, were then compared to the "platonic" class of image. For example, if we are looking for the platonic class of "Cat" and the algorithm returned the synset label of "Cat" or "Tiger Cat," we considered the classification acceptable as they both represent the class of a cat. After training, we obtained a reasonable classification rate of 0.76, which is close enough to the algorithm's current maximum accuracy rate.

3.3.7 Generation of Explanation Visualizations

In order to actualize the concepts of adjacency and MC during the experimental task, we utilized several visualization and presentation methods to provide the participant with AI EV. The low and high MC visualizations were both implemented using the Integrated Gradients method (Sundararajan et al., 2017, 2019). These visualizations highlight in green the areas of the stimulus image that positively impact the model's classification. The medium MC visualization was implemented using the Grad-CAM class activation function (Selvaraju et al., 2020), this method displays a heatmap that highlights activated regions important for the classification of the image.

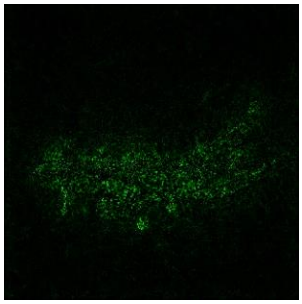
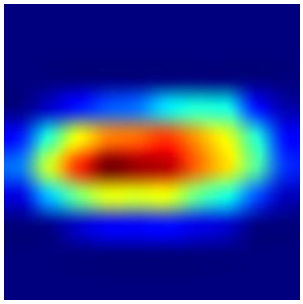
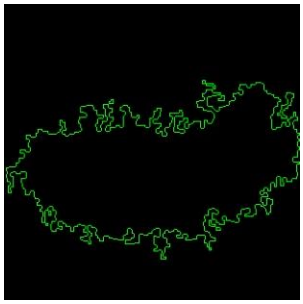
Adjacency representations were implemented using the Grad-CAM class activation visualization overlaid upon the original image and the Integrated Gradients visualization overlaid upon a grayscale version of the original image ensuring that the attribution colors did not blend with those of the image as recommended by Sundararajan et al. (2019). Non-adjacent representations were implemented by showing a black background under each visualization. In total, six combinations of explanation visualizations per image were generated **(Table 3.1)**.

3.3.8 Analysis

Using SAS 9.4, we performed repeated-measures ANOVA for the dependent variable Perceived Confidence (PConf) with both Adjacency, MC, and the interaction of those two as within-subject factors. PConf is a discrete variable taking values from 1 to 7, where 7

represents the highest value of confidence between the user and the system. Adjacency is a categorical binary variable (i.e., adjacent and non-adjacent), and MC is a categorical ordinal variable (i.e., low, medium, and high). Since a classification's validity can greatly impact a user's perceived confidence in the system, we decided to use the variable Classification as a covariate in the model. Classification is a Boolean variable indicating if the model's classification is correct or not. All post hoc comparison t-tests have been Bonferroni corrected.

Table 3.1. Example of each type of EV for the classification "Airplane".

		Morphological Clarity (MC)		
		Low (Cloud of points)	Medium (Heatmap)	High (Outline)
Adjacency	Adjacent			
	Non-adjacent			

Note: Image selected from the ImageNet dataset (Jia Deng et al., 2009).

3.4 Results

The results show a statistically significant main effect of Adjacency ($F(1, 205) = 246.96$, $p < .001$), and MC ($F(2, 410) = 5.69$, $p = .004$). Additionally, we observed a significant interaction between MC and Adjacency ($F(2, 410) = 6.10$, $p = .003$). The covariate variable Classification also has a main effect on PConf ($F(1, 205) = 768.13$, $p < .001$).

Post hoc comparisons of simple effects **(Table 3.2)** reported that adjacent EV (M = 5.01, SD = 0.06) results in higher confidence than non-adjacent EV (M = 4.55, SD = 0.06; t(205) = 15.71, p < .001), providing support for H1. Concerning MC, post hoc comparisons showed that low MC (Cloud of Points) EV (M = 4.85, SD = 0.06) results in higher PConf than both medium MC (Heatmap) (M = 4.77, SD = 0.06; t(410) = 2.36, p = .019), and high MC (Outline) (M = 4.73, SD = 0.06; t(410) = 3.27, p = .001), providing no support for H2, as the results highlight that the effect is the opposite of what we had hypothesized. The comparison between good classifications (M = 5.27, SD = 0.06) results in higher PConf than bad classification (M = 4.29, SD = 0.06; t(205) =27.72, p < .001).

Table 3.2. Post hoc comparisons Adjacency, MC, and Classification on Confidence.

Factor (1)	**Factor (2)**	**Mean difference (1 – 2)**	**DF**	***t***	***p***
Adjacent	Non-adjacent	0.46	205	15.71***	< .001
Low MC (Cloud of Points)	Medium MC (Heatmap)	0.08	410	2.36*	0.019
Low MC (Cloud of Points)	High MC (Outline)	0.12	410	3.27**	0.001
Medium MC (Heatmap)	High MC (Outline)	0.03	410	0.91	0.363
Good classification	Bad classification	0.98	205	27.72***	< .001

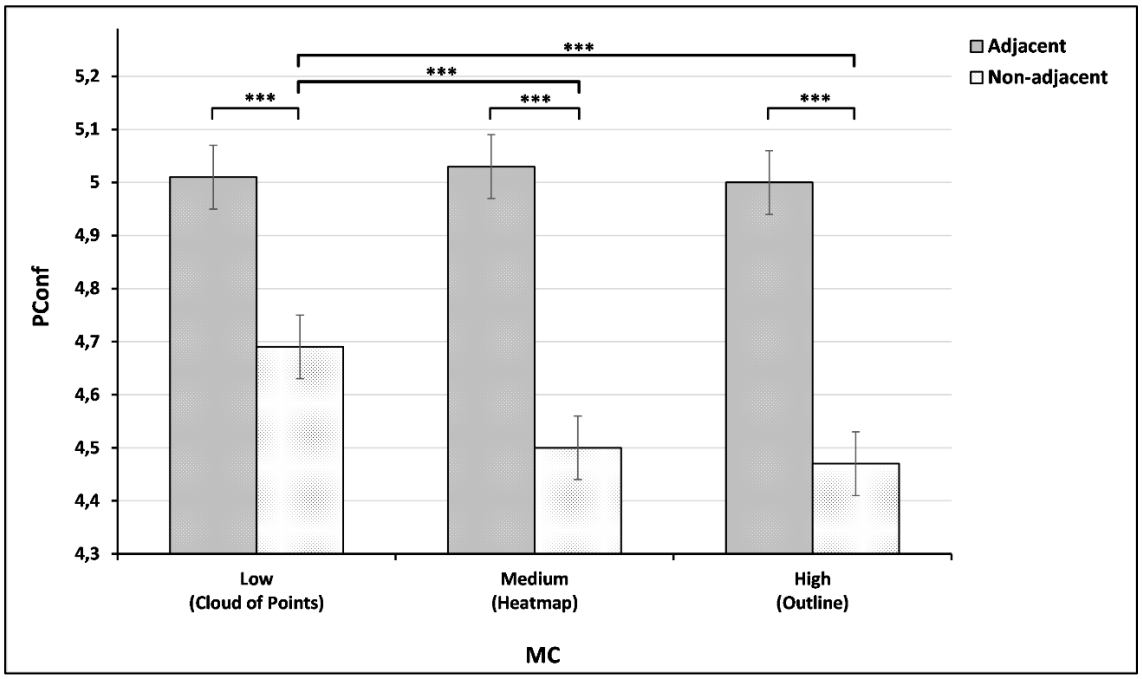
*p < .05, **p < .01, ***p < .001

As shown in **Table 3.3**, adjacent-high MC EV did not significantly differ from the two other adjacent representations. Therefore, H3 is also not supported. Surprisingly, non-adjacent-low MC EV (M = -0.25, SD = 0.09) results in a significantly greater positive impact on PConf than both non-adjacent-medium MC EV, and non-adjacent-high MC EV. **Figure 3.3** illustrates that all three adjacent EV types have a significantly greater positive impact on PConf than their non-adjacent equivalent. It also shows that low MC EV has a significantly greater positive impact on trust than medium and high MC EV.

Table 3.3. Post hoc comparisons of the interaction between Adjacency and MC on Confidence.

Adjacency	MC level (1)	MC level (2)	Mean difference (1 – 2)	DF	*t*	*p*
Adjacent	Low (Cloud of Points)	Medium (Heatmap)	-0.024	410	-0.48	0.632
	Low (Cloud of Points)	High (Outline)	0.009	410	0.18	0.857
	Medium (Heatmap)	High (Outline)	0.033	410	0.66	0.511
Non-adjacent	Low (Cloud of Points)	Medium (Heatmap)	0.191	410	3.82***	< .001
	Low (Cloud of Points)	High (Outline)	0.223	410	4.44***	< .001
	Medium (Heatmap)	High (Outline)	0.032	410	0.63	0.530

*p < .05, **p < .01, ***p < .001



*p < .05, **p < .01, ***p < .001

Figure 3.3. Comparison of adjacent and non-adjacent visualizations based on their MC from low to high.

3.5 Discussion

The results indicate a clear difference when comparing EVs, such that adjacency strongly influenced user confidence. Looking at this significant difference in more detail, we found that adjacent EVs had a greater positive effect on user confidence than non-adjacent EVs, confirming our hypothesis. However, contrary to our initial postulation in which we surmised that a high MC would provide more than sufficient information and accuracy of representation to positively influence user confidence, low MC EVs resulted in the highest confidence between all three levels of MC. Furthermore, looking at the results from the interaction of both adjacency and MC, the findings indicate no clear leading method of information presentation; in that, all three adjacent EVs have an approximately equivalent impact on user confidence. Moreover, the results indicate that in situations of non-adjacency, there is a significant difference between low and medium MC as well as low and high MC, where the latter resulted in lower confidence in the system.

3.5.1 The Impact of Adjacency on User Confidence

We first hypothesized that adjacent visualizations would have a greater significant impact on user confidence than non-adjacent EVs, as adjacency is posited to provide a superior cognitive fit with the task at hand in terms of providing visual scan patterns for pertinent information potentially leading to higher confidence through a reduction in epistemic uncertainty. The results indicate, at least in part, that our hypothesis holds true and that adjacency does indeed have a significant impact on user confidence. This effect can potentially be explained by the lower cognitive effort required to correspond between the highlighted areas of the explanation with the original image. Users were able to promptly identify which areas of the image influenced the AI system in its decision, helping them compare AI classification reasoning with their own. Indeed, Dennis and Carte (1998) posited that adjacent visualizations led to more accurate decision-making in less time when the user is faced with a spatial task, resulting in a better cognitive fit. However, non-adjacent EVs have the advantage of aiding the user to inspect the details of the explanation without the distraction of the underlying image or to review the original image to verify what the attributions highlight in order to form a fresh opinion (Sundararajan et al., 2019). Moreover, non-adjacent EVs ask the user to continuously gaze back and forth between

the original image and the EV. Thus, in terms of cognitive fit, we suggest that users need to mentally associate the original image zones with the AI decision visualization, creating a more complex mental map requiring more cognitive resources.

3.5.2 The Conditional Effect of MC on User Confidence

We further hypothesized that high MC visualizations would have a more significant positive effect on user confidence. This visualisation technique focuses and highlights important regions of the images being classified, reducing the noise in how the explanation visualization represents the network's results, thus diminishing the information load required. However, our results surprisingly indicate the opposite to be true, showing in this case that low MC EV had a greater positive effect on user's confidence than high and medium MC EVs. An alternate interpretation for this relationship could be that low MC EV's include more information, even if that additional information would ordinarily be considered noise. Furthermore, this supplementary information may, implicitly increase the perception of transparency in the model and positively affect user confidence. This interpretation aligns with work by Brunk et al. (2019), who showed that the interpretability of black-box algorithms are perceived to be more transparent and trustworthy if additional information is present.

However, the adjacency comparisons **(Table 3.3)** indicate no significant difference between levels of MC when explanation data are presented adjacently. As can be seen in the table, in our case the main significant variance between levels of MC is between non-adjacent low MC with high and medium MC EVs. This implies that low MC EV has a greater positive effect on confidence but only in situations of non-adjacency. With these conditions in mind, and in light of our interpretations, the results strongly imply that users require more information about the explanation when the requirements of adjacency are not met. Moreover, we interpret the difference between non-adjacent MC levels to be a perceptual and processing disassociation between explanation visualization and the subject matter of the input image, in that while participants could make a certain sense of the target objects form in the case of the non-adjacent-low MC EV, the non-adjacent-medium and high MC EVs were too abstract to identify clear forms and associate them with the input image (see **Table 3.1**, showing that non-adjacent-low MC EV allows the

shape of the airplane to be distinguished contrary to the two other non-adjacent EVs). Therefore, we cannot conclude that, overall, low MC EV has a greater effect on user confidence.

Overall, the results appear to indicate that in a task context requiring no precise knowledge (i.e., identifying a simple object in an image), confidence in an AI system can be improved with explanations, providing there is adjacency between the EV and the original image and furthermore, that the EV highlights areas of the original image that correspond to the user's perceptual understanding of the task, regardless of the EV's level of precision. In contrast, when users with greater knowledge in a particular subject are tasked with identifying specific patterns requiring precision (e.g., detecting diabetic retinopathy), they are much more critical about the type and characteristics of the EV used (Sundararajan et al., 2019). However, further research is required to investigate the relationship between the task and the explanations in more detail.

3.5.3 Contribution to Theory and Implications for Practice

The theory of Cognitive fit has been used in previous research as a framework to explain the impact of different information presentation methods paired with various tasks on decision-making performance. Studies from various fields have provided validation of this theory such as computer system development (Shaft & Vessey, 2006), geographic information systems (Dennis & Carte, 1998), and e-commerce (Brunelle, 2009; Chen, 2017). Our study incorporates cognitive fit theory into XAI by assessing which combination of presentation method and type of explanation visualization provided a better "fit" and positively influenced confidence using a spatial task. The results reported in this study provide evidence showing that adjacent EVs facilitate the cognitive fit of a problem and its representation for the user, in that they help users effectively utilize their working memory while performing the task by presenting the information in a more "understandable" and structured format than non-adjacent EVs.

Studies have shown that visualized explanations of a system or model have a positive impact on confidence (Eiband et al., 2019; Meske & Bunde, 2020) and that the presence of transparent design further increases this impact (Kizilcec, 2016; Weitz et al., 2019).

We add to this design template for confidence by showing that the choice of EV type in terms of its adjacency and MC may also modulate the impact on user confidence. Furthermore, designs that impede the level of cognitive fit between interpretability and the task, such as in the case of non-adjacent EV, may have a null or even negative impact on confidence compared to cases where there is no explanation transparency. Implying that not all explanations of AI decision processes, be they visual or not, positively impact user trust in AI decision systems. Therefore, when designing AI decision tasks and interpretability interfaces the aim should be to achieve a high level of congruence between the problem-solving task and the problem representation to align with the user's mental representation and knowledge of the task.

3.5.4 Limitations & Future Work

The cognitive distance or the degree of similarity between two images representing the same label was not considered in the current study. Consequently, some pairs of images may have appeared similar (e.g., two images of an elephant's face) and others rather different (e.g., dogs of different breeds). A user presented with an AI system's classification could potentially be more confident that the system will correctly classify a similar image than if he/she is presented with a dissimilar image. This difference in similarity between pairs of images could result in a bias in the results. Therefore, future work should control the degree of similarity between pairs of images. We see two opportunities to address this challenge. Firstly, one should adopt an algorithmic perspective and compute similarity between pictures and secondly, one could task a new panel of judges to assess the perceived visual distance between images.

In this study, we only performed an analysis on the effect of Adjacency and MC on the user's confidence without including in the model the sample's demographic data. These data could have a moderating effect on the user-AI confidence (e.g., the confidence of an AI expert might be less influenced by the different EV representation than an AI novice). Moreover, we did not consider the potential impact on the confidence of the individual's propensity to trust AI. Some people are more likely to trust technology than others, and it would have been interesting to see the extent of this effect on confidence, if any.

Finally, in this study we did not control for algorithm type and design. Further work should concentrate on replicating our results for different algorithms to examine if our observations generalize. It has been demonstrated that even if the explanation algorithm is model-agnostic, intrinsic differences in the image recognition system design can lead to differences in predictive features (Ribeiro et al., 2016). Thus, future research could investigate the impact of design choices in black-box models on the interpretability explanation and its impact on the human-algorithm interaction.

While most of the work on interpretability focuses on algorithmic approaches to create interpretable models, this study shows that human-centered EVs of such algorithms can provide an avenue for new research, where explanations are designed to fit the users' capabilities and the context of use. Future studies could investigate the impact of the highlighted areas of an image to determine if the highlighted areas that show a similar thinking process to humans have a greater impact on confidence than others.

Concerns related to the lack of confidence between humans and AI are not only relevant to the field of HCI. They have far-reaching societal implications, as AI systems take critical decisions that can impact human lives. To continue developing better collaboration between human and AI agents, there is a need to focus on developing models that are interpretable, transparent, and present information in a way that is compatible with human cognitive processes (European Commission, 2021). In this regard, research in the field could be expanded to include the concurrent use of neurophysiological assessment methods such as eye-fixation related potential (Palinko et al., 2010). Studying the active brain while testing new interpretation methods would add an extra dimension that allows the testing of cognitive fit against the cognitive load of adjacent and non-adjacent low, medium, and high morphological clarity visualizations of new interpretation methods.

3.6 Conclusion

This study investigated the relationship between various types of explanation visualization of an AI system's output and user confidence in the decisions that system makes. The results show that visualization and information presentation design choices

have the potential to positively impact a user's confidence in the AI system. By drawing upon Cognitive fit theory (Vessey, 1991), we show that in this case, adjacency and the level of Morphological Clarity of the visualization has a conditional relationship upon user confidence. An explanation visualization overlaid upon the original image, no matter the level of precision of the visualization, results in a better fit between the decision task and the explanation resulting in higher confidence. On the other hand, users showed a preference for visualizations providing precision in situations of non-adjacency of the explanation visualization. Overall, our results strongly suggest that the careful consideration and application of Cognitive fit theory, adjacency methods and explanation visualizations containing low morphological clarity to AI interface and task design may help improve the confidence relationship between a user and an AI decision support system.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018-April, 1–18. https://doi.org/10.1145/3173574.3174156

Adipat, Zhang, & Zhou. (2011). The Effects of Tree-View Based Presentation Adaptation on Mobile Web Browsing. MIS Quarterly, 35(1), 99. https://doi.org/10.2307/23043491

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58(December 2019), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. Journal of Experimental Psychology, 80(3p2), 1.

Brunk, J., Mattern, J., & Riehle, D. M. (2019). Effect of transparency and trust on acceptance of automatic online comment moderation systems. 2019 IEEE 21st Conference on Business Informatics (CBI), 1, 429–435.

Chen, C.-W. (2017). Five-star or thumbs-up? The influence of rating system types on users' perceptions of information quality, cognitive effort, enjoyment and

continuance intention. Internet Research, 27(3), 478–494. https://doi.org/10.1108/IntR-08-2016-0243

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017-Janua, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Cofta, P. (2009). Designing for Trust. In Handbook of Research on Socio-Technical Design and Social Networking Systems (Vol. 731, Issue 9985433, pp. 388–401). IGI Global. https://doi.org/10.4018/978-1-60566-264-0.ch026

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. The Lancet Digital Health, 1(8), e390. https://doi.org/10.1016/S2589-7500(19)30197-9

Dennis, A. R., & Carte, T. A. (1998). Using Geographical Information Systems for Decision Making: Extending Cognitive Fit Theory to Map-Based Presentations. Information Systems Research, 9(2), 194–203. https://doi.org/10.1287/isre.9.2.194

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Ml, 1–13. http://arxiv.org/abs/1702.08608

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 1–6. https://doi.org/10.1145/3290607.3312787

European Commission. (2021). Proposal for a Regulation of the European Parliament ans of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM, 206 final.

Fairclough, S. H., Karran, A. J., & Gilleade, K. (2015). Classification Accuracy from the Perspective of the User. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 3029–3038. https://doi.org/10.1145/2702123.2702454

Gillespie, B., Muehling, D. D., & Kareklas, I. (2018). Fitting product placements: Affective fit and cognitive fit as determinants of consumer evaluations of placed brands. Journal of Business Research, 82, 90–102. https://doi.org/10.1016/j.jbusres.2017.09.002

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89. https://doi.org/10.1109/DSAA.2018.00018

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Explainable Artificial intelligence (XAI), How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. NeuroIS 2021 Proceedings, 1–10.

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPRW.2009.5206848

Jia, R., Steelman, Z., & Reich, B. H. (2017). Using Mechanical Turk Data in IS Research: Risks, Rewards, and Recommendations. Communications of the Association for Information Systems, 41, 301–318. https://doi.org/10.17705/1CAIS.04114

Keller, K. L., & Staelin, R. (1987). Effects of Quality and Quantity of Information on Decision Effectiveness. Journal of Consumer Research, 14(2), 200. https://doi.org/10.1086/209106

Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395. https://doi.org/10.1145/2858036.2858402

Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, Humanness, and Trust: Rethinking Trust in Technology. Journal of the Association for Information Systems, 16(10), 880–918. https://doi.org/10.17705/1jais.00411

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Luhmann, N. (2001). Familiarity, confidence, trust: Problems and alternatives. Réseaux, 108(4), 15. https://doi.org/10.3917/res.108.0015

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology. ACM Transactions on Management Information Systems, 2(2), 1–25. https://doi.org/10.1145/1985347.1985353

Meske, C., & Bunde, E. (2020). Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12217 LNCS (pp. 54–69). Springer International Publishing. https://doi.org/10.1007/978-3-030-50334-5_4

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39–41. https://doi.org/10.1145/219717.219748

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. 1(1). http://arxiv.org/abs/1811.11839

Müller, R., Kessler, F., Humphrey, D. W., & Rahm, J. (2021). Data in context: How digital transformation can support human reasoning in cyber-physical production systems. 1–40. https://arxiv.org/pdf/2103.17095.pdf

Nourani, M., Honeycutt, D. R., Block, J. E., Roy, C., Rahman, T., Ragan, E. D., & Gogate, V. (2020). Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8. https://doi.org/10.1145/3334480.3382967

Nuamah, J. K., Seong, Y., Jiang, S., Park, E., & Mountjoy, D. (2020). Evaluating effectiveness of information visualizations using cognitive fit theory: A neuroergonomics approach. Applied Ergonomics, 88(June 2019), 103173. https://doi.org/10.1016/j.apergo.2020.103173

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10, April 2017, 141. https://doi.org/10.1145/1743666.1743701

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and Measuring Model Interpretability. http://arxiv.org/abs/1802.07810

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision, 128(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shaft, & Vessey. (2006). The Role of Cognitive Fit in the Relationship between Software Comprehension and Modification. MIS Quarterly, 30(1), 29. https://doi.org/10.2307/25148716

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. 34th International Conference on Machine Learning, ICML 2017, 7, 5109–5118. http://arxiv.org/abs/1703.01365

Sundararajan, M., Xu, S., Taly, A., Sayres, R., & Najmi, A. (2019). Exploring principled visualizations for deep network attributions. CEUR Workshop Proceedings, 2327.

van der Land, S., Schouten, A. P., Feldberg, F., van den Hooff, B., & Huysman, M. (2013). Lost in space? Cognitive fit and cognitive load in 3D virtual environments. Computers in Human Behavior, 29(3), 1054–1064. https://doi.org/10.1016/j.chb.2012.09.006

Vance, A., Elie-Dit-Cosaque, C., & Straub, D. W. (2008). Examining Trust in Information Technology Artifacts: The Effects of System Quality and Culture. Journal of Management Information Systems, 24(4), 73–100. https://doi.org/10.2753/MIS0742-1222240403

Vessey, I. (1991). Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature. Decision Sciences, 22(2), 219–240. https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

Vessey, I., & Galletta, D. (1991). Cognitive Fit: An Empirical Study of Information Acquisition. Information Systems Research, 2(1), 63–84. https://doi.org/10.1287/isre.2.1.63

Viera, A. J., Garrett, J. M., & others. (2005). Understanding interobserver agreement: the kappa statistic. Family Medicine, 37(5), 360–363.

Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., & Zschech, P. (2021). White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems. Nternational Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global, 0–9.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). “Do you trust me?”: Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 7–9. https://doi.org/10.1145/3308532.3329441

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3290605.3300509

Chapitre 4
Conclusion

L'objectif principal de ce mémoire était d'évaluer l'effet des visualisations de l'explication (VE) d'une décision prise par un système d'intelligence artificielle (IA) sur la certitude de l'utilisateur envers ce même système. Plus précisément, nous voulions vérifier l'effet de l'adjacence et de la clarté morphologique des VE sur la certitude. Dans un second temps, nous voulions évaluer l'effet de ces mêmes facteurs sur la charge cognitive des utilisateurs lors de l'analyse des VE, en plus de vérifier si cette charge cognitive a un effet de modération sur la certitude résultante des utilisateurs envers le système. Ce mémoire a ainsi permis de prouver que les VE ont bel et bien un effet sur la certitude de l'utilisateur. Effectivement, autant l'adjacence que la clarté morphologique de la VE ont un effet sur la certitude de l'utilisateur envers le système. Ce mémoire a aussi permis d'identifier diverses caractéristiques d'une visualisation qui diminue ou au contraire augmente la charge cognitive requise par l'utilisateur. De plus, un effet de modération a été observé entre la charge cognitive et la certitude des utilisateurs dans le système. Ces résultats permettent d'avoir une meilleure compréhension des bonnes pratiques pour la conception de VE compréhensibles et utilisables par l'utilisateur, entrainant ultimement une meilleure collaboration entre l'utilisateur et le système d'IA.

Deux expériences ont été menées pour permettre de répondre aux divers objectifs de ce mémoire. La première expérience en laboratoire, menée en août 2020 auprès de 19 participants, évaluait la charge cognitive requise par les utilisateurs lors de l'analyse des six types de VE à l'étude. L'expérience était décomposée en 60 itérations, où chacune présentait successivement une image suivie de sa classification par le système et de la VE. Le participant devait évaluer à chaque itération son niveau de certitude quant à la capacité du système de classifier, dans le futur, une image similaire à celle venant d'être présentée. Des mesures de pupillométrie évaluant la dilation de la pupille des participants ont été utilisées pour estimer la charge cognitive demandée à l'utilisateur. Les réponses à la question évaluant la confiance perçue de l'utilisateur ont permis ensuite d'évaluer la relation entre la charge cognitive et la certitude de l'utilisateur. Les participants recevaient

une compensation sous la forme d'une carte cadeau Amazon d'une valeur de 50 $ pour avoir participé à l'étude. L'article présenté au chapitre 2 présente les résultats obtenus à la suite de cette expérience.

La deuxième expérience menée auprès de 350 participants durant l'été 2020 consistait en une enquête en ligne distribuée sur la plateforme MTurk d'Amazon. Chacun des participants devait répondre à un questionnaire Qualtrics en ligne demandant en moyenne 15 minutes à répondre. La tâche consistait en 50 itérations, dont chacun d'elle présentait sur la même page une image, la VE et la classification de l'image par le système d'IA. Une image semblable à la première était aussi présentée et le participant devait évaluer son niveau de certitude par rapport à la capacité du système de classifier correctement la deuxième image présentée. À la suite des 50 itérations, les participants devaient répondre à quelques questions sociodémographiques. Puisque la qualité des réponses des participants sur les plateformes de distribution d'enquête en ligne telle que MTurk est grandement variable, plusieurs contrôles d'attention ont été placés à travers le questionnaire. Ces derniers permettaient de vérifier si le participant répondait correctement aux questions. Un grand nombre de réponses a ainsi été filtré nous laissant avec 206 participations utilisables pour l'analyse. Les participants ont ensuite reçu une compensation de 2 $ pour leur participation. Cette expérimentation a permis de rédiger l'article présenté au chapitre 3 de ce mémoire.

4.1 Rappel des questions de recherche

L'objectif de ce mémoire était d'étudier l'effet de différentes VE d'un système de reconnaissance d'image sur la certitude de l'utilisateur envers le système ainsi que la charge cognitive de l'utilisateur. Les questions de recherche de ce mémoire par articles étaient donc les suivantes :

Article 1 :

- *Dans quelle mesure l'adjacence et la clarté morphologique de la visualisation de l'explication influencent-elles la charge cognitive requise par l'utilisateur utilisant ces explications ?*

- *Y a-t-il une corrélation entre la charge cognitive requise par l'utilisateur utilisant ces explications et la certitude de l'utilisateur envers le système d'IA et ses décisions ?*

Article 2 :

- *Dans quelle mesure la visualisation d'une explication de la décision d'un système d'IA affectera-t-elle la certitude de l'utilisateur envers ce système et ses décisions?*

- *Dans quelle mesure l'adjacence et la clarté morphologique de la visualisation de l'explication favorisent-elles ou dégradent-elles la certitude de l'utilisateur envers le système d'IA et ses décisions ?*

4.2 Principaux résultats

Les résultats suivants sont présentés selon l'ordre des questions énumérées à la section précédente.

Les résultats présentés dans la première étude montrent qu'en effet l'adjacence et la clarté morphologique ont un effet sur la charge cognitive de l'utilisateur lors de l'utilisation d'une VE. Une visualisation adjacente demande ainsi moins de charge cognitive qu'une visualisation non adjacente. Pour les VE non adjacente, l'utilisateur doit mentalement associer l'explication avec l'image originale, lui demandant ainsi un effort cognitif supérieur. Les résultats laissent présager à première vue que la visualisation ayant une clarté morphologique moyenne demande une charge cognitive inférieure aux deux autres niveaux de clarté morphologique. Cependant, après une analyse plus approfondie, on remarque que la visualisation non adjacente avec clarté morphologique moyenne résulte en une charge cognitive négative. Cependant, ce type de visualisation est très abstrait, rendant l'association de l'explication avec l'image originale très difficile, voire presque impossible. Ainsi, il est très difficile pour un utilisateur de bien analyser cette visualisation. Ceci provoque alors une charge cognitive très basse due à un effet de désengagement de l'utilisateur.

Une corrélation négative a été observée entre la charge cognitive et la certitude de l'utilisateur envers le système, signifiant qu'une visualisation demandant une charge cognitive importante, comme les visualisations non adjacentes, risque de générer une certitude plus faible envers le système. Cependant, la visualisation non adjacente et ayant une clarté morphologique basse semble générer un niveau de certitude supérieur aux deux autres visualisations non adjacentes, bien que la charge cognitive associée à cette visualisation soit la plus élevée.

Les résultats de la deuxième étude montrent que l'adjacence, la clarté morphologique ainsi que l'interaction entre ces deux facteurs ont un effet sur la certitude de l'utilisateur envers le système. En effet, une VE adjacente a un effet positif supérieur sur la certitude à celui des VE non adjacentes. Cet effet peut être expliqué par le fait que les VE adjacentes demandent un effort cognitif inférieur à l'utilisateur puisque l'association entre l'explication et l'image originale a déjà été faite. L'utilisateur peut donc identifier plus facilement quelles parties de l'image ont pu influencer le système dans sa prise de décision, et ainsi comparer son propre raisonnement avec celui du système. Ensuite, les VE ayant une clarté morphologique basse ont un effet positif supérieur sur le niveau de certitude des utilisateurs par rapport aux deux autres niveaux de clarté morphologique. En observant les interactions entre l'adjacence et la clarté morphologique, on observe qu'il n'y a cependant pas de différence significative au niveau de la certitude entre les trois VE adjacentes. La différence se situe plutôt au niveau des visualisations non adjacentes, où celle ayant une basse clarté morphologique entraine une certitude nettement supérieure aux visualisations ayant une clarté morphologique moyenne et élevée. Ces résultats semblent indiquer que les utilisateurs exigent plus d'information et de précision dans une situation où une VE non adjacente leur est présentée. La plus grande précision fournie par les visualisations avec clarté morphologique basse permet ainsi aux utilisateurs d'associer mentalement l'explication avec l'image originale plus facilement. Les deux autres niveaux, en situation de non adjacence, sont trop abstraits pour permettre à l'utilisateur d'identifier les données de l'explication avec l'objet de l'image originale.

4.3 Contributions théoriques et pratiques de l'étude

Dans un premier temps, la théorie du Cognitive Fit (Vessey, 1991) a été proposée dans le but d'étudier l'effet de représentation tabulaire et graphique sur la performance d'un individu lors de la réalisation d'une tâche. Cette théorie a été adaptée au cours des dernières années dans divers domaines de recherche tels que les systèmes d'information géographique (SIG) (Dennis & Carte, 1998), le commerce en ligne (Brunelle, 2009; Chen, 2017), les bases de données (Bizarro, 2015), et les interfaces mobiles (Adipat et al., 2011; Chua & Chang, 2016; Urbaczewski & Koivisto, 2008) en utilisant des diverses représentations de l'information. Nous contribuons à cette théorie en l'intégrant dans la recherche en XAI. Nous proposons ainsi des lignes directrices par rapport à l'apparence d'une VE qui favorise le développement d'une représentation mentale optimisée du problème par l'utilisateur, améliorant ainsi les performances lors de la réalisation de la tâche. Cette amélioration de la performance se manifeste en diminuant la charge cognitive de l'utilisateur et en augmentant la certitude de ce dernier envers le système.

Les diverses techniques d'interprétabilité et de visualisation de l'explication développées à travers la recherche en XAI ont un impact positif sur la certitude de l'utilisateur envers le système (Cofta, 2009; Eiband et al., 2019; Glikson & Woolley, 2020; Lee & See, 2004; Meske & Bunde, 2020). Ce mémoire ajoute à cette découverte en montrant que l'effet sur la certitude de l'utilisateur est modulé par les choix de présentation des VE. En effet, certains choix de présentation peuvent favoriser le développement de la certitude de l'utilisateur envers le système. Au contraire d'autres choix de présentation peuvent avoir un effet nul, voire négatif, sur la certitude de l'utilisateur en comparaison à une situation où il n'y aurait pas de transparence (c.-à-d. aucune visualisation de l'explication présentée). Les résultats de ce mémoire proposent donc qu'une attention particulière doive être portée aux choix de présentation de l'explication, et ce en tenant compte du contexte d'utilisation et de la tâche à réaliser. Tout cela dans l'optique d'avoir une VE favorisant le développement d'un niveau de certitude adéquat dans le système.

De plus, les résultats préliminaires de ce mémoire semblent s'accorder avec la littérature existante dans la recherche en système d'information quant à la relation entre la charge cognitive de l'utilisateur et la confiance ou certitude de ce dernier (Gupta et al., 2019;

Khawaji et al., 2014; Samson & Kostyszyn, 2015). Bien que ces dernières études n'évaluent pas toutes la confiance ou la certitude de l'utilisateur envers le système, tous semblent montrer qu'une charge cognitive inférieure semble entrainer une confiance ou certitude supérieure de l'utilisateur. Ce mémoire ajoute à la littérature existante en évaluant l'effet de la charge cognitive de l'utilisateur sur la certitude dans de futures prédictions ou classifications du système d'IA.

La recherche en XAI est encore à ses débuts et ses découvertes peuvent difficilement être intégrées pour l'instant dans un contexte professionnel, puisque la plupart des systèmes et modèles interprétables sont peu compréhensibles et utilisables pour un utilisateur. Ce mémoire permet donc d'ouvrir la voie vers l'utilisation de techniques d'interprétabilité dans le milieu professionnel en fournissant des lignes directrices permettant la conception de visualisation orientée utilisateur. À plus grande échelle, cette étude peut aussi servir d'exemple dans le développement et l'évaluation de techniques d'interprétabilité utilisées dans des domaines tels que la radiologie et l'ophtalmologie, où des systèmes d'IA sont utilisés pour aider à détecter des anomalies sur des imageries médicales. La conception et l'évaluation de VE orientée utilisateur, tout en considérant le contexte et la tâche, seraient ainsi grandement bénéfiques pour le professionnel en l'aidant dans la réalisation de ses tâches complexes et ayant un impact critique sur la santé d'êtres humains.

4.4 Limites et pistes de recherches futures

Il est important de considérer quelques limites aux expérimentations conduites dans le cadre de ce mémoire.

Pour les études présentées dans ce mémoire, les données démographiques reliées aux échantillons de participant n'ont pas été incluses dans l'analyse des résultats. Ces données pourraient avoir un effet modérateur sur les résultats. Le niveau d'éducation, l'emploi, l'âge ainsi que les connaissances préalables des participants reliées à l'IA pourraient tous relever des différences au niveau de la certitude et de la charge cognitive des participants. Un troisième article, destiné à être publié dans une revue scientifique, inclura une analyse sociodémographique de l'échantillon en plus de mesures supplémentaires permettant

d’estimer la charge cognitive des utilisateurs. Cet article permettra ainsi de confirmer ou d’infirmer les résultats obtenus dans ce mémoire.

De plus, certains utilisateurs sont par défaut plus enclin à faire confiance à une nouvelle technologie ou à être certain des décisions prises par un système que d’autres, que ce soit par leurs croyances ou leur éducation par exemple. Il serait donc important d’évaluer dans une étude à venir la propension du participant à faire confiance à une nouvelle technologie ou à être certain des décisions prises par un système et ainsi de contrôler l’influence de ce facteur sur les résultats.

De plus, les classifications d’image ainsi que la génération d’explication se sont faites en utilisant seulement un algorithme de reconnaissance d’image. Même si les techniques de visualisation utilisées dans ce mémoire peuvent être employées avec différents systèmes, certaines variations dans la classification de l’image peuvent survenir donc venir modifier l’explication résultante. De futures études pourraient donc vérifier si nos résultats se répètent en utilisant d’autres systèmes de reconnaissance d’image et du même coup observer si le choix de l’algorithme a aussi un effet sur la certitude de l’utilisateur envers le système.

Le design expérimental des deux études présentées dans ce mémoire utilise une seule question, répétée entre 50 et 60 fois dépendant de l’étude, pour mesurer la certitude de l’utilisateur envers le système d’IA. En effet, nous avons choisi de présenter une seule question par itération pour ne pas trop alourdir le questionnaire. Dans une situation où on aurait demandé au participant de répondre à plus d’une question par itération, les participants auraient été d’avantage porté à abandonner ou à répondre de manière aléatoire pour accélérer l’achèvement du questionnaire. Cependant, des études en systèmes de recommandation centrés sur l’utilisateur, un domaine de recherche de l’IA, décrivent l’importance de mesurer la confiance des utilisateurs à l’aide de plusieurs questions (Knijnenburg & Willemsen, 2015; Pu et al., 2011). Une future étude pourrait ainsi se pencher sur cette recommandation en utilisant par exemple une échelle standardisée évaluant la certitude des utilisateurs envers le système d’IA à chaque itération, en diminuant cependant le nombre d’itérations totales imposées à chaque participant. Cette

nouvelle étude pourrait ensuite comparer ses résultats avec ceux présentés dans ce mémoire et observer si différentes conclusions intéressantes en ressortent.

Ce mémoire s'est concentré sur l'évaluation d'explication présentée sous forme de visualisation utilisant diverses couleurs et différents niveaux de précision pour tenter d'éclairer l'utilisateur sur le pourquoi de la prédiction du système. Cependant, présenter une explication d'un système de reconnaissance d'images sous forme d'une visualisation n'est pas l'unique manière valide de présenter une explication. En effet, Hendricks et al. (2016) propose une technique d'explication textuelle. Cette technique se concentre sur les propriétés discriminatives de l'objet présenté dans l'image en expliquant à l'aide de phrase complète pourquoi la prédiction faite par le système est appropriée. Un exemple d'explication d'une classification d'une image présentant un oiseau serait : « Ceci est un vacher bronzé, car c'est un oiseau noir avec un œil rouge et un bec noir pointu ». Ce genre d'explication étant facilement compréhensible et utilisable par tout utilisateur, il serait donc intéressant dans une étude ultérieure d'évaluer l'effet sur la certitude et la charge cognitive d'une telle présentation de l'explication et y comparer les résultats présentés dans ce mémoire. Il serait même pertinent de comparer divers scénarios entre eux utilisant à la fois des explications textuelles et visuelles. La finalité serait de mieux définir les caractéristiques et exigences d'une explication d'un système d'IA qui maximise la certitude et minimise la charge cognitive de l'utilisateur, l'aidant ainsi dans sa prise de décision et favorisant une meilleure collaboration entre l'humain et les systèmes d'IA.

Bibliographie

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018-April, 1–18. https://doi.org/10.1145/3173574.3174156

Adipat, Zhang, & Zhou. (2011). The Effects of Tree-View Based Presentation Adaptation on Mobile Web Browsing. MIS Quarterly, 35(1), 99. https://doi.org/10.2307/23043491

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. In ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Attard-Johnson, J., Ó Ciardha, C., & Bindemann, M. (2019). Comparing methods for the analysis of pupillary response. Behavior Research Methods, 51(1), 83–95. https://doi.org/10.3758/s13428-018-1108-6

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58(December 2019), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. Journal of Experimental Psychology, 80(3p2), 1.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, 91(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bigras, É., Léger, P.-M., & Sénécal, S. (2019). Recommendation Agent Adoption: How Recommendation Presentation Influences Employees' Perceptions, Behaviors, and Decision Quality. Applied Sciences, 9(20). https://doi.org/10.3390/app9204244

Bizarro, P. A. (2015). Effect of different database structure representations, query languages, and task characteristics on information retrieval. Journal of Management Information and Decision Science, 18(1), 27–52.

Brunelle, E. (2009). The moderating role of cognitive fit in consumer channel preference. Journal of Electronic Commerce Research, 10(3).

Brunk, J., Mattern, J., & Riehle, D. M. (2019). Effect of transparency and trust on acceptance of automatic online comment moderation systems. 2019 IEEE 21st Conference on Business Informatics (CBI), 1, 429–435.

Chen, C.-W. (2017). Five-star or thumbs-up? The influence of rating system types on users' perceptions of information quality, cognitive effort, enjoyment and continuance intention. Internet Research, 27(3), 478–494. https://doi.org/10.1108/IntR-08-2016-0243

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017-Janua, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Chua, W. Y., & Chang, K. T. T. (2016). An investigation of usability of push notifications on mobile devices for novice and expert users. Proceedings of the Annual Hawaii International Conference on System Sciences, 2016-March. https://doi.org/10.1109/HICSS.2016.703

Cofta, P. (2009). Designing for Trust. In Handbook of Research on Socio-Technical Design and Social Networking Systems (Vol. 731, Issue 9985433, pp. 388–401). IGI Global. https://doi.org/10.4018/978-1-60566-264-0.ch026

Dastin, J. (2018). Amazon reportedly scraps internal AI recruiting tool that was biased against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. The Lancet Digital Health, 1(8), e390. https://doi.org/10.1016/S2589-7500(19)30197-9

Dennis, A. R., & Carte, T. A. (1998). Using Geographical Information Systems for Decision Making: Extending Cognitive Fit Theory to Map-Based Presentations. Information Systems Research, 9(2), 194–203. https://doi.org/10.1287/isre.9.2.194

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Ml, 1–13. http://arxiv.org/abs/1702.08608

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 1–6. https://doi.org/10.1145/3290607.3312787

European Commission. (2021). Proposal for a Regulation of the European Parliament ans of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM, 206 final.

Fairclough, S. H., Karran, A. J., & Gilleade, K. (2015). Classification Accuracy from the Perspective of the User. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 3029–3038. https://doi.org/10.1145/2702123.2702454

Gillespie, B., Muehling, D. D., & Kareklas, I. (2018). Fitting product placements: Affective fit and cognitive fit as determinants of consumer evaluations of placed brands. Journal of Business Research, 82, 90–102. https://doi.org/10.1016/j.jbusres.2017.09.002

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89. https://doi.org/10.1109/DSAA.2018.00018

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goodhue, D. L., & Thompson, R. L. (1995). Task-Technology Fit and Individual Performance. MIS Quarterly, 19(2), 213. https://doi.org/10.2307/249689

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating Visual Explanations. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 9908 LNCS (pp. 3–19). https://doi.org/10.1007/978-3-319-46493-0_1

Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Explainable Artificial intelligence (XAI), How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. NeuroIS 2021 Proceedings, 1–10.

Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Towards user-centric AI explanations: increasing confidence in AI through visualizations of an AI system explanation. Soumis à La Conférence ICIS 2021.

Isabella, S. L., Urbain, C., Cheyne, J. A., & Cheyne, D. (2019). Pupillary responses and reaction times index different cognitive processes in a combined Go/Switch incidental learning task. Neuropsychologia, 127, 48–56. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2019.02.007

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-Mediated Communication. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300469

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPRW.2009.5206848

Jia, R., Steelman, Z., & Reich, B. H. (2017). Using Mechanical Turk Data in IS Research: Risks, Rewards, and Recommendations. Communications of the Association for Information Systems, 41, 301–318. https://doi.org/10.17705/1CAIS.04114

Keller, K. L., & Staelin, R. (1987). Effects of Quality and Quantity of Information on Decision Effectiveness. Journal of Consumer Research, 14(2), 200. https://doi.org/10.1086/209106

Khawaji, A., Chen, F., Zhou, J., & Marcus, N. (2014). Trust and Cognitive Load in the Text-Chat Environment: The Role of Mouse Movement. Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design, 324–327. https://doi.org/10.1145/2686612.2686661

Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395. https://doi.org/10.1145/2858036.2858402

Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating Recommender Systems with User Experiments. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender Systems Handbook (pp. 309–352). Springer US. https://doi.org/10.1007/978-1-4899-7637-6_9

Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, Humanness, and Trust: Rethinking Trust in Technology. Journal of the Association for Information Systems, 16(10), 880–918. https://doi.org/10.17705/1jais.00411

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Luhmann, N. (2001). Familiarity, confidence, trust: Problems and alternatives. Réseaux, 108(4), 15. https://doi.org/10.3917/res.108.0015

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology. ACM Transactions on Management Information Systems, 2(2), 1–25. https://doi.org/10.1145/1985347.1985353

Meske, C., & Bunde, E. (2020). Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12217 LNCS (pp. 54–69). Springer International Publishing. https://doi.org/10.1007/978-3-030-50334-5_4

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39–41. https://doi.org/10.1145/219717.219748

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. 1(1). http://arxiv.org/abs/1811.11839

Müller, R., Kessler, F., Humphrey, D. W., & Rahm, J. (2021). Data in context: How digital transformation can support human reasoning in cyber-physical production systems. 1–40. https://arxiv.org/pdf/2103.17095.pdf

Niezgoda, M., Tarnowski, A., Kruszewski, M., & Kamiński, T. (2015). Towards testing auditory–vocal interfaces and detecting distraction while driving: A comparison of eye-movement measures in the assessment of cognitive workload. Transportation Research Part F: Traffic Psychology and Behaviour, 32, 23–34. https://doi.org/https://doi.org/10.1016/j.trf.2015.04.012

Nourani, M., Honeycutt, D. R., Block, J. E., Roy, C., Rahman, T., Ragan, E. D., & Gogate, V. (2020). Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8. https://doi.org/10.1145/3334480.3382967

Nuamah, J. K., Seong, Y., Jiang, S., Park, E., & Mountjoy, D. (2020). Evaluating effectiveness of information visualizations using cognitive fit theory: A neuroergonomics approach. Applied Ergonomics, 88(June 2019), 103173. https://doi.org/10.1016/j.apergo.2020.103173

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA ’10, April 2017, 141. https://doi.org/10.1145/1743666.1743701

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and Measuring Model Interpretability. http://arxiv.org/abs/1802.07810

Pu, P., Chen, L., & Hu, R. (2011). A User-Centric Evaluation Framework for Recommender Systems. Proceedings of the Fifth ACM Conference on Recommender Systems, 157–164. https://doi.org/10.1145/2043932.2043962

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samson, K., & Kostyszyn, P. (2015). Effects of Cognitive Load on Trusting Behavior – An Experiment Using the Trust Game. PLOS ONE, 10(5), 1–10. https://doi.org/10.1371/journal.pone.0127680

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision, 128(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shaft, & Vessey. (2006). The Role of Cognitive Fit in the Relationship between Software Comprehension and Modification. MIS Quarterly, 30(1), 29. https://doi.org/10.2307/25148716

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. 34th International Conference on Machine Learning, ICML 2017, 7, 5109–5118. http://arxiv.org/abs/1703.01365

Sundararajan, M., Xu, S., Taly, A., Sayres, R., & Najmi, A. (2019). Exploring principled visualizations for deep network attributions. CEUR Workshop Proceedings, 2327.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. Patterns, 1(4), 100049. https://doi.org/https://doi.org/10.1016/j.patter.2020.100049

Urbaczewski, A., & Koivisto, M. (2008). The Importance of Cognitive Fit in Mobile Information Systems. Communications of the Association for Information Systems, 22. https://doi.org/10.17705/1cais.02210

van der Land, S., Schouten, A. P., Feldberg, F., van den Hooff, B., & Huysman, M. (2013). Lost in space? Cognitive fit and cognitive load in 3D virtual environments. Computers in Human Behavior, 29(3), 1054–1064. https://doi.org/10.1016/j.chb.2012.09.006

Vance, A., Elie-Dit-Cosaque, C., & Straub, D. W. (2008). Examining Trust in Information Technology Artifacts: The Effects of System Quality and Culture. Journal of Management Information Systems, 24(4), 73–100. https://doi.org/10.2753/MIS0742-1222240403

Vessey, I. (1991). Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature. Decision Sciences, 22(2), 219–240. https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

Vessey, I., & Galletta, D. (1991). Cognitive Fit: An Empirical Study of Information Acquisition. Information Systems Research, 2(1), 63–84. https://doi.org/10.1287/isre.2.1.63

Vieira, L. N. (2016). How do measures of cognitive effort relate to each other? A multivariate analysis of post-editing process data. Machine Translation, 30(1–2), 41–62. https://doi.org/10.1007/s10590-016-9188-5

Viera, A. J., Garrett, J. M., & others. (2005). Understanding interobserver agreement: the kappa statistic. Family Medicine, 37(5), 360–363.

Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., & Zschech, P. (2021). White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems. Nternational Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global, 0–9.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). "Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 7–9. https://doi.org/10.1145/3308532.3329441

Wickens, C. D. (2008). Multiple resources and mental workload. Human Factors, 50(3), 449–455. https://doi.org/10.1518/001872008X288394

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3290605.3300509