h e g

Haute école de gestion
Genève

Artificiel ou réel ? Perception des images générées par IA

Travail de master réalisé par :

Alexandra ANDRADE

Sous la direction de :

Christian MUMENTHALER, Professeur HES

Genève, le 15 août 2024

Information science

Haute École de Gestion de Genève (HEG-GE)

Hes·SO GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

Ce travail de Master est réalisé dans le cadre du Master en Sciences de l'information de la Haute école de gestion de Genève.

L'étudiante atteste que le travail rendu est le fruit de sa réflexion personnelle, a été rédigé de manière autonome sans avoir utilisé des sources autres que celles citées dans la bibliographie et a été vérifié par un logiciel de détection de plagiat.

L'étudiante accepte, le cas échéant, la clause de confidentialité.

L'utilisation des conclusions et recommandations formulées dans ce travail, sans préjuger de leur valeur, n'engage ni la responsabilité de l'étudiante, ni celle du directeur.

Fait à Genève, le 15 août 2024

Alexandra ANDRADE

Remerciements

Je tiens à remercier Christian Mumenthaler pour sa supervision et ses conseils concernant la réalisation de ce travail, mais surtout pour son aide précieuse lorsque j'étais coincée sur R et lors de l'analyse des résultats du questionnaire.

Je souhaite remercier aussi mes collègues de travail qui depuis plus de deux ans m'ont énormément appris et m'ont toujours encouragée dans ce master. C'est grâce à leur bienveillance et compréhension lors de mes deadlines que j'ai pu jongler entre mes études et ma vie professionnelle sans me sentir débordée.

Je remercie ma famille et mes amis qui eux aussi m'encouragent depuis des années et qui sont toujours là pour moi, dans les bons moments comme dans les moins bons. C'est grâce à leur support que j'ai pu terminer mes longues études et je sais que c'est avec leur support que j'accomplirais toutes les prochaines étapes de ma vie également.

Je tiens également à remercier chaleureusement mes camarades du Master IS pour leurs conseils, leurs encouragements et les nombreux verres et repas partagés tous ensemble. Sans elles et eux, ces trois années de Master n'auraient pas été aussi fun.

Résumé

Avec l'avènement de l'intelligence artificielle, et particulièrement l'intelligence artificielle générative, de plus en plus de questionnements se posent quant à son utilisation, notamment en ce qui concerne la génération d'images. L'objectif de ce travail est d'analyser comment nous percevons les images générées par intelligence artificielle et si nous sommes capables de les distinguer d'images réelles.

Ce travail commence tout d'abord par une revue de la littérature, dans laquelle nous abordons le sujet des deepfakes en premier lieu. En effet, les deepfakes peuvent poser différentes craintes et menaces. Il devient de plus en plus difficile d'avoir confiance en ce que nous voyons, créant une peur d'une pandémie de désinformation, mais en nous assurant de la fiabilité de nos sources, nous pouvons contourner ce problème. Différentes techniques existent également pour détecter les deepfakes, que ce soit à travers de programmes spécialisés ou par le manque de clignement d'yeux de la personne présente dans la vidéo.

En deuxième lieu, nous plongeons dans la perception d'images générées par intelligence artificielle et nous voyons, à travers de multiples expériences différentes, que l'IA est capable de générer des images si hyperréalistes, que certains visages de personnes réelles sont considérés comme synthétiques en comparaison. Cela s'expliquerait par le fait que l'IA génère des visages plus communs et donc perçus comme plus humains. Certaines personnes avaient même plus confiance en leur jugement lorsqu'elles jugeaient des images IA comme réelles que lorsque les images étaient réellement réelles.

Nous présentons ensuite notre recherche, ses hypothèses et notre méthodologie. En effet, nous avons tout d'abord créer un corpus d'images IA selon 4 catégories (Animaux, Paysages, Tableaux et Visages) puis récolté des images réelles correspondantes. Nous avons ensuite créé un questionnaire que nous avons fait passer à des participant-e-s.

Nos résultats montrent que de manière globale, les participant-e-s étaient capables de distinguer des images IA d'images réelles. C'est également le cas par catégorie, sauf dans la catégorie des visages, pour laquelle les résultats n'étaient pas significatifs et rejoignent donc les résultats des études abordées dans la revue de la littérature. Nous mentionnons également les quelques limitations de notre étude et de quelle façon nous pourrions l'améliorer. Finalement, nous esquissons une ébauche d'idée pour un éventuel tableau de bord qui nous permettrait de partager nos résultats avec un plus grand public.

Mots-clés : Intelligence artificielle, deepfake, perception visuelle, hyperréalisme, science de l'information, désinformation

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

Ces dernières années, il semble que le terme « intelligence artificielle », ou IA, se soit exponentiellement propagé sur toutes les lèvres. De nombreux articles paraissent dans plusieurs journaux, médias et sur le web, expliquant ce qu'est l'IA, reportant sur les avancées de la technologie ou sur les nouvelles applications de celle-ci dans des domaines parfois surprenants. En effet, l'IA a déjà commencé à être utilisée dans la gestion de déchets (Orbisk 2023), la création de bières ou même la protection d'abeilles (Wolter, Savani 2024). Loin de se limiter aux industries et contextes majoritairement technologiques, l'IA possède une force d'adaptation et d'innovation qui ne semble pas prête de s'arrêter de sitôt.

L'intérêt du public général pour l'IA a particulièrement surgi lors du lancement du chatbot ChatGPT d'OpenAI en novembre 2022. ChatGPT n'est pas uniquement de l'intelligence artificielle, c'est de l'intelligence artificielle *générative*. L'IA générative, à l'inverse de l'IA non générative, est « capable de générer du contenu inédit » (Adobe Firefly 2023) d'après des prompts que nous lui fournissons. L'IA peut donc générer non seulement du texte, mais également des images et, plus récemment, des vidéos.

Les débuts de la génération d'images n'étaient pas fameux car souvent l'IA incorporait des incohérences dans ce qu'elle générait comme, par exemple, des mains ou des jambes supplémentaires. Cependant, l'IA s'est considérablement améliorée et de plus en plus les images générées sont réalistes et presque indiscernables d'images produites par des appareils photographiques. Cette capacité de réalisme soulève des questions sur la capacité humaine à distinguer entre une image créée par l'intelligence artificielle ou une image créée par un être humain.

À travers ce travail, nous souhaitons explorer cette capacité de distinction, notamment sur une courte durée telle qu'une première impression. En effet, les images IA étant de plus en plus courantes dans notre vie quotidienne, il est important de comprendre à quel point nous sommes capables de les reconnaitre. Ce besoin de discernement est d'autant plus important lorsque nous prenons en considération que nous sommes dans l'ère des deepfakes qui peuvent malheureusement être utilisés à des fins malhonnêtes.

La première partie de notre travail est constitué d'une revue de la littérature abordant notamment ce sujet des deepfakes et des éventuels problèmes que ceux-ci pourraient causer, mais également la perception humaine des images, et plus spécifiquement des visages, générées par intelligence artificielle.

Dans la deuxième partie, nous présenterons notre étude pour laquelle nous avons compilé un corpus d'images selon différentes catégories, incluant à la fois des images réelles que nous avons rassemblé et des images que nous avons générées par IA à travers différents sites de génération d'image. Nous avons ensuite présenté ces images de manière aléatoire, et sans préciser si les images étaient réelles ou générées, pendant uniquement une seconde à travers un questionnaire que nous avons fait passer à des participant-e-s. Il leur est ensuite demandé dans un premier temps de juger à quel point l'image leur a paru être réelle ou générée par IA. Puis dans un deuxième temps, d'évaluer leur propre niveau de confiance par rapport au jugement fait lors de la première question.

Nous espérons apporter une meilleure compréhension de la perception des images générées par intelligence artificielle et éventuellement fournir des conseils et recommandations pour une interaction plus sûre et critique des images que nous voyons sur les réseaux et sur nos écrans.

2. Revue de la littérature

2.1 L'ère des deepfakes et de la désinformation

L'intelligence artificielle générative semble particulièrement incroyable, et elle l'est, car elle peut nous permettre de gagner du temps dans certaines tâches ou nous aider à brainstormer des idées. Cependant, l'IA fait également surgir certains questionnements et certaines craintes quant à son utilisation qui peut parfois être détournée de manière malhonnête. C'est surtout le cas dans le contexte de la génération d'images et de vidéos deepfakes.

À travers les réseaux sociaux et grâce aux améliorations technologiques des appareils photos et téléphones portables, il n'a jamais été aussi simple de créer, éditer et propager des vidéos, selon Li, Chang et Lyu. C'était déjà le cas en 2018, lorsqu'ils ont écrit leur article, mais en 2024 nous pouvons bien imaginer que le phénomène a pris de l'ampleur, notamment grâce au réseau social Tik Tok. En effet, grâce à celui-ci, les utilisateurs peuvent facilement filmer des vidéos, les éditer et même y appliquer des filtres divers et variés. Tik Tok est l'une des applications les plus utilisées dans le monde, avec 1,5 milliards d'utilisateurs actifs par mois en 2023 (Iqbal 2024), et, avec son format de vidéos courtes, a même inspiré d'autres grands réseaux sociaux comme Instagram et Youtube dont les réels et shorts se rapprochent beaucoup de ce qui se voit sur Tik Tok. La popularisation des réseaux sociaux vidéos qui « provide organizations and individuals with the tools and technology to create and post content make it very easy to distribute deepfakes » (Kietzmann et al. 2020, p.137).

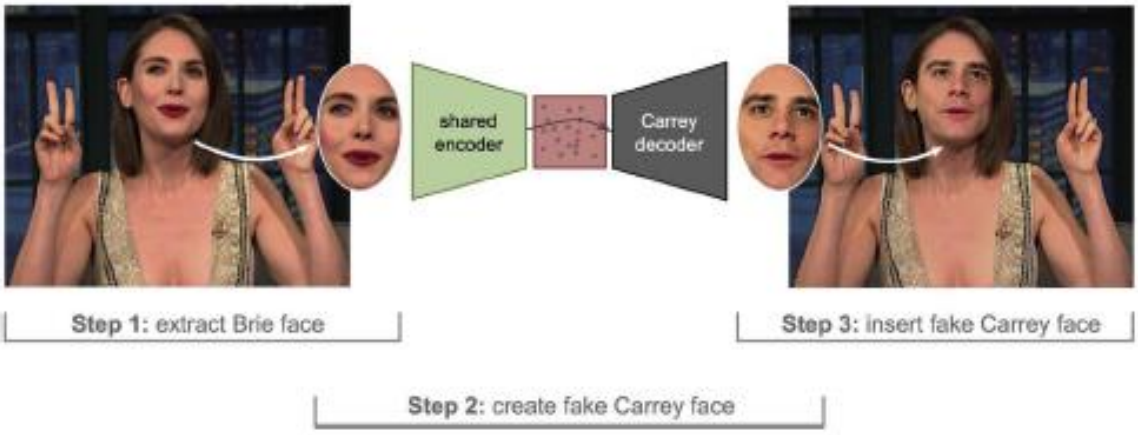
Cependant, c'est bien en 2018 que commence l'ère des deepfakes, avec la sortie et la diffusion publique du software DeepFake. Ce programme permet la falsification de vidéos en remplaçant le visage de la personne filmée, par celui de quelqu'un d'autre ou en remplaçant par son propre visage, mais en le modifiant pour que la personne paraisse dire autre chose que ce qu'elle dit réellement. C'est par exemple le cas avec l'une des premières vidéo deepfake dont le public a eu connaissance, celle de Barack Obama dans laquelle il « warns about the dangers of deepfakes, something that Obama never actually did » (Karnouskos 2020, p.138).

C'est grâce aux réseaux antagonistes génératifs (GANs) que les programmes de création de deepfakes peuvent échanger des visages de cette façon et générer des vidéos qui n'ont jamais existé. Les GANs étant très « data-hungry » (Kietzmann et al. 2020, p.139) un grand nombre d'images, se comptant dans les dizaines de milliers, doivent être fournies au modèle afin que celui-ci puisse s'entrainer (Li, Chang, Lyu 2018, p. 1). C'est pour cette raison que les célébrités sont les plus touchées par la création de deepfakes à leur image, car « an extensive library of images and videos already exists to train the networks » (Kietzmann et al. 2020, p. 139).

2.1.1 Les deepfakes, comment ça marche ?

Il nous semble important d'expliquer ici brièvement comment fonctionnent les deepfakes. Pour cela, nous allons nous baser sur l'article de Kietzmann et al. qui y ont consacré une partie, en utilisant comme exemple une vidéo deepfake dans laquelle le visage d'Alison Brie est remplacé par celui de Jim Carrey. Le procédé est résumé ci-dessous par la Figure 1.

Figure 1: Processus de création d'un deepfake



(Kietzmann et al. 2020, p. 138)

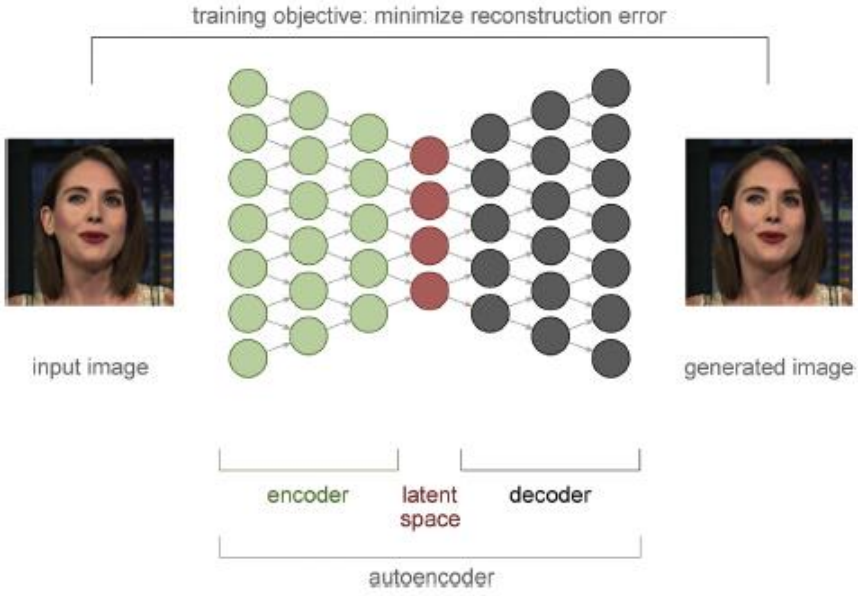
Les deepfakes sont créés grâce au deep learning « that can be used to train [deep neural networks (DNNs)] reminiscent of neurons in the brain » (Kietzmann et al. 2020, p. 138). Ces « neurones » sont appelés des unités et accomplissent chacune une tâche simple, mais ensemble elles sont capables d'accomplir des opérations plus complexes tel que reconnaitre des visages. « The computations of DNNs are dictated by the strength of the connection of their respective units » (Kietzmann et al. 2020, p. 139), qui avec de l'entrainement vont faire en sorte que les DNNs reconnaissent les visages.

Cependant, afin de pouvoir générer des visages spécifiques, les DNN doivent pouvoir déterminer les caractéristiques spécifiques de chaque visage afin de pouvoir les générer.

> *« The process of first recognizing a comparably small number of facial characteristics in the input and then generating real-looking faces as output is accomplished in three subparts: an* encoder*, a* latent space*, and a* decoder *» (Kietzmann et al. 2020, p. 139).*

L'*encoder* va compresser l'image en caractéristiques essentielles, puis celles-ci sont ensuite représentées dans le *latent space*. Dans cet espace, le DNN « can learn more general facial characteristics rather than memorizing all input examples of specific people » (Kietzmann et al. 2020, p. 139). Finalement, le *decoder* va décompresser l'image afin de la reconstruire. L'architecture du DNN est représentée à la figure 2 ci-dessous :

Figure 2: Représentation de l'autoencoder



(Kietzmann et al. 2020, p. 140)

L'astuce pour créer des deepfakes est d'utiliser le même *encoder* pour différents *autoencoders* qui vont « learn to use general features that the faces […] have in common » (Kietzmann et al. 2020, p. 141). Cependant, comme nous l'avons mentionné plus haut, pour que cette astuce fonctionne bien, une bibliothèque extensive d'images est nécessaire afin que l'*encoder* puisse avoir suffisamment de caractéristiques communes à identifier.

La technologie pour créer des deepfakes est complexe, mais le matériel nécessaire pour le faire est assez simple, nécessitant uniquement un PC et une carte graphique (Karnouskos 2020, pp. 138-139). « Because of the low learning curve [and] public access to the technology, deepfakes can be created easily even by home users and without the need for deep technical expertise » (Karnouskos 2020, p. 139). Ceci couplé avec la possibilité de propager rapidement ces images à travers les réseaux sociaux fait que tout cela est « new and exciting but also worrying » (Karnouskos 2020, p. 139)

2.1.2 Les craintes liées aux deepfakes

Malgré l'excitation qu'un tel développement technologique nous permette de créer si facilement des images et vidéos de choses qui ne sont jamais arrivées, il y a malheureusement toujours des personnes pour utiliser les nouvelles technologies pour faire du mal et exploiter d'autres personnes. « Technological progress […] promotes the good and bad in people, moving us forward and backward at the same time » (Kietzmann et al. 2020, p. 141). Même lorsque l'utilisation de deepfakes est faite à des fins humoristiques et est clairement détectable en tant que deepfake, comme c'est le cas du deepfake Brie/Carrey, techniquement, Brie et Carrey n'ont pas donné leur autorisation pour que leur image soit utilisée de cette façon.

Dans des cas plus sérieux, des personnes peuvent découvrir que des deepfakes pornographiques ont été créés en utilisant leur image contre leur gré. Cela a été le cas en septembre 2023, en Espagne, lorsqu'un groupe d'adolescentes mineures ont découvert que des images d'elles nues avaient été générées par IA et avaient commencé à être partagées dans la ville. L'une de ces jeunes filles a même été victime d'une tentative d'extorsion par un

adolescent qui lui a ensuite envoyé une de ces photos deepfake lorsqu'elle a refusé de lui donner de l'argent (Guy 2023).

Malheureusement, cet usage illégal de deepfakes, présent depuis les débuts de la technologie, est l'une des principales conséquences négatives de la propagation de la technologie deepfake, comme il est mentionné par plusieurs auteurs (Karnouskos 2020, p. 143; Kietzmann et al. 2020, p. 142; Fallis 2021, pp. 623-624; Harris 2021, pp. 13385-13388). Ce type de crime est d'ailleurs genré puisque, bien que pas les seules, se sont souvent les femmes qui sont victimes de telles actions. De manière plus importante, les cas de pornodivulgation, qui est la diffusion non consensuelle d'images ou vidéos à caractère sexuel, « is expected to increase considering the high quality as well as easiness that deepfakes bring into the table » (Karnouskos 2020, p. 143) et cela est d'autant plus sinistre lorsque nous savons qu'environ un américain sur 25 en est victime.

Une autre menace des deepfakes, l'une des premières auxquelles nous pensons, est la possibilité d'une pandémie de désinformation et une perte de confiance totale en ce que nous voyons dans les médias, posant une menace à la connaissance car les « deepfakes reduce the amount of information that videos carry to viewers » (Fallis 2021, p. 624).

Plusieurs philosophes pensent que les deepfakes sont une menace épistémique, qui peut « easily lead people to acquire false beliefs » (Fallis 2021, p. 625). En effet, Fallis argumente que, puisqu'il nous est impossible d'être visuellement témoin de tout ce qui se passe autour du monde, nous dépendons donc beaucoup des vidéos pour enrichir notre connaissance. Malgré le fait que d'autre sources d'information pourraient être disponibles pour réfuter un deepfake, les alternatives ne sont pas toujours « equally reliable ». Nous n'avons pas toujours la possibilité d'être directement témoin de ce qui se passe et les photos sont psychologiquement moins impactantes car contrairement à un photo, une vidéo « extends beyond a single instant, and […] contains context that single photos lack » (Harris 2021, p. 13378).

La menace épistémique n'est pas uniquement présente dans le cas où une vidéo est un deepfake, mais aussi dans le cas de vidéos réelles. Une grande partie de la population est désormais consciente de l'existence des deepfakes et de la possibilité que ce qu'elle voit dans les médias puisse être faux. Ceci peut créer un climat de suspicion où, par conséquent, les personnes pourraient être systématiquement méfiantes des vidéos, même lorsqu'elles sont authentiques. Tout cela pour se protéger contre la possibilité d'être trompé. Cette situation est préoccupante car les vidéos sont pourtant considérées comme la meilleure manière d'acquérir des connaissances. Si la confiance dans les vidéos disparait, ces dernières pourraient complètement perdre leur crédibilité mais surtout perdre leur capacité à transmettre des informations et de la connaissance (Fallis 2021, pp. 626-627).

Dans son article, Fallis contre-argumente certaines objections qui pourraient être faites à son encontre. Par exemple, il pourrait être argumenté que les deepfakes ne sont pas si parfaits et indistincts de vidéos réelles. Une vidéo dans laquelle un politicien apparait en train de parler une langue qu'il ne parle pas du tout sera démasquée rapidement comme un deepfake. Fallis répond à cela que même si aujourd'hui les deepfakes ne sont pas parfaits, la technologie s'améliore très rapidement et cela pourrait clairement devenir un problème dans un futur proche (Fallis 2021, p. 634; Harris 2021, p. 13375).

Pourtant, cela ne veut pas dire que les vidéos vont transmettre moins de connaissances réelles puisque même avant l'avènement des deepfakes la falsification de vidéos existait déjà. Nous pouvons citer les films propagandistes des Nazis, datant de la Deuxième Guerre Mondiale, qui montraient à quel point les juifs étaient bien traités sous leur joug. Ici, même si la vidéo a réellement été tournée et est dans un sens « vraie », nous savons très bien que son contenu est actuellement complètement faux et qu'il ne faut absolument pas y croire. De plus, il n'a pas fallu attendre l'existence des deepfakes pour que les gens soient « worried about videos being fake » (Fallis 2021, p. 626) puisque plusieurs personnes n'ont pas cru aux vidéos montrant les premiers pas de l'Homme sur la lune en 1969 (Fallis 2021, p. 626).

Même avant les deepfakes, il était facile de manipuler une vidéo sans pour autant que le contenu de celle-ci soit faux. Parmi les techniques utilisées pour cela, il y a la suppression de parties de la vidéo ou tout simplement une modification du cadrage afin de cacher une partie de ce qui a été filmé (Harris 2021, p. 13377). Un exemple récent de cette dernière technique est la propagation d'une vidéo propagée dans le but d'essayer de décrédibiliser le président américain Joe Biden en le faisant passer pour un vieil homme qui n'a plus toute sa tête. En effet, Biden a été filmé lors de la récente rencontre du G7 en train de regarder ailleurs, semblant distrait, alors qu'une photo commémorative était sur le point d'être prise. Le président a alors dû être rappelé à attention. Plusieurs médias, notamment républicains, se sont empressés de diffuser la vidéo afin de discréditer Biden. En réalité, en regardant la version plus large de la vidéo, Biden était simplement en train de féliciter des parachutistes (Leingang 2024).

Tous ces exemples prêtent à imaginer que dans le futur, nous ne pourrons plus faire confiance à aucune vidéo que nous voyons dans les médias, de peur que celle-ci soit en fait un deepfake. Cependant, il faut prendre en compte les sources qui partagent ces vidéos et images, car ce sont bien celles-ci qui influenceront leur crédibilité et non pas le contenu en soi : « An audience may find even the most realistic video evidence unconvincing when it is delivered by a dubious source » et inversément « an audience may find even weak video evidence compelling so long as it is delivered by a trusted source » (Harris 2021, p. 13374).

Il nous suffirait donc de vérifier que les sources qui partagent ou créent les vidéos sont des sources dignes de confiance. Pour les personnes lambda, cette vérification semble assez simple puisqu'il suffirait de se baser sur la réputation des sources ou d'utiliser des outils disponibles en ligne qui permettent la vérification (Ministère de l'économie, des finances et de la souveraineté industrielle et numérique 2024). Cependant, il est plus difficile pour les médias de faire de même. En effet, leur rôle est déjà de partager et d'informer sur la vérité. Pour cela, les médias se doivent déjà de contrôler les sources de ce qu'ils partagent, mais avec les deepfakes ils devront être encore plus vigilants et consciencieux dans leurs vérifications (Harris 2021, pp. 13382-13384).

Pour Harris, le problème que pourraient avoir les deepfake n'est pas épistémique, mais plutôt psychologique. Nous avons déjà mentionné les deepfakes pornographiques qui sont de manière évidente néfastes et une claire atteinte à l'intégrité des personnes victimes d'un tel acte. Cependant, même si les deepfakes ne sont pas réalistes ou sont clairement faux, cela peut tout de même être problématique car même sans le vouloir et même en sachant que c'est faux, des associations mentales peuvent tout de même être faites. En effet, il a été prouvé par différentes études, que « even brief exposures to fictious video clips depicting members of a racial group engaging in either aggressive or harmonious activity affects subsequent measures

of implicit associations » (Harris 2021, p. 13386). Les deepfakes pourraient donc, même si évidemment faux, influencer la perception que nous pourrions avoir d'une personne et même « reinforce existing xenophobic or racist stereotypes by problematically depicting members of marginalized groups » (Harris 2021, p. 13388).

Nous avons donc présenté différentes menaces que posent les deepfakes, mais dans la section suivante nous présenterons différentes solutions et techniques pour reconnaître des deepfakes et ne pas être trompés par ceux-ci.

2.1.3 Détecter les deepfakes

Comme nous l'avons déjà mentionné plusieurs fois, malgré le fait que la technologie deepfake soit impressionnante et très réaliste, elle n'est pas encore parfaite et il existe des moyens et des techniques pour reconnaitre les deepfakes.

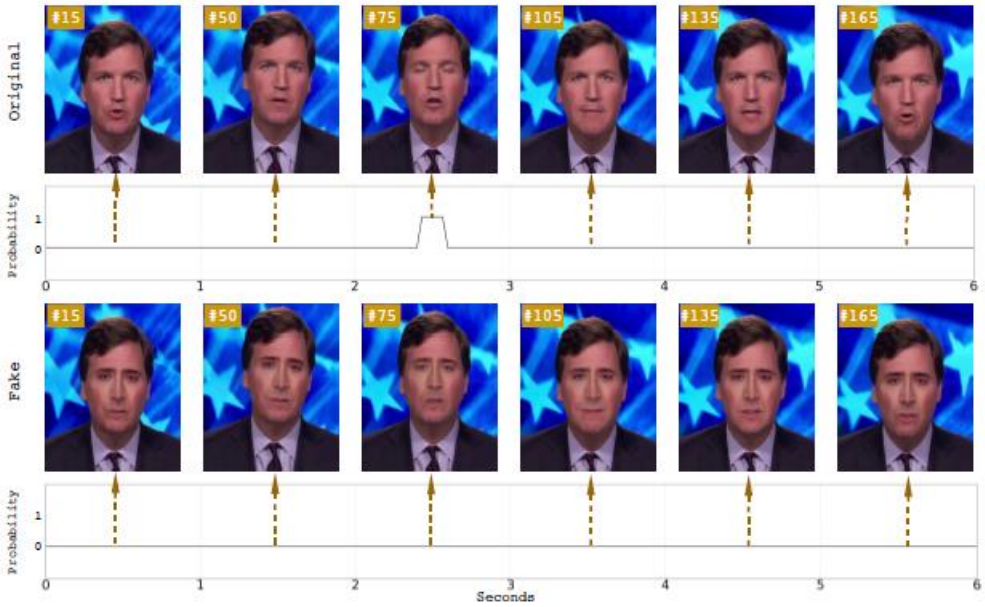
> *« While traditional media forensic methods based on cues at the signal level […], physical level […], or semantic level […] can be applied for this purpose, the AI generated fake face videos pose challenges to these techniques » (Li, Chang, Lyu 2018, p. 1).*

L'expérience de Li et al. a pour but d'étudier et détecter l'absence de signaux physiologiques humains, tels que la respiration ou le mouvement des yeux, dans les vidéos générées par IA afin de pouvoir les reconnaitre. Les auteurs ont décidé de se concentrer uniquement sur le clignement des yeux et utilisent un nouveau modèle de deep learning se basant sur un réseau de neurones convolutifs (CNN pour convolutional neural network en anglais) associé à un réseau de neurones récurrent (RNN pour recursive neural network). Le modèle CNN, ne peut distinguer l'ouverture et la clôture des yeux qu'en se basant uniquement sur des séquences uniques et ne prend pas en compte des données temporelles. C'est pourquoi le RNN, qui lui prend en compte ces données temporelles et peux conserver de l'information, est associé au CNN pour créer un nouveau modèle. Celui-ci est appelé un réseau de neurone convolutifs récurrents à long terme (LRCN pour long-term recurrent convolutional neural network). Pour résumé, un LRCN va tirer parti des forces de chacun des deux autres réseaux et va :

> *« [incorporate] the temporal relationship between consecutive frames, as eye blinking is a temporal process which is from opening to closed, such that LRCN memorize the long term dynamics to remedy the effect by noise introduced from single image » (Li, Chang, Lyu 2018, p.1-3).*

La raison pour laquelle le clignement des yeux est parfois absent des vidéos générées par IA est parce que les bases de données utilisées pour l'entrainer possèdent rarement des visages de personnes avec les yeux fermés. C'est pourquoi l'absence de clignement des yeux est l'un des principaux signes qu'une vidéo a été générée par intelligence artificielle. En effet, en moyenne, une personne cligne des yeux 17 fois par minute, c'est-à-dire 0.283 clignements par seconde, augmentant si la personne est en pleine conversation ou diminuant si elle est en train de lire. Dans un exemple de deepfake, représenté à la Figure 3, il est noté qu'en 6 secondes, la personne n'a pas cligné des yeux une seule fois ce qui est « abnormal from the physiological point of view » alors qu'au contraire un clignement a lieu dans la vidéo originale (Li, Chang, Lyu 2018, p. 2).

Figure 3: Exemple d'absence de clignement des yeux dans le deepfake



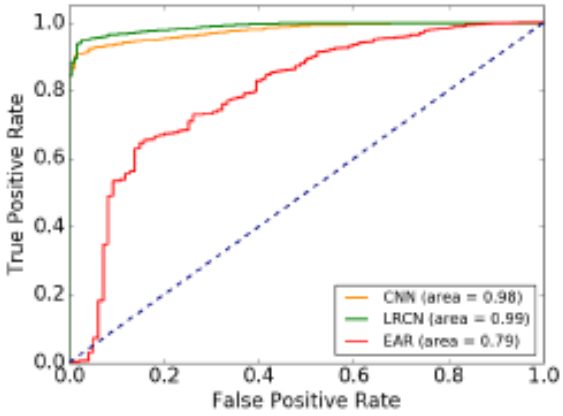
(Li, Chang, Lyu 2018, p. 2)

La méthode utilisée par Li et al. consiste tout d'abord en une étape de préparation, dans laquelle les visages sont détectés dans chaque séquence de la vidéo. Ces visages sont ensuite alignés dans un système qui détectera chaque mouvement de tête et changement d'orientation. Une région, correspondant à chaque œil, est ensuite extraite pour former une séquence. Finalement, le clignement d'œil est détecté en déterminant le degré d'ouverture de l'œil dans chaque séquence de la vidéo en utilisant le modèle LRCN expliqué précédemment (Li, Chang, Lyu 2018, p. 3).

Leur expérience consiste ensuite à entrainer le modèle LRCN, puis à comparer ses performances avec deux autres méthodes : Eye Aspect Ratio (EAR) et CNN. La méthode EAR se base sur la distance entre les paupières pour analyser et déterminer l'ouverture des yeux et est une méthode plus rapide, mais puisqu'elle dépend de la détection des points des yeux, elle est un peu moins fiable. Le résultat de leur étude est présenté par le graphe de la Figure 4, et celui-ci nous montre que globalement, la méthode LRCN est la meilleure méthode pour détecter le clignement des yeux et donc la meilleure méthode à utiliser si nous voulons démasquer des deepfakes en nous basant sur la caractéristique de l'absence de clignement des yeux (Li, Chang, Lyu 2018, pp. 5-6).

Cependant, et les auteurs le mentionnent dans leur conclusion, l'absence de clignement des yeux n'est pas la seule preuve de manipulation d'image, des clignements trop rapides ou fréquents peuvent aussi être suspects. De plus, cette absence étant déjà notée, les créateurs de deepfakes peuvent y prêter particulièrement attention et soit mieux entrainer leurs modèles pour que ceux-ci apprennent à incorporer des clignements d'yeux, soit simplement à les éditer par la suite en post-production (Li, Chang, Lyu 2018, p. 6).

Figure 4: Résultat de la comparaison entre les méthodes LRCN, CNN et EAR



(Li, Chang, Lyu 2018, p. 5)

D'autres manières de détecter des deepfakes incluent également le manque de synchronisation entre les mouvements de la bouche et les paroles prononcées, les différences de luminosité ou encore la qualité d'image inconsistante entre le sujet et le background. Evidemment, si la technologie pour créer des deepfakes s'améliore, les softwares pour détecter dits deepfakes s'améliorent également et permettent de détecter des inconsistances qui ne pourraient être perçues à l'œil nu (Harris 2021, p. 13382).

Une autre manière de détecter des deepfakes, et qui ne compte pas sur le fait de détecter des caractéristiques telles que nous avons mentionné ou l'utilisation d'un software, est l'utilisation de technologies qui peuvent tracer « the genealogy of media content » (Harris 2021, p. 13382). Une telle technologie serait, par exemple, la technologie blockchain. Celle-ci est notamment devenue connue pour son rôle dans les transactions de cryptomonnaie et d'authentification des NFTs (Barber 2023, p. 40). Cependant, son utilisation présente un défaut majeur qui est son coût environnemental élevé. Harris pense que ceci n'est pas un si grand problème car selon lui, la plupart des personnes qui créent des deepfakes, ne désirent pas particulièrement en crée qui ne sont pas du tout détectables, tant que ce n'est pas à l'œil nu. Le fait que leur deepfakes peuvent être reconnus en tant que tels par un software ou autre moyen non visible directement sur la vidéo, ne leur pose pas de problème et nous permet donc d'éviter d'utiliser trop souvent des technologies blockchain coûteuses énergétiquement (Harris 2021, pp. 13382-13383).

Malgré tous ces moyens, même si nous ne pouvons pas facilement contrôler si une vidéo est un deepfake ou pas, nous pourrons normalement toujours faire confiance à certains médias réputés pour leur fiabilité. Même si des logos et caractéristiques vidéos spécifiques à une source peuvent être copiés, les réseaux sociaux et chaînes de télévision officielle ne peuvent pas être utilisés pour diffuser des deepfakes. Ainsi, nous pouvons avoir confiance que ces sources ne présenteraient pas des deepfakes comme étant authentiques (Harris 2021, p. 13384). Souvent, les sources fiables et officielles sont munies d'une marque d'authentification, souvent un petit vu sur les réseaux sociaux, et nous permettent donc de vérifier l'authenticité de nos sources. Il est également possible aujourd'hui de signaler du contenu que nous suspectons comme étant du deepfake (Barber 2023, p. 40).

Kietzmann et al. dans leur article proposent quatre points d'action pour mitiger les risques de la propagation de deepfakes, qu'ils appellent « the R.E.A.L framework ». La première action consiste en l'enregistrement (Record) de la vie d'une personne en traquant ses communications, localisations et activités, afin que si une vidéo deepfake est faite de la personne, elle pourra prouver que ce n'est pas elle dans la vidéo. Cela pose évidemment des problèmes de privacité, mais les auteurs semblent penser que cela vaut la peine pour contrer des deepfakes. La deuxième action est d'exposer (Expose) les deepfakes à travers différentes technologies, parmi lesquels certaines que nous avons déjà mentionné précédemment. La troisième action serait de plaider (Advocate) pour une protection juridique « in instances of defamation, malice, breach of privacy, or emotional distress caused by deepfakes, as well as in cases of copyright infringment, impersonation and fraud » (Kietzmann et al. 2020, p. 145). Les auteurs se demandent si l'implication des réseaux sociaux, à travers lesquels la plupart des deepfakes sont propagés, ne devrait pas être étudiée, voire que ceux-ci soient punis par la loi dans le cas où ils auraient connaissance des deepfakes mais ne prendraient aucune mesure pour lutter contre. La quatrième et dernière action est d'utiliser (Leverage) et améliorer la confiance entre les marques et les client-e-s, afin que ceux-ci restent critiques de ce qu'ils voient et ne soient pas hâtifs dans leur jugement basé sur un deepfake négatif de la marque (Kietzmann et al. 2020, pp. 143-145).

L'éducation est également un bon moyen de prévenir contre les deepfakes, en expliquant ce qu'est un deepfake, quels en sont les éventuels menaces et comment apprendre à les reconnaître et « how to avoid being fooled without simply doubting all images » (Barber 2023, p. 40).

Dans cette section dédiée aux deepfakes, nous avons tout d'abord expliqué la technologie derrière la création de deepfakes qui utilise les réseaux antagonistes génératifs et les deep neural networks qui sont entrainés avec des milliers d'images qui seront ensuite utilisées pour générer du contenu deepfake. Nous avons ensuite survolé quelques craintes et menaces que peuvent poser les deepfakes, notamment épistémologiques et liées à la désinformation, mais aussi les problèmes psychologiques que peuvent développer les victimes de deepfakes négatifs ainsi que l'atteinte à leur intégrité. Pour terminer, nous avons présenté quelques techniques pour détecter des deepfakes, incluant l'absence de clignement des yeux, de variation de lumière et de qualité d'image, l'utilisation de différentes technologies et softwares spécialisés dans la détection de deepfakes, ainsi que l'importance de l'éducation et de la vérification des sources d'où nous tirons nos informations.

Le portrait que nous avons dressé peut paraitre plutôt sombre, mais en appliquant les quelques stratégies présentées pour éviter d'être trompés, nous pensons que l'avènement des deepfakes n'est pas si terrible. Malgré le fait que certain-e-s chercheuses et chercheurs voient le futur de manière pessimiste, nous pensons que tant que nous faisons attention à bien vérifier nos sources et sommes conscients des menaces potentielles, nous serons capables d'éviter un drame informationnel.

2.2 Perception des images générées par IA

L'essor et les avancées technologiques impressionnantes de l'intelligence artificielle ont conduit à une récente prolifération des images générées par IA, que ce soit de l'art ou des images censées être réalistes, comme, par exemple, des visages qui sont parfois indiscernables de visages de personnes réelles. Dans la section précédente, nous avons parlé de deepfakes, de la façon dont ils sont créés et quelles étaient les potentielles menaces d'une

propagation de deepfakes. La différence entre des deepfakes et des images générées par IA est que les deepfakes sont basés sur une vidéo déjà existante et l'IA est utilisée pour modifier la vidéo existante. Alors que les images générées par IA sont entièrement créées par l'IA, basées sur le matériel utilisé pour entrainer celle-ci.

Dans cette section de notre revue de la littérature, nous allons nous concentrer sur l'intelligence artificielle et particulièrement sur la perception que nous, en tant qu'êtres humains, avons des images générées par l'IA. Sommes-nous capables de reconnaître si une image est réelle ou générée par IA ?

2.2.1 L'IA à la rescousse des enfants disparu-e-s

Avant d'entrer en détail sur la perception et la capacité à discerner les images réelles d'images générées par intelligence artificielle, nous souhaitons également mentionner que le réalisme des images générées par IA, que nous soyons capables de les reconnaitre ou pas, peut être bénéfique humanitairement. En effet, la technologie GAN peut être utilisée dans les cas de disparition d'enfants par la génération de photos sur lesquelles les enfants disparu-e-s sont plus âgé-e-s, selon l'âge que ces enfants auraient au moment de la génération de l'image.

D'après plusieurs rapports, près de 30% des victimes de traffic humain sont des enfants, dont une majorité sont des filles, comme c'est le cas en Inde où le « victim ratio is 1 : 6 for boys to girls » (Chandaliya, Nain 2022, p. 1). Beaucoup des enfants kidnappé-e-s sont généralement très jeunes et donc changent beaucoup avec les années qui passent et pourraient ne pas être reconnaissables en les comparant uniquement avec des photos, parfois de mauvaise qualité, de lorsqu'iels étaient petit-e-s. C'est pour cette raison que Chandaliya et Nain ont développé un modèle, appelé ChildGAN, basé sur un *autoencoder* variationnel (VAE) et un GAN qui peut « automatically generate visually realistic face photos, while attaining enhanced face-recognition, age-estimation, and gender-preservation rates » (Chandaliya, Nain 2022, p. 2).

L'*autoencoder* variationnel est un modèle d'apprentissage automatique qui possède un encodeur qui va transformer les données *input* en une représentation latente et un décodeur qui va reconstruire ces données basées sur la représentation latente. À la différence d'un *autoencoder* normal, le VAE « introduces an additional condition that forces the latent representation *z* to follow a Gaussian distribution » (Dahmani et al. 2019, p. 2599). En résumé, un VAE favorise une couverture maximal de l'espace et permet une meilleure combination entre différents vecteurs latents (Dahmani et al. 2019, p. 2599).

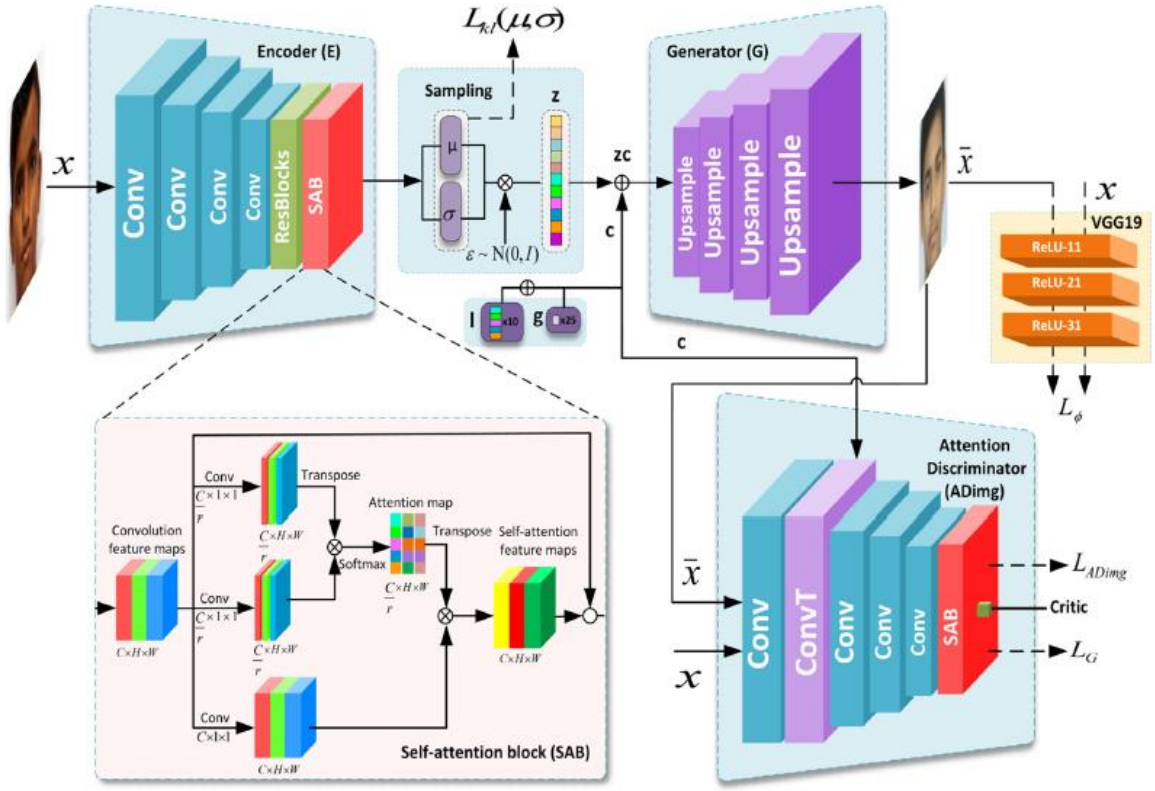
Le modèle ChildGAN est composé de trois parties qui sont représentées à la Figure 5 ci-après. La première est l'encodeur qui va transformer l'image d'entrée *x* en un vecteur d'identité *z* à l'aide de quatre couches de convolution. Chaque couche réduit la taille de l'image et augmente le nombre de cartes caractéristiques. Un réseau résiduel avec neuf blocs résiduels et un bloc d'auto-attention est ensuite utilisé après les couches de convolution pour simuler le vieillissement et capturer les dépendances à long terme entre les régions du visage. Le résultat est un tenseur de taille 512 x 8 x 8 converti en une couche entièrement connectée avec 32'768 caractéristiques, puis réduit à un vecteur de 50 dimensions. Le vecteur d'âge et de sexe est combiné avec *z* pour former un vecteur final de 125 dimensions, permettant de changer l'âge tout en conservant les caractéristiques identitaires du visage (Chandaliya, Nain 2022, p. 3).

La deuxième partie est ce que les auteurs appellent générateur mais qui en fait agit comme un décodeur. Celui-ci va concaténer le vecteur *z* avec un vecteur conditionnel *c* afin de les transformer en 32'768 caractéristiques puis en un tenseur de taille 512 x 8 x 8. Ce tenseur est

ensuite passé à travers quatre couches de convolution pour agrandir l'image (Chandaliya, Nain 2022, p. 3).

Finalement, la troisième partie importante du modèle est le discriminateur d'attention qui est composé de quatre couches de convolution qui réduisent la taille des images et augmentent le nombre de canaux. Sa fonction est surtout de forcer le générateur à produire des visages réalistes (Chandaliya, Nain 2022, p. 3).

Figure 5: Représentation du modèle ChildGAN avec ses trois parties principales



(Chandaliya, Nain 2022, p. 5)

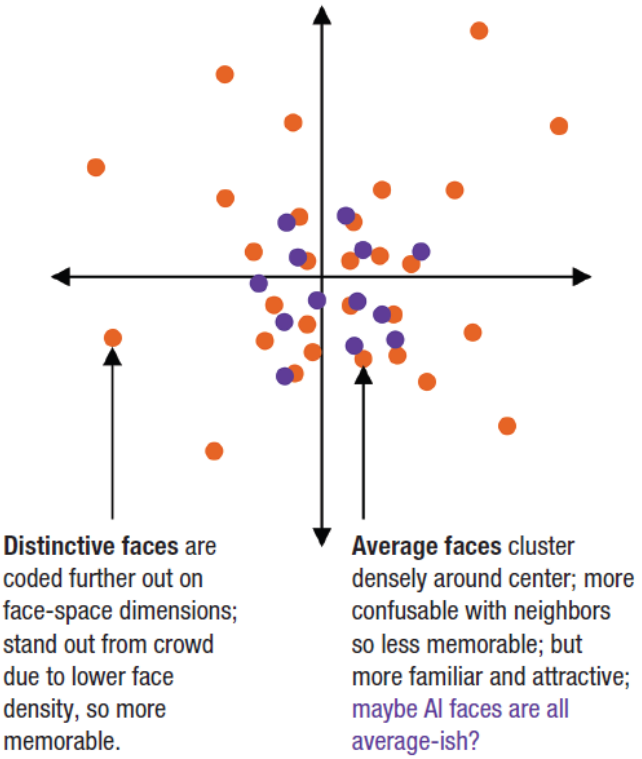
Les chercheurs ont ensuite comparé leur modèle ChildGAN avec d'autres modèles connus et ont déterminé que leur modèle générait des images correctes que ce soit en vieillissant ou en rajeunissant et qu'il était capable de garder une certaine cohérence dans les caractéristiques physiques du visage. Même lorsque le visage était obstrué par des lunettes ou des chapeaux, par exemple, le modèle était toujours capable de produire des résultats satisfaisants. De plus, même si de base ce modèle était basé sur des image d'enfants indien-ne-s, en ajoutant plus de matériel varié pour son entrainement, ses résultats sont également bons avec des enfants d'autre ethnies (Chandaliya, Nain 2022, pp. 8-9).

2.2.2 L'hyperréalisme des visages générés par IA

L'utilisation d'intelligence artificielle pour retrouver des enfants disparus est donc l'une des nombreuses façons que les algorithmes générateurs de visages peuvent être utilisés. Cependant, de plus en plus d'études paraissent, parlant d'hyperréalisme de l'IA. Les images, notamment les visages, générées sont si réalistes, que l'être humain n'est même plus capable de les distinguer d'images réelles et peut aller jusqu'à percevoir les image IA comme étant plus réalistes que des images réelles.

Il existe une hypothèse en psychologie, appelée *face-space theory*, qui nous dit que les caractéristiques des visages sont représentés cognitivement par des points dans un espace multidimensionnel où plus les points sont proches, plus les visages sont similaires, et plus les points sont éloignés plus les visages sont différents (Valentine, Lewis, Hills 2016, p. 1998). De plus, plus les points sont proches du centre, plus les caractéristiques sont considérées comme communes ou prototypiques. À l'inverse, plus les points sont éloignés du centre plus les caractéristiques sont uniques et distinctives. De manière générale, les visages humains sont « normally distributed within this space in such a way that more average features […] are statistically overrepresented » (Miller et al. 2023, pp. 1390-1391). Les algorithmes génératifs étant entrainés sur ces caractéristiques et étant généralement biaisés envers les caractéristiques les plus communes statistiquement, il fait donc sens que les visages générés par IA soient considérés comme « communs ». De plus, comme nous pouvons le voir dans la figure 6, ces visages ayant plus de caractéristiques communes, les points représentant les visages générés par IA, représentés en violet dans le schéma, auront plus tendance à être proche du centre du *face-space* (Miller et al. 2023, pp. 1390-1391).

Figure 6: Représentation des visages humains (orange) et générés par IA (violet) dans le face-space



(Miller et al. 2023, p. 1391)

Ceci pourrait expliquer pourquoi, d'après plusieurs expériences, les visages générés par IA sont plus difficilement distinguables des visages réels. C'est ce que rapportent en tout cas les résultats de la recherche qu'ont mené Nightingale et Farid. Pour chacune des expériences, il a été demandé à différents groupes de participant-e-s, composés entre 219 et 315 personnes par groupe, de juger des visages réels et synthétiques. Ces visages étaient sélectionnés parmi une base de données contenant 800 visages, 400 synthétiques et 400 réels similaires aux synthétiques, et équitablement répartis en termes d'ethnie et de genre. Pour chaque

expérience, 128 images différentes de cette base de données ont été sélectionnées pour chacune des expériences, assurant ainsi une plus grande variété dans les échantillons testés (Nightingale, Farid 2022, pp. 1-3).

Dans la première expérience, les participant-e-s devaient simplement juger si les visages leur semblaient réels ou non. Les résultats montrent une moyenne de réussite de 48,2%, c'est-à-dire que les participant-e-s ont été capable de déterminer de manière correcte si un visage était réel ou pas à 48% en moyenne. Les auteurs ont également examiné les résultats dépendamment du genre et de l'ethnie des visages présentés et ont trouvé que les visages d'hommes blancs étaient moins souvent jugés correctement comme réels ou synthétiques. Ils expliquent ceci par le fait que les visages d'hommes blancs sont prédominants dans les bases de données d'entrainement des algorithmes et de ce fait, générés de manière bien plus réaliste par l'IA (Nightingale, Farid 2022, p. 1).

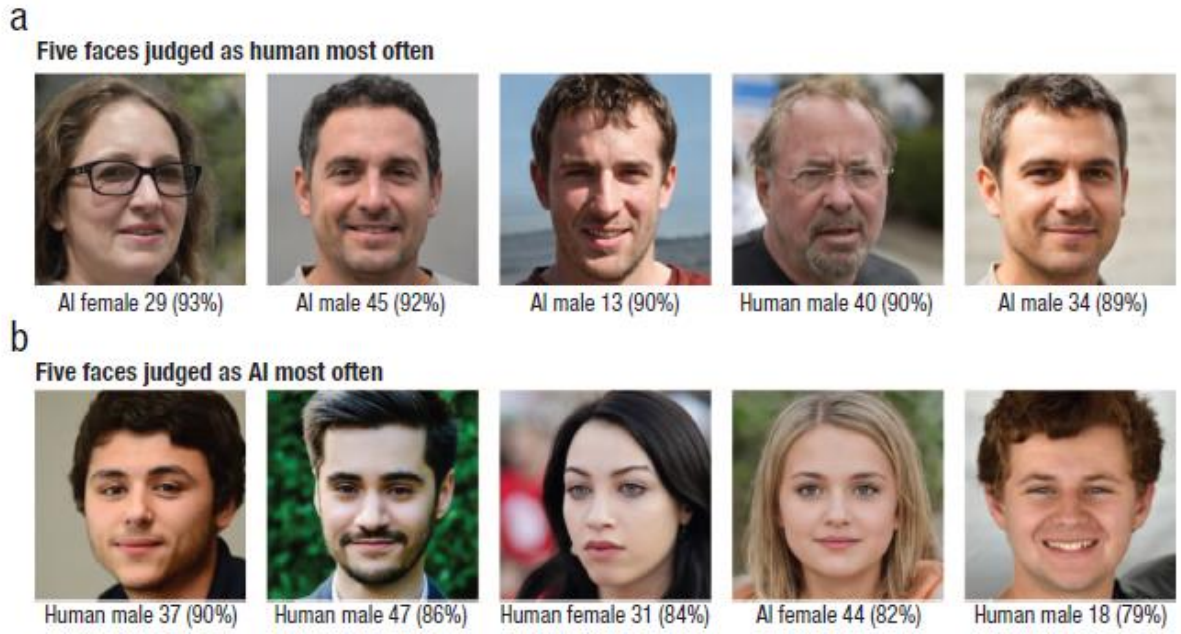
Similairement à la première expérience, la deuxième expérience consiste également à déterminer si des visages sont réels ou générés par IA. La différence est que les participant-e-s ont été entrainé-e-s et reçoivent des commentaires après chaque image. Après le jugement d'une première moitié du set, 64 images, leur taux d'exactitude atteignait les 59,3% en moyenne, une amélioration de 10% par rapport à la première expérience. Après la deuxième partie de l'expérience, aucune amélioration n'est notée avec la moyenne qui reste similaire, à 58,8%. Encore une fois, les auteurs ont remarqué que les visages d'ethnie blanche étaient plus difficilement correctement déterminés. De plus, « the lack of improvement over time suggests that the impact of feedback is limited, presumably because some synthetic faces simply do not contain perceptually detectable artifacts » (Nightingale, Farid 2022, p. 1).

Pour résumer les résultats de la recherche de Nightingale et Farid, il est désormais difficile de différencier si des visages sont réels ou générés par une IA, le cas étant plus flagrant avec des visages d'hommes blancs car ceux-ci sont plus récurrents que d'autres types de visages dans les bases de données d'entrainement des algorithmes (Nightingale, Farid 2022, p. 1).

Miller et al. ont voulu reprendre et analyser plus loin les résultats de Nightingale et Farid en se concentrant uniquement sur les visages de personnes blanches et avec uniquement des participant-e-s d'ethnie blanche afin d'éviter « out-group effects in humanness ratings » (Miller et al. 2023, p. 1393). Pour leur expérience, les auteurs ont demandé aux participant-e-s de leur expérience de juger si un visage leur paraissait réel ou généré par IA, puis de juger la confiance qu'iels avaient en leur réponse et, finalement, de nommer quels éléments de l'image avaient contribué à leur choix à travers une question ouverte (Miller et al. 2023, pp. 1392-1393).

L'effet d'hyperréalisme des visages synthétiques blancs noté par Nightingale et Farid a pu être répliqué dans l'expérience de Miller et al. En effet, les auteurs, avec un set d'images différents et des participant-e-s différent-e-s, ont enregistré que les visages synthétiques ont été déterminés en tant que visages humains à 65,9%, alors que les visages humains n'ont atteint que les 51,1%. Dans la Figure 7 ci-dessous sont présentés les cinq visages jugés comme plus humains et les cinq visages jugés comme plus synthétiques. Il est intéressant de noter la force de l'hyperréalisme car parmi les cinq visages considérés comme les plus réels, quatre sont synthétiques. Et inversement, parmi les cinq visages considérés comme synthétiques, quatre sont réels (Miller et al. 2023, p. 1394).

Figure 7: Images jugées comme les plus réels (a) et les plus synthétiques (b)



(Miller et al. 2023, p. 1394)

À l'inverse de ce que les auteurs avaient prédits, les participant-e-s ayant jugé incorrectement les visages synthétiques comme étant réels étaient généralement plus confiants en l'exactitude de leur choix, « indicating that the tendency for AI hyperrealism is exacerbated by overconfidence » (Miller et al. 2023, p. 1395). Les caractéristiques dont les participant-e-s se sont aidés pour déterminer si un visage était synthétique a permis de créer un framework de 21 thèmes principaux et 20 sous-thèmes. Parmi les thèmes principaux, nous retrouvons en majorité la peau et les rides, la qualité de l'image dans lequel nous avons les sous-thèmes clarté et flou ainsi que rendu graphique qui sont majoritaires, et caractéristiques du visage dans lequel nous avons majoritairement les sous-thèmes yeux et dents qui ont été spécifiés par les participant-e-s (Miller et al. 2023, p. 1395). Pour que les personnes ne différencient pas entre des visages synthétiques ou humains, Miller et al. théorisent qu'il doit exister « some visual differences between AI and human faces, which people misinterpret » (Miller et al. 2023, p. 1395).

L'âge et le genre de la personne participant à l'expérience peut également influencer la perception, comme ont été étonnés de voir Tucciarelli et al. Lors d'une de leur expérience, et comme pour beaucoup d'autres recherches similaires, les auteurs ont demandé à des participant-e-s de juger des visages en tant que réels ou synthétiques. Puis, lors de l'analyse des résultats, en plus de l'analyse globale, les chercheurs et chercheuses ont également calculé si une différence de genre et d'âge était significative statistiquement. D'après leurs résultats, la réponse est positive : les hommes ont jugé les visages, que ceux-ci soient synthétiques ou réels, comme étant réels plus souvent que les femmes (Tucciarelli et al. 2022, pp. 4-5).

Il a également été noté que plus l'âge des participant-e-s était élevé, plus la probabilité de déterminer un visage synthétique comme étant réel augmentait. Cela pourrait confirmer le fait que les personnes nées après 1980, et souvent référées comme des *digital natives*, interagissent différemment avec la technologie et les médias et semblent être moins souvent

victimes de désinformation et posséder une meilleure *digital literacy*. L'âge n'est évidemment pas l'unique facteur qui peut influencer les compétences numériques d'une personne « psychological factors, social influence, and actual use of digital technologies » (Tucciarelli et al. 2022, p. 15) peuvent également avoir un impact. Tucciarelli et al. n'ont pas recherché plus loin ce phénomène, mais leurs résultats, comme l'ont montré d'autres études également, vont dans le sens qu'effectivement des personnes plus jeunes seraient moins susceptibles de se faire piéger par ce qu'elles verraient dans les médias et sur les réseaux sociaux. Cela expliquerait donc pourquoi les personnes plus jeunes ayant participé aux expériences de Tucciarelli et al. trouvaient les visages synthétiques moins réels (Tucciarelli et al. 2022, pp. 4-5 et p.15).

Revenons-en cependant à l'étude de Miller et al. cherchant à déterminer quelles caractéristiques étaient déterminantes selon leurs participant-e-s à juger un visage comme synthétique ou réel. Les auteurs ont donc mis en place une deuxième expérience avec un nouveau set de participant-e-s, devant juger des visages réels et synthétiques sur la base de 14 attributs dont 4 provenaient de la *face-space* théorie, 9 provenaient de la première expérience et correspondaient aux 9 attributs les plus mentionnés, ainsi que l'attribut de l'âge. Cependant, les participant-e-s n'étaient pas au courant que des visages synthétiques faisaient partis des visages à juger et celles et ceux qui s'en sont rendu compte n'ont pas été pris en compte dans l'analyse (Miller et al. 2023, p. 1395-1396).

Les résultats de l'expérience confirment que les visages jugés comme réels étaient considérés comme « more proportional, alive in the eyes, and familiar ; and less memorable, symmetrical, attractive, and smooth-skinned » (Miller et al. 2023, p.1396). Dans la continuité des résultats de la première expérience, les visages AI ont été jugés plus réels et, en accord avec l'hypothèse des auteurs, les visages AI « were significantly more average (less distinctive), familiar, and attractive, and less memorable than human faces » (Miller et al. 2023, p. 1397) et explique donc pourquoi les visages synthétiques se retrouvent plus souvent au centre du *face-space*.

Les visages synthétiques sont donc plus « oubliables » car plus communs. Il est intéressant de mettre en contraste une étude datant de 2015, peu avant l'avènement des réseau antagonistes génératifs, qui a examiné le même résultat, mais basé sur une théorie différente et presque contraire. Balas et Pacella suggèrent que les visages synthétiques, ou artificiels comme ils les appellent dans leur étude, « constitute a class of "other group" faces and may […] be processed less effectively » (Balas, Pacella 2015, p. 332). Leur hypothèse se base sur le fait que les êtres humains ont tendance à mieux se rappeler des visages de leurs propre ethnie, ou en tout cas d'ethnies qu'ils rencontrent souvent. « Other-race faces are generally harder to distinguish from one another than own-race-faces and elicit different neural responses than own-race faces » (Balas, Pacella 2015, p. 331).

Leur expérience est présentée aux participant-e-s comme une expérience de mémoire des visages. Divisé-e-s aléatoirement en deux groupes, l'un ayant uniquement des visages réels et l'autre uniquement des visages artificiels, chaque participant-e a d'abord passé une phase de pré-test durant laquelle, chacun-e a vu 45 visages pendant deux secondes et a dû les catégoriser par sexe aussi rapidement que possible. Dans l'expérience même, il a été dit aux participant-e-s que 90 visages leur seraient présentés, dont les 45 qu'iels avaient déjà vu quelques minutes auparavant. Au lieu de déterminer le genre du visage, les participant-e-s avaient tout le temps nécessaire pour déterminer si ce visage leur avait été montré dans la

phase de pré-test. Comme espéré par les chercheurs, le résultat est que les visages artificiels sont moins mémorables que les visages réels (Balas, Pacella 2015, pp. 332-334).

Il est intéressant de mettre en lien l'étude de Balas et Pacella avec celle de Miller et al. car malgré l'intervalle de huit ans entre les deux et l'étude de technologies bien différentes, les deux recherches mettent en lumière la difficulté des êtres humains à mémoriser des visages synthétiques.

2.2.3 La confiance accordée aux visages IA

En plus d'être hyperréaliste et indistinguable de visages réels, plusieurs études ont montré que nous serions plus enclins à accorder notre confiance à des visages synthétiques, et même sans voir un visage quelconque, nous accordons également plus notre confiance à des conseils provenant de l'intelligence artificielle.

En effet, dans l'étude de Nightingale et Farid dont nous avons déjà parlé précédemment, leur troisième et dernière expérience avait pour thème la confiance accordée aux visages synthétiques. Durant l'expérience, il a été demandé aux participant-e-s de juger non pas si un visage leur paraissait réel ou pas, mais plutôt le niveau de confiance que chaque visage leur inspirait, sur une échelle de 1 à 7, 7 étant très digne de confiance. Les résultats montrent que les visages synthétiques paraissent 7,7% plus dignes de confiance, une différence qui est significative statistiquement (Nightingale, Farid 2022, p.1-2).

Le fait que les visages synthétiques soient considérés comme plus dignes de confiance peut être expliqué par le fait que ceux-ci sont généralement plus « communs », comme nous l'avons mentionné précédemment, et que les visages communs ont une tendance à paraitre plus dignes de confiance (Nightingale, Farid 2022, p. 2).

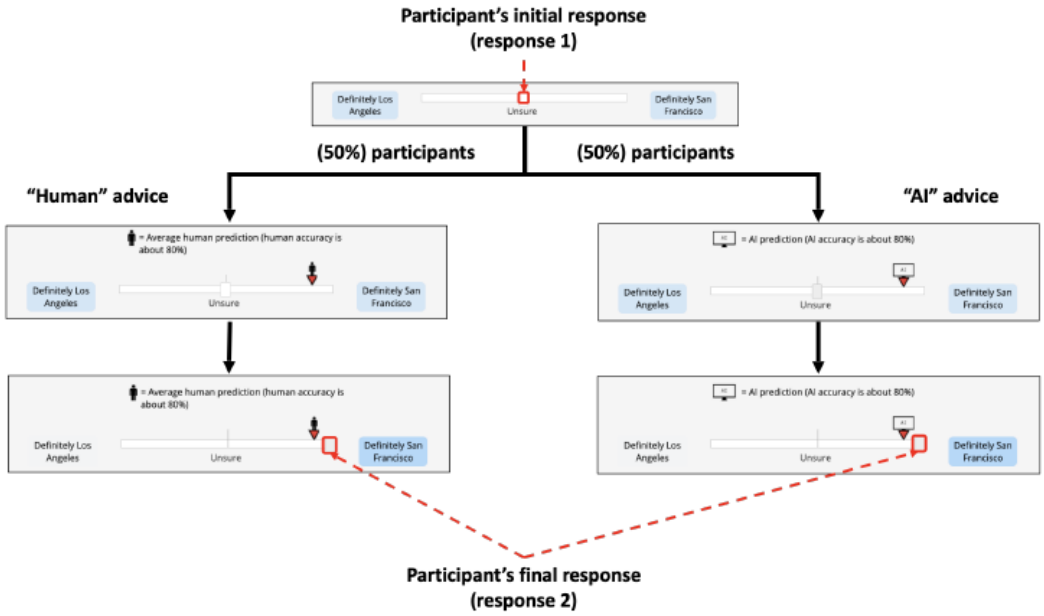
Avec l'avènement de ChatGPT ces dernières années, plusieurs études se sont également intéressées à la confiance accordée aux conseils prodigués par l'intelligence artificielle. Selon certaines, « people utilize automated or AI advice more than advice from peers », ce qui est contredit pas d'autres qui « have mixed results where AI advice has less or similar utilisation » (Vodrahalli et al. 2022, p. 764). Pour tenter d'en avoir le cœur net, Vodrahalli et al. ont mené leur propre étude.

Leurs participant-e-s ont été divisé-e-s en deux groupes, l'un recevant des conseils provenant d'êtres humains, et l'autre recevant des conseils générés par IA. La question est d'abord présentée une première fois, sans aucun conseil, et les participant-e-s doivent sélectionner sur un slider le choix qui leur parait le plus correct. Une fois la réponse enregistrée, la question réapparait une deuxième fois, cette fois avec le conseil de la réponse à choisir. Les labels utilisés pour les conseils étaient identiques tant pour les conseils humains que pour les conseils IA, sauf que le symbole utilisé pour indiquer le conseil variait entre une personne et un ordinateur (voir figure 8) (Vodrahalli et al. 2022, p. 764).

Les questions étaient toutes basées sur le même principe. Si le thème de la question était Art, l'image d'un tableau était présentée avec deux possibilités de période à choisir. Similairement, si le thème était Ville, l'image d'une ville était présentée et il fallait choisir, parmi deux possibilités, quelle ville était représentée. Dans le thème Sarcasme, un texte provenant de Reddit était présenté aux participant-e-s et les participant-e-s devaient sélectionner quel choix était le sarcasme. Finalement, dans le quatrième et dernier thème Recensement, un tableau

de données donnait les informations d'un individu et les participant-e-s devaient décider si la personne gagnait plus ou moins de 50'000$ par an (Vodrahalli et al. 2022, pp. 765-766).

Figure 8: Visualisation de l'étude de Vodrahalli et al. 2022



(Vodrahalli et al. 2022, p. 764)

Les résultats de l'expérience sont résumés au Tableau 1 et présentent la précision (accuracy) des réponses des participant-e-s après avoir reçu conseil ainsi que le taux d'activation (Activation Rate) correspondant au pourcentage de participant-e-s qui ont changé leur réponse après avoir reçu conseil. Nous pouvons voir que les taux sont généralement plus élevés pour les conseils IA pour les thèmes d'Art, des Villes et du Recensement, mais que dans le thème du Sarcasme, les conseils humains sont préférés (Vodrahalli et al. 2022, p. 766). Nous pouvons donc voir que dans une certaine mesure, et selon le type de sujet, les gens ont tendance à faire plus confiance aux conseils donnés par l'IA.

Tableau 1: Résumé des résultats de l'expérience de Vodrahalli et al. 2022

Dataset	Art	Cities	Sarcasm	Census
Accuracy				
Before Advice	65.7%	72.5%	73.1%	70.5%
Δ Human	+7.1%	+3.7%	+3.3%	+2.4%
Δ AI	+11.8%	+6.2%	+3.3%	+3.5%
Activation Rate				
Human	46.0%	48.2%	39.8%	41.8%
AI	52.3%	54.5%	34.0%	47.5%
Δ	+6.4%	+6.3%	−5.8%	+5.7%

(Vodrahalli et al. 2022, p. 768)

Dans cette section dédiée à la perception de visages générés pas IA, nous avons d'abord vu comment les GANs peuvent nous aider par exemple pour retrouver des enfants disparus et

les possibilités de modèles qui peuvent être créent dans un but précis, comme l'a été le modèle ChildGAN. Nous avons ensuite discuté de l'effet d'hyperréalisme qui existe dans les visages synthétiques générés par IA et comment ils sont perçus comme plus réalistes que des visages réels, plus communs et même plus oubliables. Ces deux dernières caractéristiques sont expliquées par la théorie du *face-space* qui place les caractéristiques communes d'un visage au centre d'un espace et dans lequel se retrouvent beaucoup de caractéristiques retrouvées sur des visages synthétiques. Finalement, nous avons parlé du fait que nous semblons faire plus confiance aux visages générés par IA, voire même avoir plus confiance en des conseils provenant de l'intelligence artificielle plutôt que de personnes humaines.

2.3 Conclusion

En conclusion de cette revue de la littérature, nous avons vu que l'avancée rapide de la technologie des réseaux antagonistes génératifs et de l'intelligence artificielle de manière générale nous a fait arriver à un point aujourd'hui où nous pouvons facilement être trompés si nous ne faisons pas attention. En effet, selon plusieurs recherches, dont certaines ont été détaillées dans cette section, l'IA est désormais capable de générer du contenu, surtout des visages humains, si réalistes que nous ne sommes même pas capables de les discerner de visages humains réels. Plus encore, les visages sont si hyperréalistes que certains visages réels semblent synthétiques en comparaison.

Des études ont également montré que ces visages générés par IA paraissent plus dignes de confiance et, même sans qu'il y ait un visuel, il semble que nous ayons commencé à également plus accorder notre confiance aux conseils fournis par IA. Cette tendance peut sembler dangereuse car elle peut mettre à risque la confiance que nous pouvons accorder à ce que nous voyons sur nos écrans et dans les médias. Comme mentionné dans la section traitant de deepfakes, nous courons le risque d'être présentés avec des images et vidéos générées artificiellement bien plus souvent qu'auparavant. Il est important que nous apprenions à reconnaitre les signes d'une image ou vidéo générée par IA ou, lorsque cela est impossible dû au réalisme que l'IA génère désormais, que nous apprenions à vérifier les sources de ce que nous voyons afin de ne pas être trompés.

D'un côté plus optimiste, cette avancée technologique peut nous permettre des choses incroyables tel que générer rapidement des images ou vidéos à but éducatif comme par exemple pour visualiser une période historique lointaine. Ou, comme nous en avons parlé, aider à générer des photos vieillis d'enfants disparus afin de pouvoir plus facilement les retrouver des années plus tard, et ainsi réunir des familles.

3. Notre recherche et ses hypothèses

3.1 Inspiration pour notre recherche

Comme nous l'avons vu, selon plusieurs études, il est non seulement devenu difficile pour les êtres humains de reconnaitre des images générées par IA, mais de plus en plus nous pensons que des images pourtant réelles sont produites par intelligence artificielle. Dans notre recherche, nous souhaitons nous baser sur une théorie des premières impressions qui suggère que même des expositions très brèves nous suffisent pour que nous formions une opinion et une impression.

Pour arriver à cette conclusion, Willis et Todorov ont procédé à plusieurs expériences dans lesquelles ils ont varié le temps d'exposition de différents visages afin de déterminer quel était le temps minimum nécessaire à une personne pour qu'elle se fasse un avis sur un visage. Les différentes caractéristiques qu'ont voulu étudier les auteurs sont : « attractiveness, likeability, competence, trustworthiness, and agressiveness » (Willis, Todorov 2006, p. 593). Une expérience pour chacune de ces caractéristiques a été faite durant lesquelles des visages inconnus ont été présentés à chacun-e des participant-e-s, durant 100ms, 500ms ou 1000ms, et il leur a été demandé de juger chacun des visages selon l'une des caractéristiques ainsi que de déterminer leur niveau de confiance en leur jugement (Willis, Todorov 2006, p.593).

Les résultats de leur recherche ont montré que :

> *« as minimal an exposure as 100ms is sufficient for people to make a specific trait inference from a stranger's face. For all five traits, judgments made after 100-ms exposure to a face were highly correlated with judgments made in the absence of time constraints » (Willis, Todorov 2006, p.596).*

En ce qui concerne le niveau de confiance, les résultats montrent qu'une plus longue exposition renforce simplement les jugements faits par les participant-e-s, c'est-à-dire que « minimal exposure to faces is sufficient for people to form trait impressions, and that additional exposure time can simply boost confidence in these impressions » (Willis, Todorov 2006, p. 597).

3.2 Différence avec d'autres études et hypothèses

Notre expérience se distingue des recherches mentionnées dans la section revue de la littérature car elle est inspirée de l'étude des premières impressions de Willis et Todorov. En effet, l'objectif de notre recherche est de déterminer s'il est possible de reconnaître, en un coup d'œil, si des images sont générées par IA. Dans ce cas, cela indiquerait qu'il existe des failles dans l'algorithme des IA que l'œil humain est encore capable de discerner, même inconsciemment. Si, au contraire, les résultats de notre étude montrent une incapacité à différencier l'IA des images réelles, cela pourrait signifier la nécessité d'une plus grande vigilance face aux images que nous voyons sur le web, et en particulier sur les réseaux sociaux. Une telle vigilance et de l'éducation sont importantes pour prévenir la désinformation et les usages malhonnêtes.

De plus, la plupart des recherches que nous avons analysé ne se sont concentrées que sur un type d'image : les visages humains. Dans notre étude, nous avons également cherché à savoir si différents types d'images auraient les mêmes résultats ou si au contraire, certaines catégories seraient mieux perçues comme réelles comparé à d'autres.

Les hypothèses que nous avons formulées sont les suivantes :

1. Hypothèse 1 : Les participant-e-s sont capables de distinguer les images générées par IA des images réelles.
2. Hypothèse 2 : Le niveau de confiance des participant-e-s est plus élevé pour les images réelles.
3. Hypothèse 3 : Les participant-e-s sont capables de mieux distinguer certaines catégories (telles que les animaux et les visages) comparé à d'autres (telles que les paysages et les tableaux).

4. Méthodologie

4.1 Création et récoltes d'images

Le but de notre étude étant de comparer la perception d'images IA avec des images réelles, il nous a fallu tout d'abord créer un corpus d'images générées par intelligence artificielle. Pour notre recherche, nous avons souhaité créer un corpus d'images de thèmes variés afin d'analyser si les personnes participant à notre étude percevaient mieux certains types d'images comme étant générés par IA ou comme étant réelles. Nous avons choisi quatre catégories d'images : des animaux, des paysages, des tableaux et des visages.

Notre objectif était de générer une cinquantaine d'images photoréalistes pour chacune de nos quatre catégories, le but étant de pouvoir sélectionner les 30 meilleures. Il nous fallait donc trouver des générateurs qui nous permettraient de générer 200 images suffisamment réalistes pour pouvoir être confondues avec des images réelles.

4.1.1 Critères de sélection des générateurs

Pour cela, nous avons en premier lieu exploré et fait un point sur les différents sites qui permettent la génération d'image. Notre budget étant limité, nous nous sommes concentrés sur les sites permettant une génération gratuite. Les possibilités étaient donc restreintes car la plupart des sites soit n'offrent pas de génération gratuite, soit la qualité de téléchargement est basse ou soit ne permettent tout simplement pas le téléchargement gratuitement.

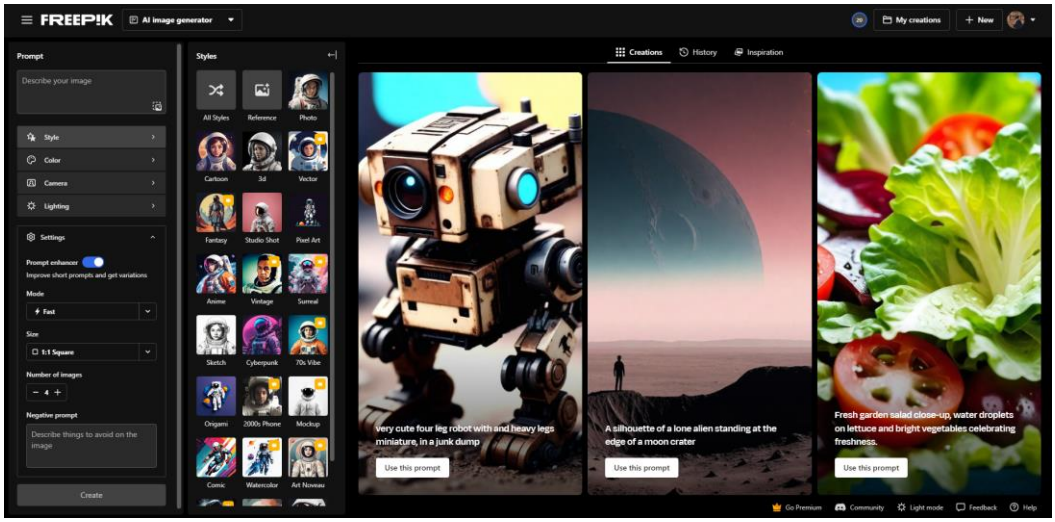
Ces critères en tête, nous avons compilé une première liste de sites web permettant la génération d'image et le téléchargement gratuit. Cependant, notre recherche se basant sur la possibilité de ne pas reconnaître la différence entre une image générée artificiellement et une image réelle, un autre critère de sélection était nécessaire : le réalisme des images générées. Celles-ci devaient être suffisamment photo-réalistes pour pouvoir éventuellement passer pour vraies. La sélection s'est donc vue réduite à trois sites qui nous paraissaient les meilleurs et que nous avons donc utilisé pour générer les images : Freepik, Runway et Face Studio.

4.1.2 Caractéristiques de Freepik et Runway

Dans la plupart, si ce n'est tous, des générateurs d'images, il est demandé à l'utilisateur d'écrire un texte descriptif de ce qui est souhaité être généré. Suivant les sites, des paramètres, plus ou moins nombreux, sont disponibles à la modification afin que l'IA puisse générer des images aussi fidèles que possible à ce qui est souhaité. Certains générateurs possèdent même un paramètre particulièrement intéressant : le negative prompt. Cette fonctionnalité permet de préciser les caractéristiques que nous souhaitons exclure ou éviter lors de la génération des images. Freepik et Runway possèdent tous les deux cette option.

Sur Freepik (Figure 9) par exemple, nous pouvons configurer plusieurs options tel que le style de l'image (cartoon, photo, abstrait, etc.), la couleur (noir et blanc, pastel, sepia, etc.), l'angle de la caméra (cinématique, panorama, portrait, etc.), ainsi que le style de lumière (dramatique, studio, etc.). Dans chacune de ces options, des possibilités gratuites sont mises à disposition, mais une grande partie sont uniquement disponibles avec un compte premium.

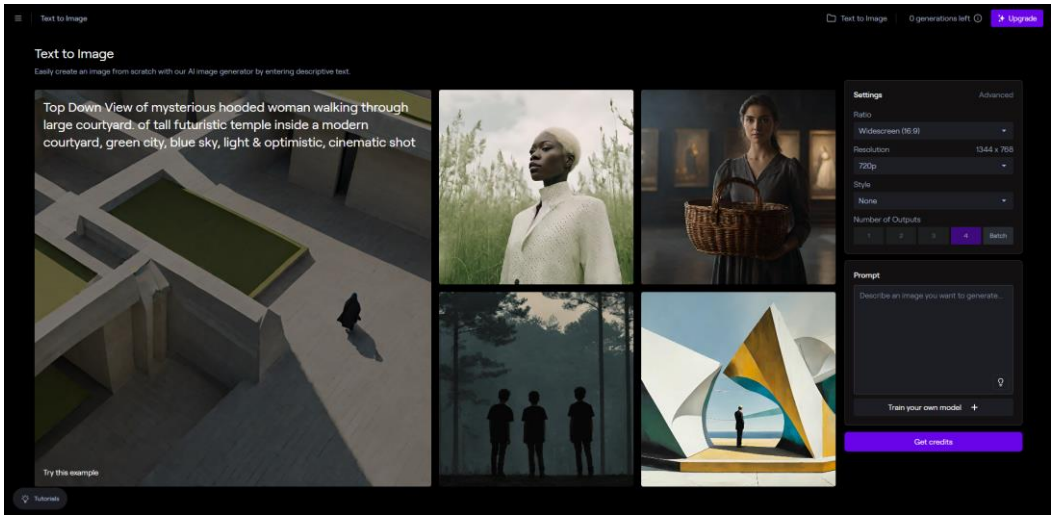
Figure 9: Interface Freepik



(*Freepik AI image generator* 2024)

Runway (Figure 10) ne possède pas autant de paramètres à modifier que Freepik. Nous ne pouvons modifier que le style d'images que nous souhaitons générer. Cependant, le choix est bien plus large sur Runway que sur Freepik car celui-ci bloque la moitié de ses options derrière un abonnement premium. Un paramètre que possède uniquement Runway est le poids du texte descriptif. En effet, nous pouvons choisir, à travers une échelle allant de 0 à 30, le degré de liberté et de créativité accordé à l'IA pour la génération. Le principe semble simple et prometteur, mais lors de la mise en pratique, nous avons remarqué que plus nous limitions la créativité de l'IA, plus le résultat était mauvais et loin de ce que nous souhaitions. Peut-être le problème provenait de notre texte qui n'était pas assez précis, mais nous n'avons donc pas spécialement utilisé cette fonctionnalité et avons gardé l'échelle sur le chiffre de base.

Figure 10: Interface Runway



(*Runway* 2024)

4.1.3 Génération d'animaux, de paysages et de tableaux

Puisque nous avions quatre catégories d'images à générer, nous avons remarqué que certains sites produisaient de meilleurs résultats dépendamment du type d'image souhaité. De manière générale, Runway a produit des images plus réalistes que Freepik qui générait souvent des

images plus stylisées et nous l'avons donc préféré lors de la génération des images d'animaux. Pour les paysages, les deux sites ont généré des images au réalisme adéquat et le nombre d'images sélectionnées est assez équilibré entre les deux. Les tableaux ont surtout été générés sur Freepik, car comme précisé plus tôt, celui-ci gardait souvent un côté plus artistique qui convenait bien à ce type d'images.

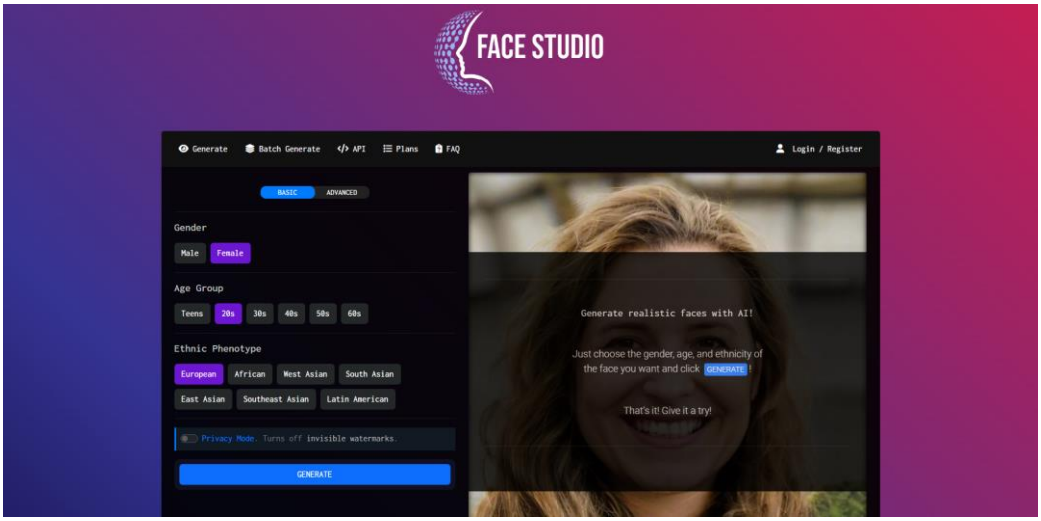
4.1.4 Caractéristiques de Face Studio et génération de visages

En ce qui concerne les visages, aucun de ces deux sites n'ont réussi à générer des images suffisamment réalistes et pertinentes à notre étude. Nous avons donc dû chercher d'autres générateurs et avons découvert Face Studio (Figure 11). Ce site permet uniquement la génération de visages photoréalistes, basés sur différents paramètres démographiques tels que le genre, l'âge et l'ethnicité. Les usager-ère-s possédant un compte payant ont accès à plus de paramètres tels que la couleur de la peau, la longueur des cheveux, le choix entre une expression joyeuse ou sérieuse et bien d'autres.

Face Studio produit des générations de visages très réalistes comparé à d'autres générateurs gratuits, qui gardent souvent un aspect dessiné. De plus, l'avantage de Face Studio est que les visages générés sont des visages communs de personnes que nous pourrions croiser dans la rue. Contrairement à d'autres générateurs qui génèrent des visages très réalistes, mais trop attirants et à la peau trop parfaite, Face Studio est capable de générer des visages avec des imperfections qui les rendent encore plus authentiques et moins artificiels.

Un avantage de Face Studio comparé à Runway et Freepik, est la possibilité de générer des images par batchs de manière gratuite. Une limite de 300 batchs par jour est fixée, mais les générations d'images individuelles sont illimitées, faisant de Face Studio un générateur bien plus généreux avec le nombre de générations gratuites. À l'inverse, Runway offre la génération de seulement 25 images par compte et Freepik offre 20 générations gratuites par jour et par compte.

Figure 11: Interface Face Studio



(*Face Studio* 2024)

4.1.5 Génération des images IA et sélection des images réelles

Une fois les sites de génération déterminés, nous avons donc commencé la génération des images en prenant note des différents textes utilisés pour la génération ainsi que des

différentes options paramétrées afin de pouvoir utiliser différentes options et trouver la plus convenable à ce que nous cherchions. Malgré le fait que nous avons toujours sélectionné les options de manière que les images générées soient photoréalistes, nous avons également toujours écrit dans notre prompt « photorealistic » ou « high quality » pour que l'IA comprenne que nous voulions insister sur ces points.

Après avoir collecté les images générées par intelligence artificielle, nous avons sélectionné des photos similaires à chacune des images générées en cherchant des photos libres de droit sur Pexels, Freepik et Unsplash. À la base, nous voulions également utiliser le même texte que lors de la génération d'images pour la recherche de photos réelles. Cependant, un texte qui doit pourtant être aussi précis et descriptif que possible pour l'IA, ne donne pas toujours les résultats escomptés sur les sites de recherche d'images. En effet, simplifier les termes de recherche nous a permis d'obtenir des résultats plus précis et pertinents et rendu le processus de sélection plus efficace. Les images de tableaux ont été recherchées à travers Wiki Art, qui permet d'effectuer des recherches par courants artistiques, périodes et artistes. Pour les photos de visages, nous les avons sélectionnées parmi les centaines mises à disposition dans la base de données Flickr-Faces HQ (FFHQ) créée par Tero Karras, Samuli Laine et Timo Aila (Karras, Laine, Aila 2019).

Au total, notre corpus final d'images qui seront utilisées durant notre questionnaire était de 240 images équitablement réparties en images réelles et générées par IA, et mélangeant animaux, paysages, tableaux et visages.

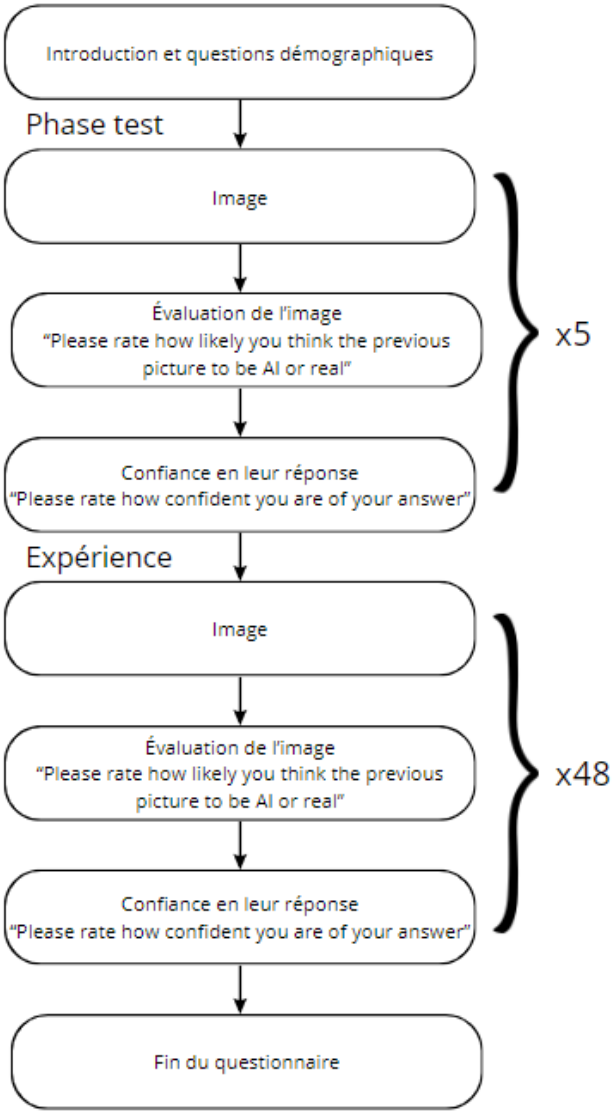
4.2 Création et passation du questionnaire

À travers notre étude, nous désirons analyser la perception humaine des images générées par intelligence artificielle et déterminer si l'être humain peut reconnaitre qu'une image est réelle ou générée par IA lors d'une première impression. Pour cela nous avons décidé de récolter nos données à travers un questionnaire pour plusieurs raisons.

Tout d'abord, un questionnaire permet de recueillir des données rapidement et auprès d'un grand nombre de personnes. Les questionnaires étant auto-administrés, il n'existe pas d'interaction entre les enquêteur-ice-s et les participant-e-s, réduisant ainsi la possibilité de biais. Les questionnaires permettent également aux participant-e-s de garder leur anonymat et simplifie l'étape d'anonymisation des données pour les chercheur-euse-s. Nous avons décidé d'utiliser Qualtrics comme outil de création de notre questionnaire.

Notre questionnaire, dont un schéma est disponible à la Figure 12, se compose tout d'abord d'un message d'introduction dans lequel nous expliquons qui nous sommes, le but de notre étude et le déroulement du questionnaire. Après cela, il est demandé aux participant-e-s de donner leur consentement à ce que leurs réponses soient utilisées dans notre étude. S'en suit une section de questions démographiques demandant aux participant-e-s de nous donner leur genre et leur âge. Les participant-e-s ont ensuite une partie test afin de pouvoir comprendre et s'habituer au déroulement de l'étude, puis commence la plus grande partie du questionnaire, l'expérience même. Le questionnaire se termine avec un dernier message de notre part, remerciant les participant-e-s du temps accordé à notre expérience.

Figure 12: Schéma de notre questionnaire



4.2.1 Plan expérimental

Notre corpus ayant un nombre assez élevé d'images, nous avons décidé de créer cinq questionnaires identiques excepté que les images présentées seraient différentes. En effet, afin d'éviter que les participant-e-s, par fatigue et impatience, ne soient plus assez concentré-e-s à la fin du questionnaire, nous avons préféré diviser de manière semi-aléatoire les images afin que chaque questionnaire n'en comporte que 48. La division s'est faite semi-aléatoirement car il fallait tout de même que chaque questionnaire ait un nombre égal d'images pour chacune des quatre catégories et type de stimulus, IA ou réelle (R). Pour chacune des quatre catégories, nous avons donc 12 images, 6 réelles et 6 générées pas IA.

Afin d'évaluer la perception initiale des images générées par intelligence artificielle par rapport à des images réelles, chaque participant-e-s a été exposé-e à 5 images dans une partie test du questionnaire, puis à 48 images lors de la partie expérimentale. Les 48 images ont été aléatoirement alternées durant la partie expérimentale et présentées chacune pendant uniquement une seconde. Après chaque image, deux questions ont été posées aux participant-e-s. La première leur demande de juger sur une échelle allant de « extremely likely to be AI » à « Extremely likely to be real », leur perception initiale de l'image. La deuxième

question, leur demande d'indiquer le degré de confiance de leur évaluation sur une échelle allant de « Not confident at all » à « Extremely confident ».

Au total, nous avons estimé que le questionnaire prendrait environ une quinzaine de minutes à compléter. La partie test avait pour but de permettre aux participant-e-s de comprendre ce qui est attendu et de se familiariser avec le processus. Nous avons décidé de créer le questionnaire en anglais pour des raisons de diffusion et d'accessibilité à un plus grand nombre de personnes. Cela nous a permis de recevoir un grand nombre de réponses en peu de temps.

4.2.2 Passation du questionnaire et récolte des données

Pour la collecte de nos données, nous avons décidé de passer par une compagnie spécialisée dans la diffusion de sondages aux Etats-Unis, nommément Prolific (*Prolific* 2024). Ceci nous a permis d'accéder à un large échantillon de personnes, déjà enregistrées chez cette compagnie et ayant donc un intérêt à participer à des études. Une fois les questionnaires préparés, nous avons dans un premier temps envoyé qu'une seule des cinq versions. Ceci afin de pouvoir analyser en surface les résultats, de nous assurer que notre questionnaire était compréhensible par les participant-e-s, et d'apporter les modifications nécessaires aux quatre autres versions en conséquence.

Pour des raisons de temps et de budget, nous avons décidé que 30 participant-e-s par questionnaire, soit 150 au total, suffiraient à obtenir des résultats statistiques pertinents. Pour le premier questionnaire nous avons donc arrêté la collecte de données après 33 réponses. Chaque participation était rémunérée à un taux horaire de £9.00. Les participant-e-s ayant répondu à notre questionnaire avaient entre 18 et 50 ans, ce qui fait une moyenne de 32.75 ans et un écart-type *sd* = 6.4. Le questionnaire semblait bien fonctionner et nous avons donc diffusé les quatre autres versions et récolté suffisamment de données en environ une semaine.

5. Résultats

5.1 Préparation des données

Une fois toutes les données récoltées, nous sommes passés sur RStudio afin de préparer les données pour l'analyse. Dans un premier temps, nous avons retiré plusieurs colonnes du dataset, comportant parmi elles les dates auxquelles le questionnaire a été rempli, ainsi que les différentes colonnes correspondant au temps passé sur chaque question. Nous avons également décider de retirer les participants ayant passé plus de 1200 secondes (20 minutes) sur le questionnaire car nous avons estimé que ce temps était trop long. En effet, lorsque nous avons testé le questionnaire par nous-même, nous l'avons terminé en environ 15 minutes. Finalement, nous n'avons gardé que les colonnes correspondant aux résultats des questions test et de l'expérience, le code d'identification des participant-e-s et la durée totale nécessaire à chacun-e pour remplir le questionnaire.

À ce jeu de données nettoyées, nous avons ensuite ajouté les informations concernant les images notées : le type de stimulus (AI ou R), la catégorie (animaux, paysages, tableaux ou visages), le nom du fichier image afin de pouvoir savoir exactement de quelle image il est question, ainsi qu'à quel questionnaire appartiennent les réponses (1 à 5 ainsi que les images de la partie test). Le même traitement a été appliqué à chaque questionnaire, puis les cinq questionnaires ont été fusionnés pour n'avoir qu'un seul jeu de données. Cette fusion nous a permis de commencer l'analyse de données avec un dataset complet et uniforme.

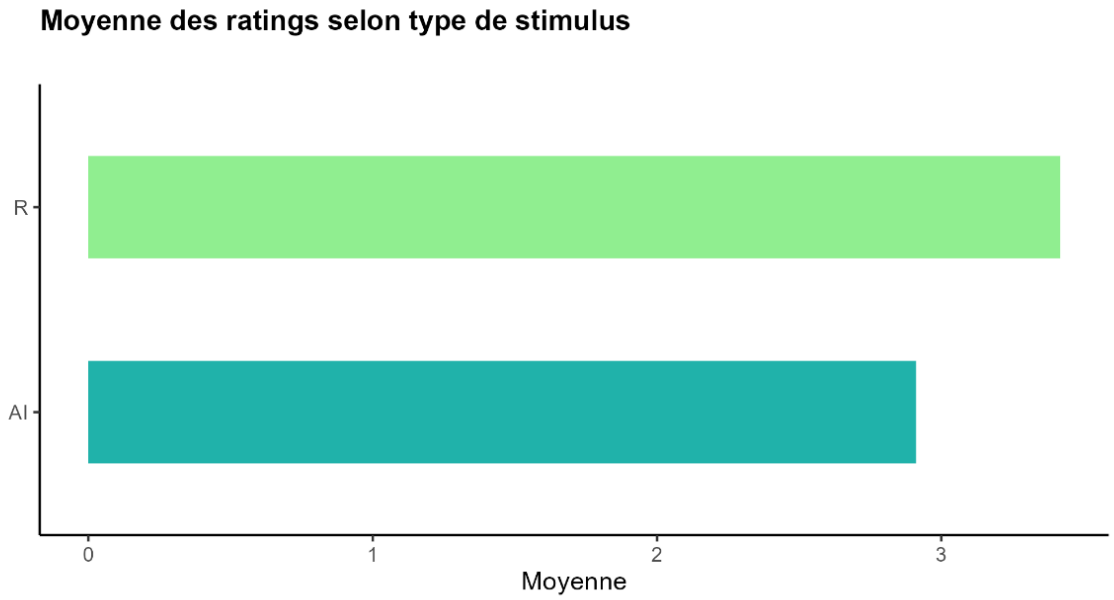
5.2 Résultats par rapport à nos hypothèses

5.2.1 Résultat global des évaluations d'images

Notre première hypothèse était que les participant-e-s seraient capables de distinguer les images générées par IA des images réelles. Pour tester cela, nous avons fait un GLM qui analyse la relation entre le type de stimulus et les résultats des évaluations attribuées aux images par les participant-e-s. Dans celui-ci, le prédicteur est le type de stimulus (AI ou R) et la variable dépendante est l'évaluation des participant-e-s.

En ce qui concerne la moyenne des évaluations attribuées par nos participants aux images IA, notre test révèle que notre moyenne des images réelles (M = 3.39, SD = 1.24) est significativement plus élevée que la moyenne des images AI (M = 2.92, SD = 1.15), t = 17.08, $p < .05$. Nous avons visualisé cela à travers un graphique simple à la Figure 13.

Figure 13: Visualisation de la moyenne globale des évaluations

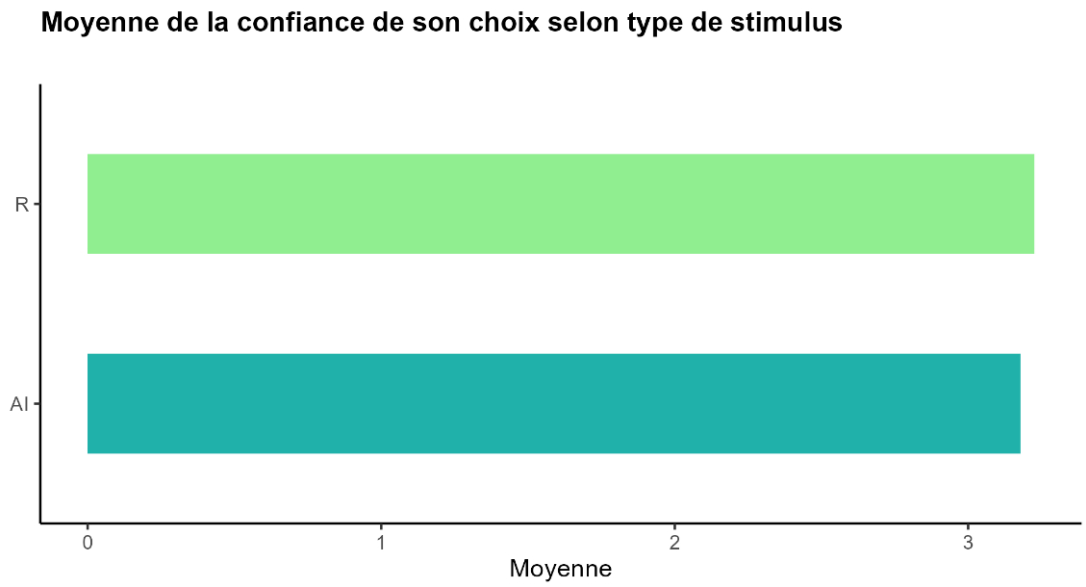


5.2.2 Résultat global de la confiance des participant-e-s en leur choix

Selon notre deuxième hypothèse, nous pensions que les participant-e-s auraient un taux de confiance plus élevé pour les images réelles (R) que pour les images IA. De même que pour notre première hypothèse, nous avons fait un GLM qui analyse la relation entre un prédicteur et une variable dépendante qui sont les mêmes que pour le premier GLM.

En ce qui concerne la confiance que nos participant-e-s avaient en leur réponse, la moyenne de leur confiance en leur jugement des images AI (M = 3.19, SD = 1.10) n'est pas significativement inférieure à la moyenne de leur confiance en leur jugement d'images réelles (M = 3.20, SD = 1.11), t = 0.52, p = .60. Nous l'avons également visualisé par un graphique à la Figure 14.

Figure 14: Visualisation de la moyenne globale de la confiance des participant-e-s en leur choix



5.2.3 Résultats selon les différentes catégories

Notre troisième hypothèse était que certaines catégories seraient mieux distinguées comparées à d'autres. En l'occurrence, que les participant-e-s détermineraient mieux les images réelles (R) des images IA dans les catégories « animaux » et « visages ». Nous avons continué avec nos GLM que nous avons fait par catégorie cette fois et toujours avec les mêmes prédicteur et variable dépendante.

5.2.3.1 Catégorie « Animaux »

Au niveau de l'évaluation des images, notre test révèle que la moyenne de leur jugement des images AI (M = 2.73, SD = 1.28) est significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.64, SD = 1.10), t = 16.27, p < .05. Pour la confiance, nous observons que la moyenne des images AI (M = 3.30, SD = 1.08) n'est au contraire pas significativement inférieure à la moyenne de la confiance des images réelles (M = 3.31, SD = 1.09), t = 0.16, p = .86.

5.2.3.2 Catégorie « Paysages »

Au niveau de l'évaluation des images, notre test révèle que la moyenne de leur jugement des images AI (M = 2.86, SD = 1.14) est significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.30, SD = 1.13), t = 8.58, p < .05. Pour la confiance, nous observons que la moyenne des images AI (M = 3.03, SD = 1.07) n'est au contraire pas significativement inférieure à la moyenne de la confiance des images réelles (M = 3.12, SD = 1.08), t = 1.83, p = .06.

5.2.3.3 Catégorie « Tableaux »

Au niveau de l'évaluation des images, notre test révèle que la moyenne de leur jugement des images AI (M = 2.66, SD = 1.22) est significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.15, SD = 1.13), t = 8.86, p < .05. Pour la confiance, nous observons

que la moyenne des images AI (M = 3.17, SD = 1.16) est significativement supérieure à la moyenne de la confiance des images réelles (M = 3.05, SD = 1.15), t = -2.25, p = .02.

5.2.3.4 Catégorie « Visages »

Au niveau de l'évaluation des images, notre test révèle que la moyenne de leur jugement des images AI (M = 3.40, SD = 1.17) n'est pas significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.48, SD = 1.18), t = 1.46, p = .14. Pour la confiance, nous observons que la moyenne des images AI (M = 3.24, SD = 1.08) est au contraire significativement inférieure à la moyenne de la confiance des images réelles (M = 3.37, SD = 1.10), t = 2.50, p = .01.

5.2.3.5 Moyennes des évaluations d'images et de la confiance par catégorie

Afin que les résultats par catégorie soient lisibles et facilement comparables entre eux, nous les avons compilés dans le Tableau 2 et visualisés aux Figures 15 et 16.

Tableau 2: Moyenne des évaluations et confiance en leur évaluation par catégorie

		Évaluation	*Confiance*
Catégorie	*Type de stimulus*	*M (SD)*	*M (SD)*
Animaux	AI	2.73 (1.28)	3.30 (1.08)
	R	3.64 (1.10)	3.31 (1.09)
Paysages	AI	2.86 (1.14)	3.03 (1.07)
	R	3.30 (1.13)	3.12 (1.08)
Tableaux	AI	2.66 (1.22)	3.17 (1.16)
	R	3.15 (1.13)	3.05 (1.15)
Visages	AI	3.40 (1.17)	3.24 (1.08)
	R	3.48 (1.18)	3.37 (1.10)

Figure 15: Visualisation de la moyenne des évaluations par catégorie

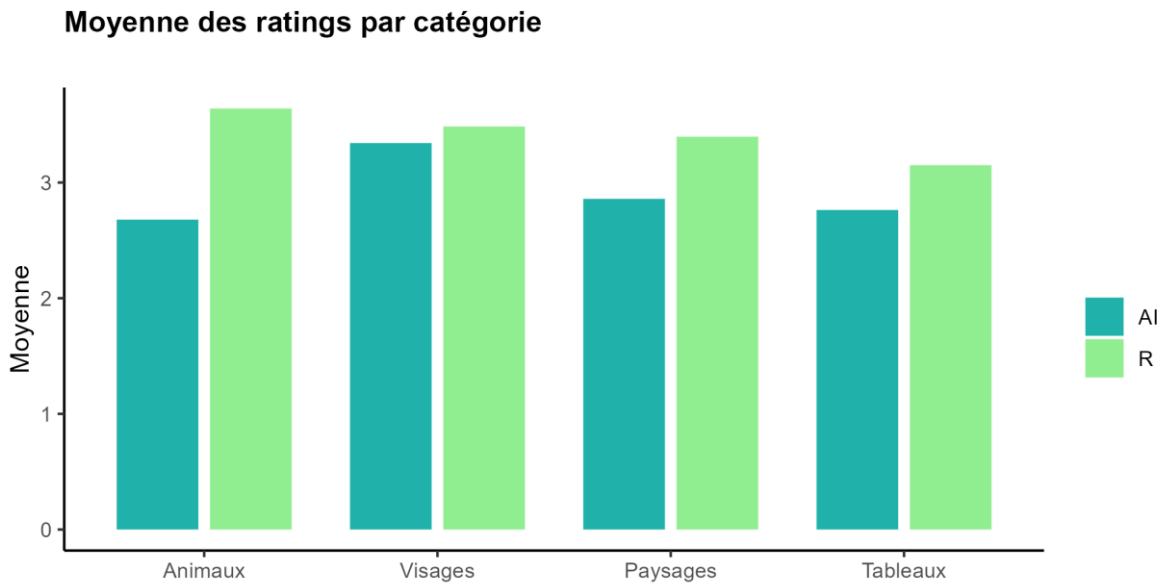
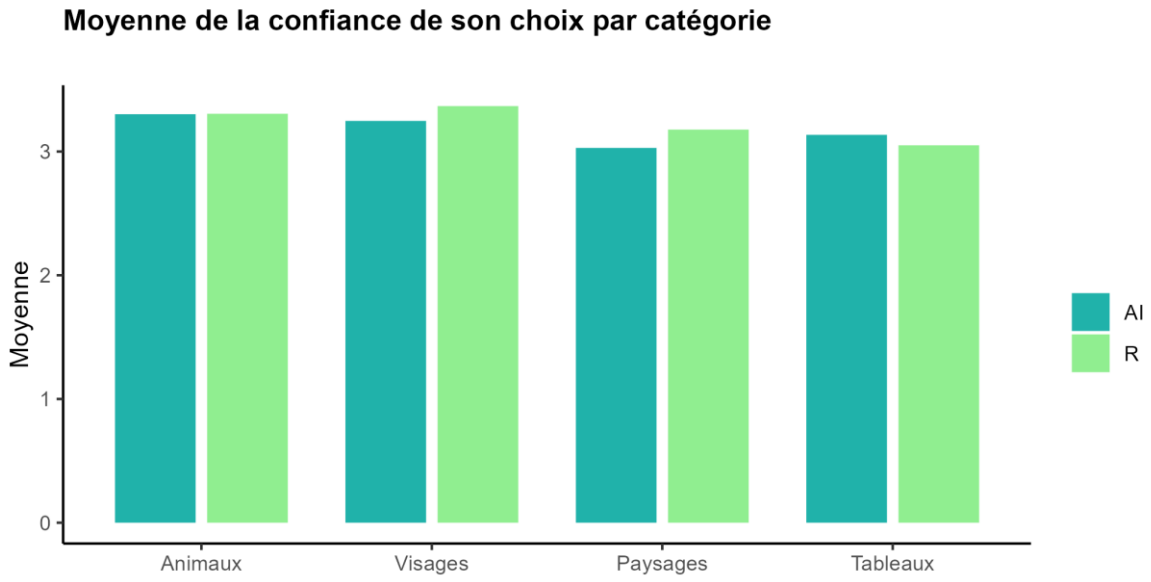


Figure 16: Visualisation de la moyenne de la confiance par catégorie



5.2.4 Résultats selon le genre des participant-e-s

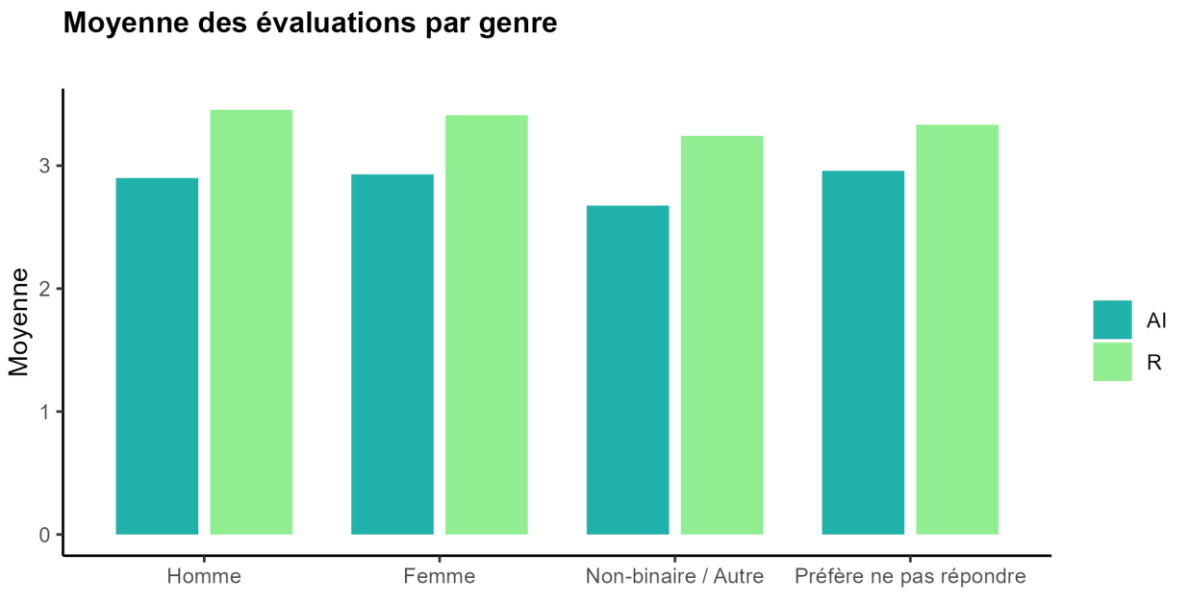
Nous avons également eu la curiosité de voir si des différences entre différents genres et différentes tranches d'âges étaient significativement visibles, comme il l'a été noté dans l'une des études dont nous avons discuté dans la revue de la littérature. Pareillement aux résultats par catégorie, nous avons combiné les résultats par genre dans le Tableau 3 pour une meilleure lisibilité et facilité de comparaison.

Tableau 3: Moyenne des évaluations et confiance en leur évaluation par genre

		Évaluation	*Confiance*
Genre	*Type de stimulus*	*M (SD)*	*M (SD)*
Hommes	AI	2.91 (1.29)	3.26 (1.19)
	R	3.43 (1.19)	3.25 (1.17)
Femmes	AI	2.92 (1.22)	3.15 (1.07)
	R	3.37 (1.13)	3.18 (1.09)
Non-binaire / Autre	AI	2.70 (1.07)	2.95 (0.83)
	R	3.25 (0.98)	2.83 (0.82)
A préféré ne pas répondre	AI	2.95 (1.28)	3.54 (0.83)
	R	3.33 (1.17)	3.59 (0.96)

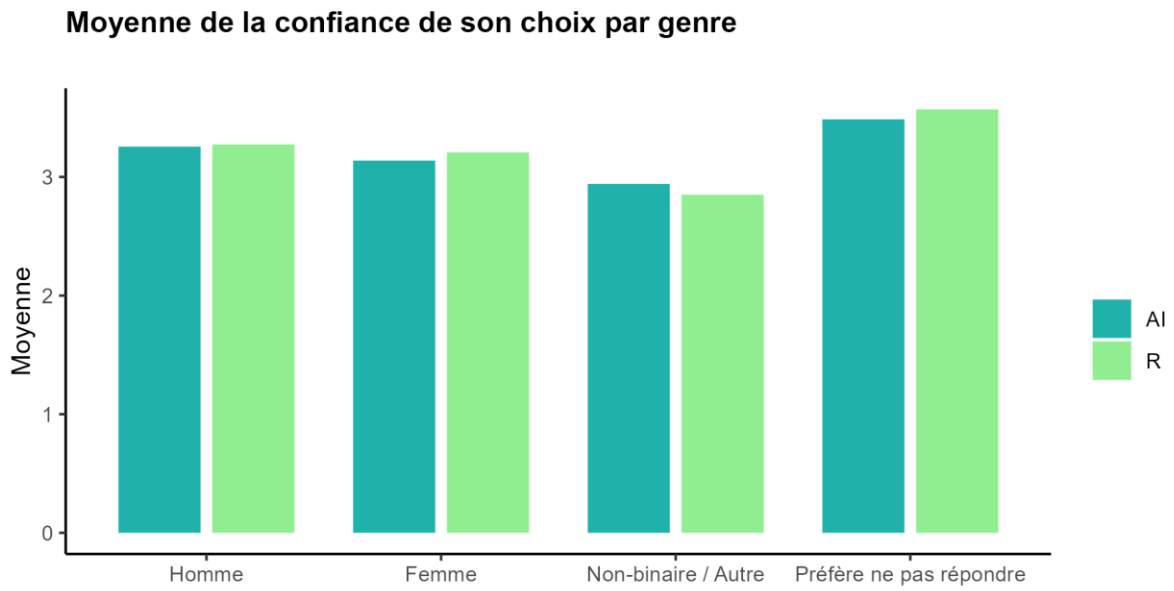
Au niveau des évaluations données, notre test montre que les participants masculins (n = 48) ont une moyenne de leur jugement des images AI (M = 2.91, SD = 1.29) qui est significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.43, SD = 1.19), t = 10.41, p < .05. Les participantes féminines (n = 86) ont une moyenne de leur jugement des images AI (M = 2.92, SD = 1.22) qui est également significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.37, SD = 1.13), t = 12.78, p < .05. Les participant-e-s non-binaires (n = 5) ont une moyenne de leur jugement des images AI (M = 2.70, SD = 1.07) qui est elle aussi significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.25, SD = 0.98), t = 4.36, p < .05. Les personnes qui n'ont pas voulu préciser leur genre (n = 3) ont une moyenne de leur jugement des images AI (M = 2.95, SD = 1.28) qui n'est tout juste pas significativement inférieure à la moyenne de leur jugement d'images réelles (M = 3.33, SD = 1.17), t = 1.96, p = .051. Nous avons visualisé ces résultats à la Figure 17.

Figure 17: Visualisation de la moyenne des évaluations par genre



Au niveau de la confiance en leur jugement, notre test montre que les hommes ont une moyenne de leur confiance en leur jugement des images AI (*M* = 3.26, *SD* = 1.19) qui n'est pas significativement supérieur à la moyenne de leur confiance en leur jugement d'images réelles (*M* = 3.25, *SD* = 1.17), *t* = -0.21, *p* = .83. Les femmes ont une moyenne de leur confiance en leur jugement des images AI (*M* = 3.15, *SD* = 1.07) qui n'est pas significativement inférieure à la moyenne de leur confiance en leur jugement d'images réelles (*M* = 3.18, *SD* = 1.09), *t* = 1.02, *p* = .30. Les participants non-binaires ont une moyenne de leur confiance en leur jugement des images AI (*M* = 2.95, *SD* = 0.83) qui n'est pas significativement supérieure à la moyenne de leur confiance en leur jugement d'images réelles (*M* = 2.83, *SD* = 0.82), *t* = -1.15, *p* = .25. Finalement, les personnes qui n'ont pas voulu préciser leur genre ont une moyenne de leur confiance en leur jugement des images AI (*M* = 3.54, *SD* = 0.83) qui n'est pas significativement inférieure à la moyenne de leur confiance en leur jugement d'images réelles (*M* = 3.59, *SD* = 0.96), *t* = 0.32, *p* = .74. Nous avons visualisé ces résultats à la Figure 18.

Figure 18: Visualisation de la moyenne de la confiance par genre



5.2.5 Résultats selon l'âge des participant-e-s

De même que pour le genre, nous avons voulu voir s'il existait une différence significative entre différents groupes d'âges. Nous avons donc divisé les participants par décennie, ajoutant le participant de 18 ans à la tranche 18-30 ans. Les résultats des évaluations des participant-e-s ainsi que de la confiance en leur évaluation sont visibles au Tableau 4.

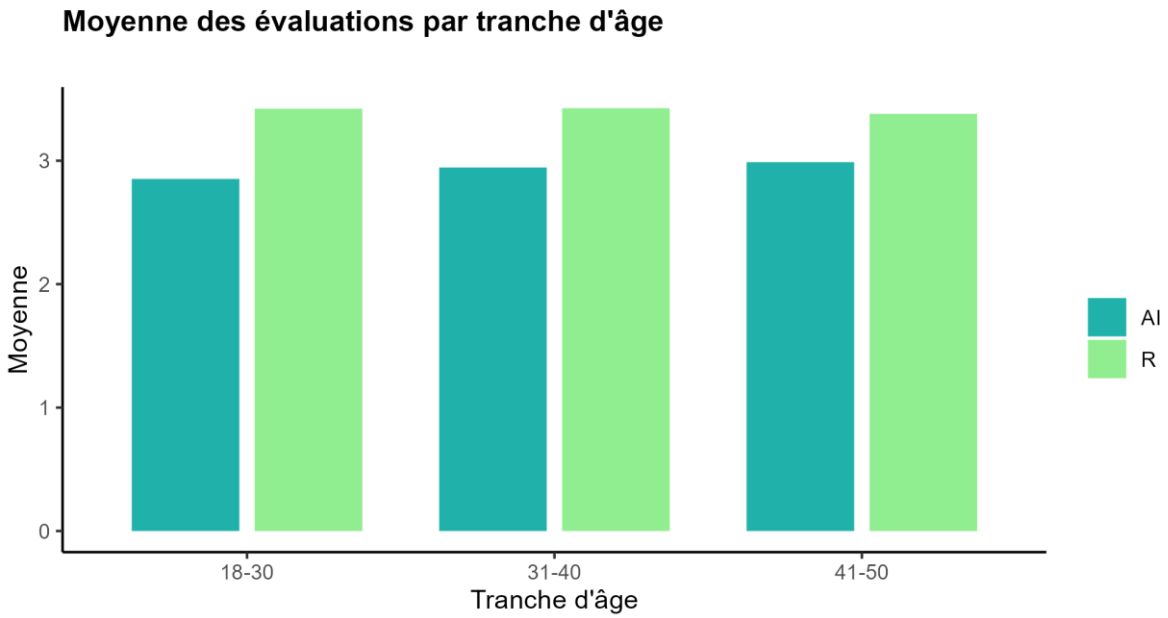
Tableau 4: Moyenne des évaluations et confiance en leur évaluation par tranche d'âge

		Évaluation	*Confiance*
Tranche d'âge	*Type de stimulus*	*M (SD)*	*M (SD)*
18 - 30	AI	2.86 (1.21)	3.18 (1.06)
	R	3.40 (1.11)	3.20 (1.05)
31 - 40	AI	2.94 (1.27)	3.21 (1.13)
	R	3.38 (1.18)	3.22 (1.15)
41 - 50	AI	3.02 (1.25)	3.14 (1.16)
	R	3.36 (1.18)	3.12 (1.16)

Au niveau des évaluations données, notre test montre que les 18-30 ans (n = 61) ont une moyenne de leur jugement des images AI (M = 2.86, SD = 1.21) qui est significativement

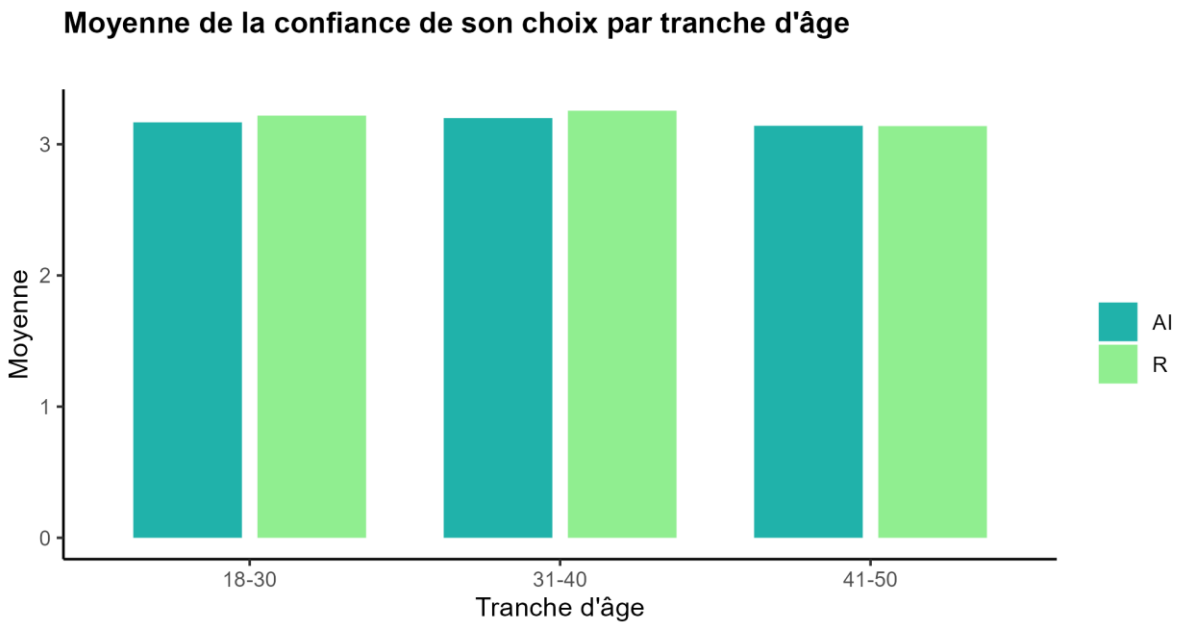
inférieure à la moyenne de leur jugement d'images réelles (*M* = 3.40, *SD* = 1.11), *t* = 13.27, *p* < .05. Les 31-40 ans (*n* = 63) ont une moyenne de leur jugement des images AI (*M* = 2.94, *SD* = 1.27) qui est également significativement inférieure à la moyenne de leur jugement d'images réelles (*M* = 3.38, *SD* = 1.18), *t* = 10.36, *p* < .05. Les 41-50 ans (*n* = 18) ont une moyenne de leur jugement des images AI (*M* = 3.02, *SD* = 1.25) qui est elle aussi significativement inférieure à la moyenne de leur jugement d'images réelles (*M* = 3.36, *SD* = 1.18), *t* = 4.37, *p* < .05. Nous avons visualisé ces résultats à la figure 19. Nous avons visualisé ces résultats à la Figure 19.

Figure 19: Visualisation de la moyenne des évaluations par tranche d'âge



Au niveau de la confiance en leur jugement, notre test montre que les 18-30 ans ont une moyenne de leur confiance en leur jugement des images AI (*M* = 3.18, *SD* = 1.06) qui n'est pas significativement inférieure à la moyenne de leur confiance en leur jugement d'images réelles (*M* = 3.20, *SD* = 1.05), *t* = 0.69, *p* = .48. Les 31-40 ans ont une moyenne de leur confiance en leur jugement des images AI (*M* = 3.21, *SD* = 1.13) qui n'est pas significativement inférieure à la moyenne de leur confiance en leur jugement d'images réelles (*M* = 3.22, *SD* = 1.15), *t* = 0.27, *p* = .78. Les 41-50 ans ont une moyenne de leur confiance en leur jugement des images AI (*M* = 3.14, *SD* = 1.16) qui n'est pas significativement supérieure à la moyenne de leur confiance en leur jugement d'images réelles (*M* = 3.12, *SD* = 1.16), *t* = -0.26, *p* = .79. Nous avons visualisé ces résultats à la Figure 20.

Figure 20: Visualisation de la moyenne de la confiance par tranche d’âge



6. Discussion

6.1 Interprétation de nos résultats et comparaison avec d'autres études

6.1.1 Résultats globaux

Si nous rappelons notre première hypothèse, nous pensions que les participant-e-s à notre étude seraient capables de différencier des images générées par IA d'images réelles. En regardant les résultats de notre questionnaire, nous pouvons voir que de manière générale, notre hypothèse se confirme car nos participant-e-s, de manière significative, n'ont pas été trompé-e-s par les images générées par IA et ont déterminé comme réel des images en effet réelles.

Nous n'avons pas demandé aux participant-e-s quelles caractéristiques leur permettaient de faire leur choix, mais nous imaginons que d'éventuelles imperfections ou manque de réalisme pourraient avoir permis à la distinction des images IA. La tendance de certains générateurs que nous avons utilisé à générer des images donnant l'impression d'être des peintures, et que nous avons mentionné dans la section 4 de ce travail, pourraient également en être la raison.

Ce résultat est cependant contraire aux conclusions auxquelles les différentes études que nous avons mentionné dans la section 2.2 sont arrivées. En effet, les résultats d'autres recherches avaient conclu sur le fait que de plus en plus l'être humain n'arrive pas à différencier des images IA d'images réelles avec une tendance à croire que des images générées par intelligence artificielle sont même plus réalistes que des images réelles.

Deux points sont cependant à prendre en considération. Premièrement, nous avons également testé la confiance que nos participant-e-s avaient en leur réponse et, de manière globale, leur niveau de confiance ne différait pas significativement entre des images IA ou réelles. Cela veut dire que bien que nos participant-e-s sont capables de distinguer les deux types d'image, iels ne sont pourtant pas sûr-e-s de leur jugement. Cela pourrait s'expliquer par le fait que les participant-e-s sont conscients de l'évolution technologique relative à la génération d'images et savent que l'IA est capable de créer des images stupéfiantes de réalisme. Nos participant-e-s pourraient donc soit avoir des doutes quant à leur capacité à reconnaître ce qui pourrait être généré par IA, soit adopter une approche plus prudente et moins catégorique dans leur jugement.

Deuxièmement, les résultats dont nous parlons correspondent aux résultats globaux de notre étude, c'est-à-dire sans différenciation aucune entre les différentes catégories d'images que nous avons (Animaux, Paysages, Tableaux et Visages). Toutefois, les articles de la section 2.2 de notre revue de la littérature se sont uniquement concentrés sur des visages et il n'est donc pas adéquat de comparer nos résultats globaux avec les leurs.

6.1.2 Résultats par catégorie

Une autre de nos hypothèses, était que les participant-e-s seraient capable de mieux faire la différence entre AI et réel selon différentes catégories. Selon nous, nous pensions que les participant-e-s auraient de meilleurs résultats dans les catégories « Animaux » et « Visages » car nous pensons que les êtres humains sont souvent en contact entre eux mais également avec leurs animaux de compagnie. Nous pensions que le contact et l'exposition fréquente à

ces êtres pourraient permettre de reconnaitre immédiatement si une image était générée par intelligence artificielle ou réelle.

Si nous analysons nos résultats par catégorie, nous pouvons voir que, comme pour les résultats globaux, les participant-e-s ont été capables de différencier les images générées par IA des images réelles dans les catégories « Animaux », « Paysages », et « Tableaux ». Nous pouvons imaginer, comme pour les résultats précédents, que des indices visuels ont aidé les participant-e-s à reconnaitre les images intelligence artificielle. Cependant, pour les catégories « Animaux » et « Paysages », la confiance manque de nouveau de significativité et montre une incertitude quant à leur choix. La catégorie « Tableaux » est la seule catégorie présentant des résultats significatifs à la fois dans l'évaluation et dans la confiance en dite évaluation. Cela montre que les participant-e-s étaient capables de différencier des œuvres réalisées par AI de véritables œuvres existantes et étaient sûr-e-s de leur choix. Il est pourtant intéressant d'observer que parmi les cinq images de la catégorie « Tableaux » considérées comme les plus réelles, deux images générées par intelligence artificielle ont pourtant réussi à se hisser parmi elles.

Figure 21: Top 5 des images de la catégorie "Tableaux" jugées comme les plus réelles

En effet, nous pouvons voir (Figure 21) que les deux images générées et représentant de l'art japonais, sont classées au 2ème et 5ème rang. Nous pensons que cela pourrait être dû au fait que nos participant-e-s viennent des Etats-Unis et ne sont peut-être pas souvent exposés à des tableaux d'art japonais. Ceci contrairement aux autres types de tableaux qui représentent de l'art occidental souvent étudié dans les cursus et possédant plus de visibilité dans les pays occidentaux. Le fait que l'IA ait ajouté de l'écriture au tableau pourrait aussi être l'une des raisons pour lesquelles les participant-e-s ont déterminé ces images comme réelles. Ne parlant pas japonais, nous ne savons pas si ce qui est écrit fait un quelconque sens, mais cela a peut-être influencé nos participants, qui ont sûrement pensé que l'IA ne pourrait pas générer des caractères si réalistes.

Au niveau des visages, nos résultats montrent que bien qu'il semblerait que les participant-e-s soient capables de reconnaître des images réelles des images générées par intelligence artificielle, la différence n'est pas significative. Nos résultats, bien que ne montrant pas exactement les mêmes conclusions que dans les articles de notre revue de la littérature, indiquent tout de même que les images IA sont suffisamment réalistes pour qu'il n'y ait pas une distinction claire. Nous pensons que le développement technologique des deepfakes est l'une des raisons pour lesquelles les visages humains semblent être la catégorie d'image la plus réaliste que peut générer l'intelligence artificielle. En effet, pour créer des deepfakes, des milliers d'images de visage d'une personne doivent être nourrit à l'IA. Cette grande quantité d'images, comparé à d'autres catégories, est sûrement la raison pour laquelle l'IA semble être plus douée pour générer des visages humains.

Cette possibilité renforce l'inquiétude que nous pourrions avoir par rapport aux deepfakes car même si nous pouvons dans certains cas encore différencier des images IA d'images réelles, ce manque de distinction significative nous rend vulnérable aux deepfakes. C'est pour cette raison que nous pensons qu'il est important d'éduquer la population quant aux risques mais également, et principalement, leur enseigner à toujours avoir un esprit critique et à s'assurer de la validité des sources desquelles les informations sont tirées.

Figure 22: Top 5 des images de la catégorie "Visages" jugées comme les plus réelles

En observant la Figure 22, nous pouvons voir que deux images générées par IA ont pu se hisser parmi les images considérées comme les plus réelles. Contrairement aux résultats de Nightingale et Farid 2022, dans lesquels les visages générés par intelligence artificielle d'hommes blancs étaient considérés comme les plus réelles, ce sont ici deux visages de personnes asiatiques qui ont réussi à tromper nos participant-e-s.

Nous avons d'ailleurs eu la curiosité de voir parmi les images générées par intelligence artificielle, lesquelles étaient le plus considérées comme réelles, toute catégorie confondue, et les cinq premières sont toutes des images de la catégorie « Visage » que nous pouvons voir à la Figure 23. Nous pouvons donc conclure qu'effectivement, les visages semblent être le type d'image que les générateurs aient le plus été entrainés à générer et avec le plus d'images d'entrainement. Cela veut donc dire que la qualité et le réalisme de la génération les rend de plus en plus indistinguables d'images de personnes réelles, ce qui rejoint les conclusions des études dont nous avons discuté.

Figure 23: Top 5 des images IA considérées comme les plus réelles

6.1.3 Résultats par genre et tranche d'âge

En observant nos résultats, nous pouvons voir qu'il n'existe pas de grande différence entre hommes et femmes car les deux genres ont jugé comme plus réelles des images effectivement réelles. En revanche, selon nos résultats, les personnes non-binaires et celles qui ont préféré ne pas répondre semblent avoir été meilleures à reconnaitre des images IA des images réelles. Il faut cependant prendre cela avec des pincettes car les échantillons de ces deux catégories

ne sont clairement pas suffisants pour arriver à une conclusion. De plus, comme précédemment, les taux de confiance ne sont pas significatifs et donc même si les participant-e-s sont capables de reconnaître des images générées par IA, iels ne sont pas toujours sûr-e-s de leurs choix.

Au niveau de la différence entre tranches d'âges, il n'y en a aucune car les trois tranches d'âges que nous avons déterminées ont toutes réussi à reconnaître les images réelles et ce de manière significative. La confiance est cependant de nouveau non-significative pour tous les âges. Cela nous montre donc que quel que soit l'angle que nous prenons pour étudier ces résultats, que ce soit globalement, par catégorie, par genre ou par âge, bien que les participant-e-s puissent reconnaitre les images générées par intelligence artificielle, leur certitude n'est jamais significative. Cela réfute notre hypothèse selon laquelle les participant-e-s auraient un niveau de confiance plus élevé pour des images réelles. En regardant les chiffres, notre hypothèse est vraie, mais ceux-ci n'étant pas significativement plus élevés que pour le niveau de confiance des images IA, nous ne pouvons donc pas la corroborer.

Il semble de plus en plus plausible que nous ne pourrons bientôt plus faire la différence à l'œil nu entre des images générées par intelligence artificielle et des images réelles. Cette théorie semble être probable au vu de nos résultats, des résultats de précédentes études mais également tout simplement du fait que la technologie s'améliore rapidement et encore plus avec toutes ces personnes qui aujourd'hui utilisent et nourrissent les IA, les aidant à progresser.

6.2 Limitations de notre étude

Désormais à la fin de notre recherche, nous réalisons quelques limitations et quelques améliorations que nous pourrions apporter à notre étude. Premièrement, la taille de l'échantillon de certaines catégories pourrait limiter l'arrivée à des conclusions, par exemple dans le cas des personnes non-binaires et celles qui ont préféré ne pas répondre. Nous aurions pu ne pas proposer le choix de ne pas répondre car les personnes ayant sélectionné ce choix pourraient soit ne pas avoir eu de choix qui les représentait parmi « Homme », « Femme » ou « Non-binaire/Autre » soit faire effectivement parti de l'une de ces trois possibilités mais refuser de nous le dire. Cette incertitude a rendu l'analyse de nos résultats plus difficiles car nous n'avons pas pu arriver à une conclusion les concernant. Dans le futur, nous pensons donc ne pas offrir ce choix, dès lors que l'analyse par genre soit pertinente à la recherche. En ce qui concerne les personnes non-binaires, et de même pour les hommes et femmes, nous aurions pu nous assurer d'avoir un nombre égal pour chaque genre, afin d'avoir plus d'assurance et de confiance en nos résultats comparant les différences entre ces groupes.

Nous aurions pu également faire de même pour les groupes par âge. Bien que nous n'ayons observé aucune différence entre les trois groupes, les 18-30 ans et les 31-40 ans ont un nombre de participant-e-s trois fois plus élevé que le groupe des 41-50 ans. Peut-être aurions-nous noté une différence entre ces groupes si les nombres étaient plus équitables. De plus, il aurait également été intéressant de recruter des participant-e-s plus âgé-e-s afin de comparer entre des générations bien différentes dans leur exposition et usage de la technologie. En effet, les plus de 60 ans pourraient ne pas être autant au courant, que des personnes plus jeunes, des avancées et capacités technologiques de l'intelligence artificielle et pourraient donc être moins capables de reconnaitre des images générées par intelligence artificielle.

Un autre point que nous pourrions améliorer dans de futures recherches, est l'utilisation de meilleurs sites de génération d'image. Si le budget le permet, payer des abonnements premiums serait bénéfique pour avoir accès au meilleur que l'IA peut générer et ainsi avoir des images de meilleure qualité et réalisme. Nous pourrions également simplement passer plus de temps à étudier les générateurs existants, en essayer plus et peut-être trouver des générateurs qui se spécialisent dans certains types d'images comme nous avons trouvé pour les images de visages avec Face Studio. Peut-être existe-t-il des générateurs qui ne créent que des images d'animaux ou de paysages.

Nos résultats démontrant que les niveaux de confiances ne sont pas significatifs, il aurait été intéressant de développer le questionnaire en y ajoutant des questions ouvertes demandant aux participant-e-s les raisons de leur incertitude. En l'état, nous ne pouvons que spéculer sur d'éventuelles raisons : certains participant-e-s, en sachant que l'IA est capable de créer des images très réalistes, n'ont peut-être pas voulu être catégorique. En demandant aux participant-e-s d'expliquer leur choix, nous pourrions présenter nos analyses et conclusions avec plus d'assurance.

6.3 Création théorique d'un tableau de bord

Avant de conclure ce travail, nous voudrions tout de même consacrer quelques instants à l'imagination d'un tableau de bord interactif qui nous permettrait de mettre en avant les résultats de notre recherche et de les rendre plus facilement accessibles au grand public.

Pour cela, nous utiliserions Quarto pour créer ce tableau de bord. Au moins deux onglets seraient disponibles. Le premier, nommé « Images » et celui qui apparaitrait automatiquement en tant que page d'accueil, pourrait avoir un carrousel des différentes images que nous avons utilisé dans notre questionnaire et que nous avons demandé aux participant-e-s de juger. Idéalement, ces images seraient présentées sous forme de GIF, apparaissant les unes après les autres, 1 seconde chacune, afin de représenter au mieux les conditions de notre questionnaire. L'entête de la section avec le carrousel pourrait être sous forme de question accrocheuse, intitulée par exemple : « Saurez-vous reconnaitre les images générées par IA ? ». Sous cette section du carrousel, nous pourrions avoir une deuxième section qui serait rétractable. Elle serait fermée par défaut, mais en cliquant sur l'entête, elle s'ouvrirait pour révéler des versions fixes des images du carrousel et présentant la réponse de quelles images ont été générées par IA.

Dans le deuxième onglet, nommé « Visualisations de nos résultats », nous aurions, comme son nom l'indique, certains des graphiques que nous avons présenté dans ce travail. Nous pourrions avoir une section dans laquelle nous présenterions les visualisations des Figures 13 et 14. Nous pourrions avoir une sidebar dans laquelle les utilisateurs et utilisatrices pourraient sélectionner laquelle des visualisations afficher. Par défaut, la visualisation de l'évaluation par type de stimulus (Figure 13) serait visible et il faudrait cocher une case dans la sidebar pour afficher la visualisation de la confiance par type de stimulus. Similairement, nous aurions une deuxième section, à côté de la première, dans laquelle nous aurions les visualisations par catégorie (Figure 15 et 16). Également dans la sidebar, les utilisateurs et utilisatrices pourraient sélectionner quelles catégories ou stimuli afficher.

Finalement, dans ce deuxième onglet, nous aurions une dernière section, sous les deux premières, contenant un court texte qui expliquerait nos résultats globaux et par catégorie, insistant sur les résultats de la catégorie « Visage » car ce sont ceux qui nous paraissent les

plus intéressants. Les résultats par genre et tranche d'âge pourraient également être mentionnés.

7. Conclusion

À travers ce travail, nous avons analysé la capacité de participant-e-s à reconnaitre des images générées par intelligence artificielle d'images réelles à travers différentes catégories. De manière globale, les personnes étaient capables de faire la distinction, ce qui confirmait notre hypothèse. Cependant, le manque de confiance des participant-e-s en leur réponse ne nous encourage pas à maintenir cette hypothèse bien longtemps car cela montre que cette capacité de distinction à l'œil nu est fragile et risque de disparaitre avec la technologie qui continue à progresser rapidement. De plus, rappelons que nos résultats n'ont pas confirmé notre hypothèse dans la catégorie « Visages ». En effet, bien que la moyenne soit plus élevée pour les images réelles, la différence avec les images générées par IA n'est pas significative.

Ce résultat s'inscrit donc parmi les résultats des différentes études que nous avons présenté dans notre revue de la littérature. L'intelligence artificielle générative est devenue si performante, qu'il est désormais difficile pour un être humain de faire la différence entre le visage d'une personne qui existe et celui d'une personne générée par IA. Plus impressionnant encore, les GANs sont même capables de générer des visages considérés comme plus réalistes que de véritables visages humains.

Cet hyperréalisme de l'IA crée quelques inquiétudes, notamment dans le contexte des deepfakes qui sont donc de plus en plus réalistes et inquiètent certaines personnes à cause des risques de désinformation et l'impact sur la confiance que nous pourrions avoir dans les médias. Récemment encore, un software, appelé Deep-Live-Cam, a enflammé GitHub avec la possibilité de créer un deepfake en direct à partir d'une seule et unique image. Bien que les créateurs aient mis en place des mesures de sécurité pour éviter les abus, ils se sont également engagés à continuer de développer et modifier le programme dans le cadre le plus éthique et légal possible, en ajoutant par exemple des watermarks si cela s'avère nécessaire (Brain Titan 2024).

Cependant, comme nous l'avons déjà mentionné, en éduquant la population quant aux risques et en leur enseignant soit comment reconnaitre des deepfakes, soit comment s'assurer des sources desquelles leurs informations sont tirées, nous pouvons mitiger les risques que posent les deepfakes. De plus, si la technologie pour créer des deepfakes s'améliore, c'est sûrement également le cas aussi de la technologie qui permet de repérer des deepfakes.

Au final, notre étude, malgré ses quelques limitations, s'inscrit dans ce qui a été et continue de se faire dans le domaine de la reconnaissance et perception d'images générées par IA. Plusieurs études se sont intéressées aux visages spécifiquement mais, en allant plus loin que notre recherche, de futures études pourraient se concentrer principalement sur différentes catégories d'images telles que des animaux, des paysages, de l'architecture ou des objets. Ou alors, au lieu de se concentrer sur la perception humaine, il pourrait être intéressant de voir si l'IA elle-même est capable de reconnaitre des images générées.

En conclusion, bien qu'il existe des risques quant à la propagation de contenu généré par IA, nous pouvons rester optimistes car des mesures et solutions existent pour les atténuer. De plus, bien que ce soit son utilisation malhonnête qui fasse le plus parler d'elle, les GANs et l'IA de manière générale restent des outils au grand potentiel qu'il nous suffit d'utiliser de manière éthique et responsable.

Bibliographie

ADOBE FIREFLY, 2023. IA générative : définition et principe de fonctionnement. *Adobe Firefly* [en ligne]. 2023. Disponible à l'adresse : https://www.adobe.com/ch_fr/products/firefly/discover/how-generative-ai-work.html [consulté le 14 juin 2024].

BALAS, Benjamin et PACELLA, Jonathan, 2015. Artificial faces are harder to remember. *Computers in Human Behavior*. Vol. 52, pp. 331-337. DOI 10.1016/j.chb.2015.06.018.

BARBER, Alex, 2023. Freedom of expression meets deepfakes. *Synthese*. Vol. 202, no 40, pp. 1-17. DOI 10.1007/s11229-023-04266-4.

BRAIN TITAN, 2024. Deep-Live-Cam: Live face-swapping and one-click video deepfake tool that replaces faces with just a…. *Medium* [en ligne]. 10 août 2024. Disponible à l'adresse : https://braintitan.medium.com/deep-live-cam-live-face-swapping-and-one-click-video-deepfake-tool-that-replaces-faces-with-just-a-d515613a94cc [consulté le 13 août 2024].

CHANDALIYA, Praveen Kumar et NAIN, Neeta, 2022. ChildGAN: Face aging and rejuvenation to find missing children. *Pattern Recognition*. Vol. 129, no 108761, pp. 1-15. DOI 10.1016/j.patcog.2022.108761.

DAHMANI, Sara et al., 2019. Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis. In : *Interspeech 2019 - 20th Annual Conference of the International Speech Communication Association*, pp. 2598-2602. Graz, Austria. ISCA. 15 septembre 2019. DOI 10.21437/Interspeech.2019-2848.

FACE STUDIO, 2024. *Face Studio* [en ligne]. 2024. Disponible à l'adresse : https://facestudio.app/ [consulté le 1 août 2024].

FALLIS, Don, 2021. The Epistemic Threat of Deepfakes. *Philosophy & Technology*. Vol. 34, no 4, pp. 623-643. DOI 10.1007/s13347-020-00419-2.

FREEPIK, 2024. Freepik AI image generator. *Freepik* [en ligne]. 2024. Disponible à l'adresse : https://www.freepik.com/pikaso/ai-image-generator?style=noStyle&submit=1 [consulté le 1 août 2024].

GUY, Jack, 2023. Outcry in Spain as artificial intelligence used to create fake naked images of underage girls. *CNN* [en ligne]. 20 septembre 2023. Disponible à l'adresse : https://www.cnn.com/2023/09/20/europe/spain-deepfake-images-investigation-scli-intl/index.html [consulté le 5 juillet 2024].

HARRIS, Keith Raymond, 2021. Video on demand: what deepfakes do and how they harm. *Synthese*. Vol. 199, no 5-6, pp. 13373-13391. DOI 10.1007/s11229-021-03379-y.

IQBAL, Mansoor, 2024. TikTok Revenue and Usage Statistics (2024). *Business of Apps* [en ligne]. 2024. Disponible à l'adresse : https://www.businessofapps.com/data/tik-tok-statistics/ [consulté le 27 juin 2024].

KARNOUSKOS, Stamatis, 2020. Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Transactions on Technology and Society*. Vol. 1, no 3, pp. 138-147. DOI 10.1109/TTS.2020.3001312.

KARRAS, Tero, LAINE, Samuli et AILA, Timo, 2019. *A Style-Based Generator Architecture for Generative Adversarial Networks* [en ligne]. 29 mars 2019. Disponible à l'adresse : http://arxiv.org/abs/1812.04948 [consulté le 20 juin 2024]. arXiv:1812.04948 [cs, stat]

KIETZMANN, Jan et al., 2020. Deepfakes: Trick or treat? *Business Horizons*. Vol. 63, no 2, pp. 135-146. DOI 10.1016/j.bushor.2019.11.006.

LEINGANG, Rachel, 2024. Videos of Biden looking lost are a viral political tactic: 'low-level manipulation'. *The Guardian* [en ligne]. 19 juin 2024. Disponible à l'adresse : https://www.theguardian.com/us-news/article/2024/jun/19/joe-biden-edited-videos [consulté le 6 juillet 2024].

LI, Yuezun, CHANG, Ming-Ching et LYU, Siwei, 2018. *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking* [en ligne]. 11 juin 2018. Disponible à l'adresse : http://arxiv.org/abs/1806.02877 [consulté le 24 avril 2024]. arXiv:1806.02877 [cs]

MILLER, Elizabeth J. et al., 2023. AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*. Vol. 34, no 12, pp. 1390-1403. DOI 10.1177/09567976231207095.

MINISTÈRE DE L'ÉCONOMIE, DES FINANCES ET DE LA SOUVERAINETÉ INDUSTRIELLE ET NUMÉRIQUE, 2024. Quels sont les outils permettant de décrypter l'information ? *Ministère de l'économie, des finances et de la souveraineté* [en ligne]. 2024. Disponible à l'adresse : https://www.economie.gouv.fr/cedef/outils-decrypter-information [consulté le 6 juillet 2024].

NIGHTINGALE, Sophie J. et FARID, Hany, 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*. Vol. 119, no 8, p. e2120481119. DOI 10.1073/pnas.2120481119.

ORBISK, 2023. Orbisk - Automatically reduce food waste in your kitchen. *Orbisk* [en ligne]. 2023. Disponible à l'adresse : https://orbisk.com/ [consulté le 13 juin 2024].

Prolific, 2024. *Prolific* [en ligne]. Disponible à l'adresse : https://www.prolific.com [consulté le 2 août 2024].

Runway, 2024. Text to Image. *Runway* [en ligne] 2024. Disponible à l'adresse : https://app.runwayml.com [consulté le 1 août 2024].

TUCCIARELLI, Raffaele et al., 2022. On the realness of people who do not exist: The social processing of artificial faces. *iScience*. Vol. 25, no 12, p. 105441. DOI 10.1016/j.isci.2022.105441.

VALENTINE, Tim, LEWIS, Michael B. et HILLS, Peter J., 2016. Face-Space: A Unifying Concept in Face Recognition Research. *Quarterly Journal of Experimental Psychology*. Vol. 69, no 10, pp. 1996-2019. DOI 10.1080/17470218.2014.990392.

VODRAHALLI, Kailas et al., 2022. Do Humans Trust Advice More if it Comes from AI?: An Analysis of Human-AI Interactions. In : *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 763-777. Oxford United Kingdom : ACM. 26 juillet 2022. ISBN 978-1-4503-9247-1. DOI 10.1145/3514094.3534150.

WILLIS, Janine et TODOROV, Alexander, 2006. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science*. Vol. 17, no 7, pp. 592-598. DOI 10.1111/j.1467-9280.2006.01750.x.

WOLTER, Frank et SAVANI, Rahul, 2024. 7 surprising ways we're already using AI. *National Museums Liverpool* [en ligne]. 2024. Disponible à l'adresse : https://www.liverpoolmuseums.org.uk/stories/7-surprising-ways-were-already-using-ai [consulté le 8 août 2024].

Annexe 1 : Top 5 des images « Animaux » jugées comme les plus réelles

1
R Animal

2
R Animal

3
R Animal

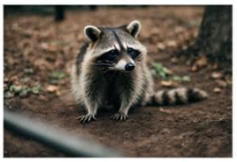
4
R Animal

5
R Animal

Annexe 2 : Top 5 des images « Animaux » jugées comme les moins réelles

1
AI Animal

2
AI Animal



3
AI Animal

4
AI Animal

5
AI Animal

Annexe 3 : Top 5 des images « Paysages » jugées comme les plus réelles

1
R Paysage

2
R Paysage

3
R Paysage

4
R Paysage

5
R Paysage

Annexe 4 : Top 5 des images « Paysages » jugées comme les moins réelles

1
AI Paysage

2
AI Paysage

3
R Paysage

4
AI Paysage

5
R Paysage

Annexe 5 : Top 5 des images « Tableaux » jugées comme les moins réelles

1
AI Tableau

2
AI Tableau

3
AI Tableau

4
AI Tableau

5
AI Tableau

Annexe 6 : Top 5 des images « Visages » jugées comme les moins réelles

1
R Visage

2
R Visage

3
AI Visage

4
R Visage

5
AI Visage