La conception de personas à l'ère de l'IA générative : vers une approche hybride homme-machine pour atténuer les biais cognitifs et algorithmiques[1]

Blangeois, Morgan

Clermont Recherche Management, Université Clermont Auvergne

morgan.blangeois@uca.fr

Résumé :

Cette communication propose une démarche exploratoire visant à associer l'expertise humaine aux capacités d'un système d'intelligence artificielle (IA) générative dans la conception de personas marketing. Elle souligne la présence de deux grandes familles de biais (cognitifs et algorithmiques) qui risquent de déformer la représentation des profils utilisateurs. Le prototype d'artefact présenté, développé en Python avec l'interface Gradio, sollicite un dialogue critique entre l'humain et la machine : le modèle d'IA propose des esquisses de personas, tandis que le concepteur humain affine les propositions et veille à repérer d'éventuels stéréotypes. Cette démarche entend former des futurs managers et concepteurs à une réflexion plus responsable, tout en leur offrant une créativité accrue. Les perspectives pédagogiques portent sur la sensibilisation aux mécanismes de biais et sur la pratique réflexive des applications d'IA. Enfin, le texte appelle à étendre et approfondir la méthodologie pour en évaluer l'efficacité empirique, tant par la recherche que par sa mise en œuvre concrète.

Mots-clés : IA générative, personas, biais, cocréation, réflexivité

[1] Candidat au prix Roland CALORI.

La conception de personas à l'ère de l'IA générative : vers une approche hybride homme-machine pour atténuer les biais cognitifs et algorithmiques

1. INTRODUCTION

1.1. CONTEXTE

L'émergence des modèles d'intelligence artificielle (IA) générative bouleverse profondément la manière dont les organisations communiquent et interagissent avec leurs publics (Korzynski et al., 2023). Capables de produire automatiquement du contenu original — texte, image ou son — à partir de données d'apprentissage (Feuerriegel et al., 2023), ces systèmes sont souvent qualifiés de « modèles de fondation » (Bommasani et al., 2022). Ils constituent en effet des briques technologiques générales, réutilisables pour de multiples applications, tout en conservant une part d'incertitude inhérente à leurs processus de génération. Comme ils s'appuient sur des données historiques, ils héritent toutefois de biais sociétaux déjà présents dans ces corpus.

Pour mieux comprendre l'IA générative et ses enjeux, il est utile de distinguer trois niveaux : le modèle, le système et l'application (Feuerriegel et al., 2023). Un modèle d'IA générative correspond à une architecture d'apprentissage automatique capable de créer de nouvelles données (textes, images, sons) en s'appuyant sur les régularités présentes dans son corpus d'entraînement. Les grands modèles génératifs capables de modéliser les données de manière polyvalente sont appelés *modèles de fondation* (Bommasani et al., 2022). Par exemple, GPT (*Generative Pre-trained Transformer*) est une famille de grands modèles de langage (LLM) utilisée pour la génération de texte. Un *système d'IA générative* intègre un modèle de fondation

et le peaufine pour des tâches spécifiques. Par exemple, un système peut utiliser GPT-4 pour générer du texte, mais il faut y ajouter des interfaces spécifiques, des flux de travail et des contrôles pour le rendre utilisable dans un contexte donné. Enfin, une *application d'IA générative* est une mise en œuvre concrète d'un système d'IA générative qui est directement utilisée par les utilisateurs finaux pour accomplir des tâches spécifiques. Par exemple, ChatGPT est une application qui utilise le système d'IA générative basé sur GPT-4, permettant aux utilisateurs d'interagir avec ce dernier pour générer du texte de manière conversationnelle. Dans notre cas, l'application en Python que nous développons est une application d'IA générative qui utilise un modèle de fondation intégré dans un système pour aider à la création de personas marketing.

Cette technologie trouve des applications concrètes dans de nombreux domaines, y compris dans le marketing et la conception de l'expérience utilisateur (UX), où l'IA générative offre la possibilité de produire rapidement du contenu riche, voire de suggérer des pistes créatives dès amont d'un projet (Chung, 2024). Parmi ces applications, la création de personas – ces portraits semi-fictifs de clients ou d'utilisateurs cibles – constitue un levier majeur (Chung, 2024). Les personas aident les équipes à cerner les besoins, motivations et comportements des usagers (Nielsen, 2019), agissant comme des guides pour concevoir et affiner une offre de produits ou de services. Ils ne devraient cependant pas être envisagés comme des représentations immuables ou exhaustives, mais bien comme des dispositifs malléables et susceptibles d'évoluer.

Pour autant, l'usage de l'IA générative introduit de nouveaux défis. D'une part, la recherche sur les biais cognitifs (Kahneman, 2011 ; Beeghly, 2015) suggère que la création manuelle de personas demeure vulnérable aux stéréotypes et aux erreurs de jugement (Goel et al., 2023). D'autre part, le recours à des modèles d'IA suscite des biais algorithmiques (Johnson, 2021 ;

Fazelpour & Danks, 2021), ce dernier pouvant perpétuer ou accentuer des préjugés hérités de ses données d'entraînement (Zhou et al., 2024), ce qui reflète et parfois aggrave les disparités et discriminations sociétales. Il faut souligner que la puissance statistique des modèles d'IA se borne à reconnaître des régularités préexistantes, sans « théoriser » par elle-même les phénomènes sociaux (Baumard, 2019), ni proposer de remèdes pour les dépasser. Face à ces enjeux, il est nécessaire d'interroger la légitimité des systèmes d'IA comme co-acteurs du processus de conception, particulièrement lorsqu'il s'agit de représenter la diversité et la complexité des profils utilisateurs. Cette démarche implique d'adopter une posture critique et réflexive pour éviter que ces outils ne renforcent les inégalités existantes (Vandangeon-Derumez & Saives, 2022).

Pour illustrer cette problématique, Johnson (2021) décrit le « proxy » : certains attributs apparemment neutres (comme le code postal) se révèlent étroitement corrélés à des facteurs sensibles (niveau socio-économique, appartenance ethnique – aux États-Unis, par exemple, les populations afro-américaines et hispaniques sont plus souvent cantonnées à des zones postales défavorisées), instaurant *de facto* des discriminations indirectes. Il devient donc essentiel de former les acteurs du marketing et du design aux techniques d'IA, mais aussi à la détection et à la prévention de ces biais, en tenant compte des contextes socio-économiques où les données prennent corps. Voilà tout l'enjeu du transfert de connaissances situées et de la créativité dans l'enseignement du management (Vandangeon-Derumez & Saives, 2022), ainsi que de la promotion d'une approche réflexive et critique, qui encourage les apprenants à investiguer les fondements mêmes des pratiques managériales.

1.2. PROBLEMATIQUE

Dans ce contexte, nous proposons une articulation homme-machine, qualifiée de « cyborg » (Dell'Acqua et al., 2023), où l'expertise humaine s'allie aux performances du système d'IA pour co-construire les personas, en reconnaissant leurs limites mutuelles et en valorisant leurs forces respectives. Nous posons la question de savoir si, grâce à une application d'IA générative, il est possible de réduire dans le même mouvement les biais cognitifs et algorithmiques, tout en maintenant l'humain au centre de la démarche et en instaurant une posture critique à l'égard des productions de la machine (Goel et al., 2023). Face à ces enjeux, notre recherche vise à explorer comment un outil d'IA générative spécifiquement conçu peut non seulement assister la création de personas mais aussi activement favoriser une démarche critique et réflexive chez le concepteur. Plus précisément, nous cherchons à répondre à la question de recherche principale suivante :

Comment une approche hybride homme-IA générative, outillée par une application dédiée (PersonaGenAI), peut-elle transformer la conception de personas marketing en favorisant l'atténuation des biais et la réflexivité critique du concepteur ?

Pour aborder cette question, nous nous appuyons sur trois sous-questions qui guident notre exploration :

1. ***Quels mécanismes spécifiques (cognitifs, algorithmiques, interactifs) sont à l'œuvre lors de la cocréation de personas via PersonaGenAI, et comment génèrent-ils ou révèlent-ils des biais ?***
2. ***Comment l'architecture et les fonctionnalités de PersonaGenAI (analyse de biais, suggestions de raffinement, options visuelles) sont-elles conçues pour inciter à une interaction critique et réflexive entre l'humain et l'outil d'IA ?***

3. ***Quelles sont les implications managériales et stratégiques d'une telle approche pour la prise de décision marketing, la gestion de l'innovation responsable et le développement des compétences managériales à l'ère de l'IA ?***

1.3. CONTRIBUTIONS ATTENDUES

Notre apport se veut à la fois méthodologique, illustratif et pédagogique. Il s'agit d'éclairer les mécanismes à l'origine des biais cognitifs et algorithmiques dans la création de personas assistée par IA, de présenter un artefact exploratoire (l'application PersonaGenAI en Python/Gradio) et de démontrer son fonctionnement via un guide illustratif. Ce guide vise à rendre tangible comment l'articulation homme-machine peut être mise en œuvre pour stimuler la réflexivité, de proposer un *artefact* exploratoire (une application en Python/Gradio) pour illustrer comment l'articulation homme-machine peut être mise en œuvre, tout en soulignant le rôle décisif de l'humain dans l'interprétation et la validation des sorties du modèle d'IA générative. Nous ouvrons également la réflexion sur l'usage de tels dispositifs dans la formation en management, en insistant sur l'importance d'une posture critique et réflexive chez les futurs managers face aux technologies algorithmiques (Vandangeon-Derumez & Saives, 2022) et sur les implications stratégiques d'une intégration responsable de l'IA.

Notre démarche demeure exploratoire et illustrative: elle ne s'appuie pas sur un protocole expérimental formel (comparaison entre conditions contrôle et expérimentale). Il s'agit plutôt d'un prototype et d'une démonstration de son mécanisme d'interaction critique, destinés à stimuler la réflexivité et la créativité, en interrogeant la place des systèmes d'IA dans la formation et la pratique managériales, tout en pointant les enjeux éthiques, managériaux et les limites de ces technologies (Orlikowski & Scott, 2023). En outre, cette exploration soulève des questions plus larges concernant la diffusion de tels outils au sein des organisations et les compétences managériales requises à l'ère de l'IA générative.

2. REVUE DE LA LITTERATURE

2.1. LA CREATION DE PERSONAS ET SES BIAIS COGNITIFS

Les personas visent à rendre compte de profils types d'utilisateurs ou de clients (Cooper, 1999 ; Nielsen, 2019). Ils représentent un outil privilégié dans les démarches de conception (*design thinking*, UX) et de ciblage marketing, car ils permettent aux équipes d'adopter la perspective de l'utilisateur final. Pourtant, de nombreux travaux soulignent leur vulnérabilité aux biais cognitifs (Kahneman, 2011) et au risque de véhiculer des stéréotypes (Beeghly, 2015), réduisant la complexité et la diversité humaines à des catégories simplifiées. Comme le soulignent Phil et Susan Turner (2011) dans leur étude sur la question « La stéréotypisation est-elle inévitable lors de la conception avec des personas ? », les personas créés sans données solides, basés principalement sur l'intuition des concepteurs, sont particulièrement enclins à refléter des stéréotypes marqués. Cooper (1999), pionnier de la méthode, mettait déjà en garde contre le risque que les « designers se conçoivent eux-mêmes » à travers leurs personas en l'absence d'une recherche utilisateur rigoureuse. Par exemple, lors de la création d'un persona marketing, le biais d'affinité peut conduire à sur-représenter des caractéristiques similaires à celles du concepteur, tandis que le biais de représentativité peut mener à assigner des traits stéréotypés basés sur des catégories démographiques simplifiées (ex : tous les « millennials » sont « *digital natives* » et écologistes). Le biais de confirmation peut également pousser à ne rechercher ou interpréter que les informations qui valident une hypothèse préconçue sur le client cible, transformant, selon Chapman & Milham (2006), les personas en « miroirs de nos suppositions » plutôt qu'en reflets de la réalité.

Ces biais, comme le souligne Beeghly (2015) à propos des stéréotypes, ne sont pas toujours intrinsèquement erronés mais deviennent problématiques lorsqu'ils enferment des populations

réelles dans des catégories rigides, ignorant la diversité et la complexité humaines. Dans un contexte marketing, cela peut engendrer des décisions inappropriées ou excluantes. D'où l'importance d'une démarche réflexive et critique lors de l'élaboration de personas, en questionnant sans cesse les présupposés et les représentations mobilisés (Vandangeon-Derumez & Saives, 2022). Les travaux fondateurs de Tversky et Kahneman (1974) sur les heuristiques de jugement et les biais, notamment ceux de représentativité ou de confirmation, éclairent les mécanismes cognitifs profonds qui rendent cette vigilance nécessaire.

On relève ainsi plusieurs biais : l'affinité (tendance à projeter ses propres préférences), la représentativité (assignation d'attributs génériques à un groupe sur la base de clichés ou stéréotypes) ou la confirmation (sélection et interprétation biaisées d'informations pour valider une hypothèse préconçue). Ces biais peuvent s'amplifier par l'usage de technologies qui, sans regard critique, procurent une illusion d'objectivité (Orlikowski & Scott, 2023), surtout si les biais algorithmiques de l'outil convergent avec les biais cognitifs de l'utilisateur. Il faut noter que même les approches basées sur les données ne sont pas exemptes de défis : Salminen et al. (2019), en analysant des personas générés automatiquement à partir de données de réseaux sociaux, ont mis en évidence la persistance de biais démographiques hérités des données sources, soulignant la nécessité d'un regard critique humain même sur les productions algorithmiques.

Dans les projets de formation, de tels biais engendrent souvent une vision réductrice des publics cibles ou la marginalisation involontaire de certains segments. Sur le plan managérial, ils entravent la réflexivité et la créativité (Vandangeon-Derumez & Saives, 2022), d'où la nécessité de mettre en place des stratégies pour atténuer leur impact dès le stade de la formation.

2.2. LES BIAIS ALGORITHMIQUES DANS L'IA GENERATIVE

Les études sur les biais algorithmiques (Fazelpour & Danks, 2021 ; Johnson, 2021) montrent que les modèles d'IA générative, malgré leur apparente « neutralité », peuvent délivrer des contenus empreints de préjugés. Ces travers découlent en partie de corpus d'entraînement biaisés ou de variables « proxy » associées à des critères sensibles (Johnson, 2021), reflétant les rapports de pouvoir et les inégalités qui structurent la société (Lambert & Gentelet, 2022). Il est crucial de reconnaître que les biais algorithmiques ne se limitent pas aux données d'entraînement (biais de données), mais peuvent aussi émerger lors de la définition des objectifs du modèle, de la mesure des performances (biais de mesure) ou de l'interprétation des résultats (biais d'évaluation) (Fazelpour & Danks, 2021). Comme le rappellent Bender et al. (2021) dans leur mise en garde contre les « dangers des perroquets stochastiques », plus un modèle est entraîné sur de vastes corpus non filtrés, plus il risque de « coder et renforcer les points de vue hégémoniques » et les discriminations associées. A ce titre, il paraît indispensable de prendre conscience qu'un système d'IA ne saurait être réduit à un outil neutre, puisqu'il incarne une vision du monde (Orlikowski & Scott, 2023), déjà véhiculée à l'origine par son concepteur.

Les travaux de Zhou et al. (2024) illustrent la façon dont la génération d'images via des modèles d'IA (Stable Diffusion, Midjourney, DALL-E) reproduit, voire accentue, les stéréotypes de genre ou de race, avec par exemple des femmes plus jeunes et souriantes ou des hommes plus âgés et plus neutres dans les mêmes dispositions professionnelles. De même, les modèles de langage peuvent générer des descriptions textuelles qui associent implicitement certaines professions, traits de personnalité ou niveaux de revenus à des groupes démographiques spécifiques, perpétuant ainsi des clichés sociaux. De même, l'analyse des modèles de langage révèle des biais préoccupants : Wyer & Black (2025) ont par exemple mis en évidence des « biais misogynes extrêmes » dans GPT-3, où les associations liées aux femmes gravitaient de manière disproportionnée autour de la violence sexualisée. Abid et al. (2021) ont quant à eux

démontré un « biais anti-musulman persistant » dans ce même modèle, associant fortement cette identité à des notions de violence. Ces biais subtils – souvent imperceptibles au premier regard – s'infiltrent dans la production créative et nuisent à la représentation plurielle. Ces observations sur les grands modèles de langage font écho à des découvertes antérieures sur les représentations sémantiques (*word embeddings*), où Bolukbasi et al. (2016) avaient déjà identifié des analogies sexistes du type « homme : programmeur = femme : femme au foyer ». Les biais présents dans les données d'entraînement des modèles de génération d'images sont également documentés, comme le montre l'analyse de Birhane et al. (2021) du dataset LAION-400M, révélant une proportion significative d'images sexualisantes ou dégradantes pour certains groupes.

De tels phénomènes ont des incidences directes sur la construction de personas par une application d'IA, susceptible de produire des portraits stéréotypés qui associent certains métiers à un unique genre ou qui véhiculent des représentations raciales partiales (Zhou et al., 2024). Le risque est d'autant plus grand que les biais algorithmiques peuvent renforcer insidieusement les biais cognitifs du concepteur si celui-ci accepte les propositions du système d'IA sans examen critique. Il est intéressant de noter, comme le suggèrent Babaei et al. (2024), que les biais de l'IA ne sont pas nécessairement identiques aux biais humains moyens ; une évaluation critique au cas par cas reste indispensable. Pourtant, Goel et al. (2023) indiquent que le système d'IA peut constituer un « point de départ » utile pour gagner du temps et nourrir la créativité, à condition que l'utilisateur adopte une vigilance critique envers ses productions. D'où l'importance, dans la formation des futurs managers, de développer des compétences pour évaluer, corriger et retravailler le contenu généré (Chung, 2024), mais aussi pour interroger les données d'apprentissage et les mécanismes du modèle d'IA générative lui-même, replacés dans leur contexte socio-historique et organisationnel.

2.3. L'APPROCHE HOMME-IA COMME REPONSE AUX DEFIS DE LA COCREATION

Pour contrer ces biais imbriqués (cognitifs et algorithmiques), plusieurs auteurs préconisent une approche collaborative où l'humain et le système d'IA coopèrent de manière itérative (Dell'Acqua et al., 2023 ; Goel et al., 2023). Tandis que le système propose un premier jet ; l'humain l'examine, le nuance, le remet en question et y réinjecte des valeurs, fort de son expérience et de sa connaissance du terrain (Vandangeon-Derumez & Saives, 2022).

Dell'Acqua et al. (2023) comparent cette dynamique à la figure du « Centaure » ou du « Cyborg », où l'humain et la machine conjuguent leurs atouts. Cette idée d'intelligence hybride trouve un écho dans l'analogie célèbre de Garry Kasparov (2017) issue du monde des échecs, où une équipe humain-machine bien coordonnée peut surpasser à la fois l'humain expert et la machine seule. Remo Pareschi (2024) formalise cette approche sous le nom de « design centaure », proposant un cadre méthodologique pour cette « fusion synergétique de l'intelligence humaine et artificielle » où les forces complémentaires sont exploitées : rapidité et analyse de l'IA d'un côté, jugement critique et éthique de l'humain de l'autre. Cette forme d'hybridation ouvre la voie à un gain de productivité et à une fécondité imaginative accrue, mais elle exige aussi une lucidité managériale : la co-construction n'a de sens que si l'on repère et corrige les biais propres au modèle d'IA, tout en interrogeant les valeurs et finalités associées (Orlikowski & Scott, 2023). Un défi majeur reste cependant d'éviter le « biais d'automatisation » (*automation bias*), c'est-à-dire la tendance humaine à accepter trop passivement les suggestions de la machine (Goddard et al., 2012), ce qui souligne l'importance d'une posture critique constante. Or, pour cela, il est nécessaire d'en connaître les caractéristiques inhérentes. L'interactivité induite s'inscrit dans une pédagogie par l'expérimentation. La confrontation

rapide avec un persona généré « par la machine » peut susciter, chez l'étudiant ou le concepteur, une réflexion critique et créative.

2.4. UN CADRE EXPLORATOIRE EN RECHERCHE EN SCIENCES DE LA CONCEPTION

La méthodologie de Recherche en Sciences de la Conception (Design Science Research, DSR) se prête souvent à la conception, au développement et à l'évaluation d'artefacts visant à résoudre un problème donné (Hevner et al., 2004 ; Peffers et al., 2007). Elle convient tout particulièrement à l'élaboration d'outils informatiques (tels qu'une application d'IA générative), via un processus méthodique d'identification du problème, de formulation des objectifs, de conception, de démonstration, d'évaluation et de diffusion. Toutefois, il importe de ne pas réduire la DSR à une pure dimension technique : elle doit s'inscrire dans une perspective plus large, qui tienne compte des incidences sociales et éthiques des artefacts (Orlikowski & Scott, 2023). Cette orientation est renforcée par l'évolution même de la DSR, qui intègre de plus en plus des approches interprétatives et participatives, reconnaissant que les artefacts s'insèrent dans des contextes humains et sociaux complexes. Ainsi, comme l'affirment Peffers et al. (2007), la validité d'un artefact n'est pas uniquement de nature technique, mais aussi organisationnelle et sociale. La dimension éthique est également cruciale : Myers & Venable (2014) ont proposé des principes éthiques spécifiques pour la DSR, soulignant la responsabilité du chercheur vis-à-vis des parties prenantes lors de la conception et de l'implantation d'artefacts, notamment en termes d'intérêt public, de consentement et d'honnêteté. *In fine*, l'adoption réelle de l'outil dépend de l'adhésion des concepteurs à ses principes, de leur capacité à déceler ses biais potentiels, mais aussi de leur aptitude à interroger la finalité même de l'artefact et ses retombées éventuelles sur les utilisateurs et la société (Lambert & Gentelet, 2022). Notre démarche s'inspire également de logiques visant à comprendre comment une

intervention (ici, notre application PersonaGenAI) interagit avec un contexte problématique (biais cognitifs et algorithmiques) pour activer des mécanismes (réflexivité critique, détection de biais) et générer des résultats souhaités (personas moins biaisés, apprentissage), une approche qui trouve des échos dans la logique CIMO (Denyer et al., 2008) souvent mobilisée en Design Science.

3. METHODOLOGIE

3.1. APPROCHE EXPLORATOIRE

Pour marquer la dimension exploratoire de notre démarche, nous précisons qu'il ne s'agit pas d'un protocole expérimental opposant un groupe de concepteurs usant d'une application d'IA générative à un groupe témoin. Notre ambition est plutôt de développer un prototype, d'en décrire le fonctionnement et d'en discuter les implications éventuelles.

En nous inscrivant dans le cadre DSR (Hevner et al., 2004 ; Peffers et al., 2007), nous avons franchi les étapes suivantes : identification du problème (biais cognitifs et algorithmiques), définition des objectifs (conception d'un outil destiné à détecter et atténuer ces biais), élaboration et développement (prototype en Python/Gradio), démonstration (présentation de l'interface), évaluation (encore informelle), puis communication (le présent article).

3.2. DESCRIPTION DE L'ARTEFACT

L'artefact développé, nommé PersonaGenAI, est une application web élaborée en Python, utilisant le framework Gradio pour son interface utilisateur interactive. Elle s'appuie sur des modèles d'IA générative accessibles via des API, principalement ceux d'OpenAI (GPT-4o-mini pour le texte, DALL-E 3 pour les images) ou, en cas d'indisponibilité de clé OpenAI, des modèles équivalents via OpenRouter (comme Mistral Large ou autres pour le texte, la

génération d'image étant alors désactivée). L'application PersonaGenAI est publiquement accessible pour démonstration sur la plateforme Hugging Face[2] et son code source principal est également disponible pour consultation[3]. L'application est structurée en plusieurs onglets guidant l'utilisateur à travers un processus itératif :

1. Configuration API : Permet à l'utilisateur de choisir et configurer l'accès aux modèles d'IA.
2. Objectif & analyse biais : L'utilisateur définit l'objectif marketing du persona. L'application utilise alors une fonction, *analyze_biases*, qui soumet cet objectif à un LLM (ex: GPT-4o-mini) avec un prompt spécifique l'instruisant d'identifier des biais cognitifs potentiels (stéréotypes, confirmation, simplification...) et de fournir des conseils d'atténuation dans un format JSON structuré. Ce retour immédiat vise à initier la réflexion critique dès le départ.
3. Image & infos base : L'utilisateur saisit les informations démographiques de base. La fonction *generate_persona_image* utilise un modèle de génération d'images (DALL-E 3) pour créer une représentation visuelle, tout en affichant un message de mise en garde sur les biais algorithmiques potentiels. Des options de personnalisation détaillées (teint, coiffure, expression, etc.) sont offertes pour permettre à l'utilisateur d'ajuster l'image et de contrer activement les stéréotypes visuels éventuels.
4. Profil détaillé & raffinement : L'utilisateur enrichit le persona avec des informations psychographiques, comportementales, etc. Pour chaque champ textuel, un bouton « 💡 » permet d'activer la fonction *refine_persona_details*. Celle-ci envoie le contenu du champ, ainsi que le contexte (objectif initial, biais identifiés précédemment), à un LLM pour obtenir 1 ou 2 suggestions concises visant à nuancer, approfondir ou dé-biaiser l'information saisie.

[2] Disponible à : https://huggingface.co/spaces/Moguiy/PersonaGenAI
[3] Disponible à : https://huggingface.co/spaces/Moguiy/PersonaGenAI/tree/main

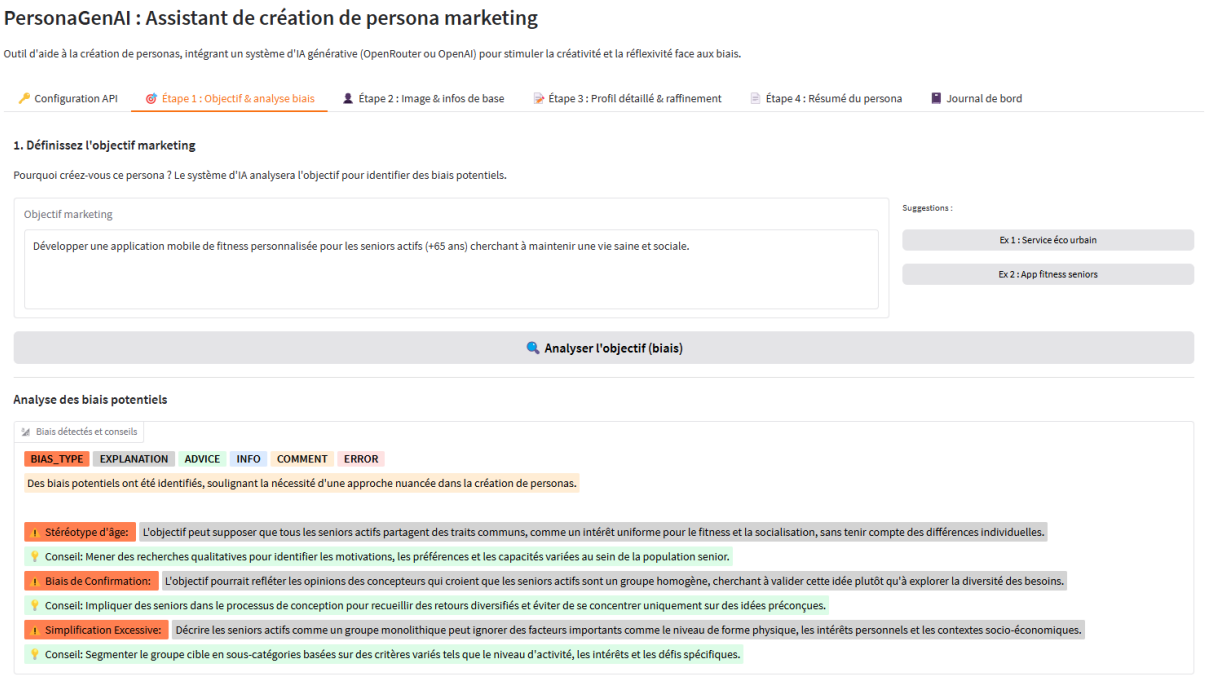
L'objectif est de stimuler l'approfondissement critique plutôt que de fournir des réponses toutes faites.

5. Résumé persona : Consolide toutes les informations et l'image en une fiche synthétique.
6. Journal de bord : Enregistre automatiquement les étapes clés, les alertes de biais, les suggestions de raffinement et les erreurs éventuelles pour une traçabilité et une analyse a posteriori du processus.

L'application n'a pas vocation à automatiser la création de personas, mais bien à servir de partenaire critique et créatif, médiatisant l'interaction entre l'intelligence humaine et l'IA générative dans une optique d'usage responsable.

4. PRESENTATION DU PROTOTYPE : DEMONSTRATION DU PROCESSUS VIA UN GUIDE ILLUSTRATIF

4.1. PHASE 1 : OBJECTIF & ANALYSE INITIALE DES BIAIS



Commentaire : La Figure 1 illustre la première étape interactive proposée par l'artefact PersonaGenAI, décrite dans la section 3.2, au sein de l'onglet « Étape 1 : Objectif & analyse

biais ». L'utilisateur saisit ici son objectif marketing initial ; dans l'exemple présenté, il s'agit de développer une application de fitness pour les « seniors actifs (+65 ans) ». L'activation de la fonction *analyze_biases* soumet ce texte à une analyse par un grand modèle de langage (ici, GPT-4o-mini ou un équivalent via OpenRouter). La section inférieure de la figure montre le résultat de cette analyse.

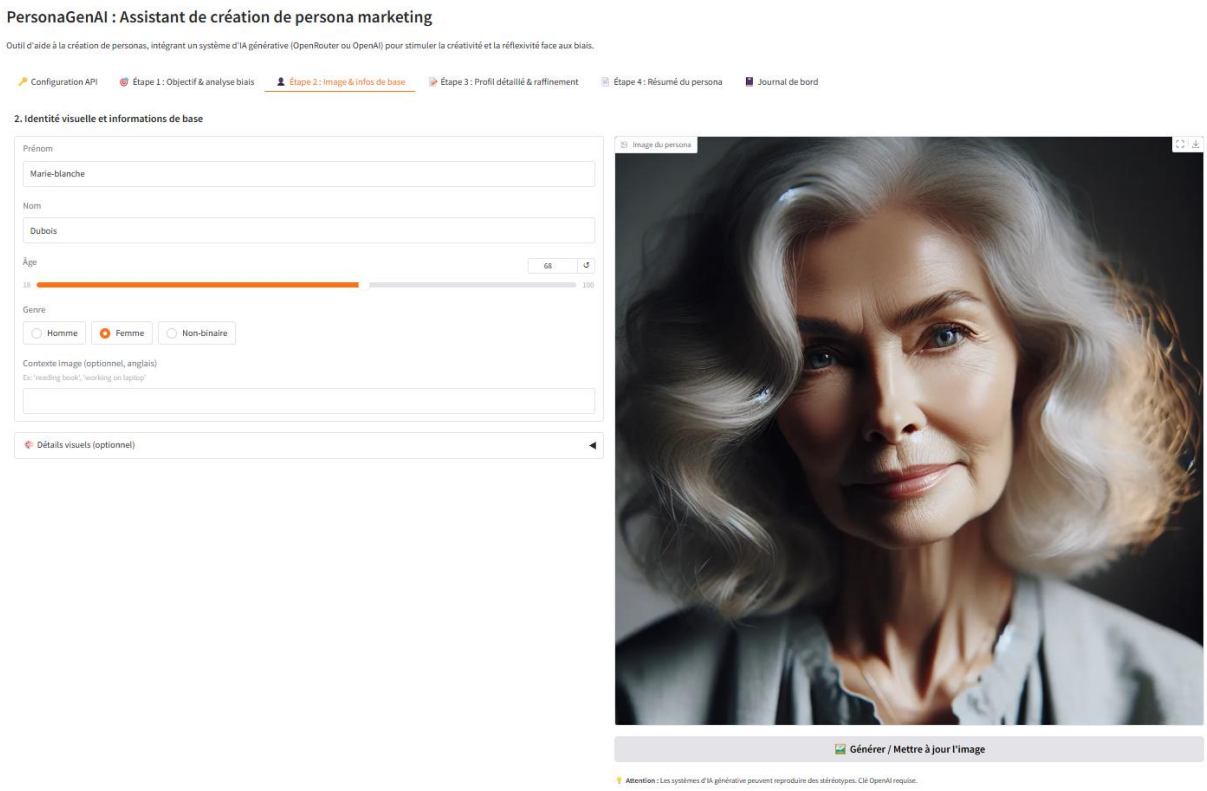
Le système a identifié trois biais cognitifs potentiels dans la formulation même de l'objectif :

1. **Stéréotype d'âge :** L'objectif pourrait présupposer une homogénéité des « seniors actifs », ignorant la diversité des profils, motivations et capacités individuelles au sein de cette tranche d'âge.
2. **Biais de confirmation :** La formulation pourrait refléter une hypothèse préconçue sur ce groupe, incitant à chercher des confirmations plutôt qu'à explorer la diversité réelle des besoins et des désirs.
3. **Simplification excessive :** Décrire les « seniors actifs » de manière monolithique risque d'occulter des facteurs importants de différenciation (niveau de revenu, intérêts personnels, contexte social, etc.).

Pour chaque biais détecté, l'outil fournit une brève explication et un conseil visant à l'atténuation (par exemple, mener des recherches qualitatives supplémentaires, segmenter la cible sur des critères plus fins, impliquer les utilisateurs dans la conception).

Cette fonctionnalité intervient avant même la définition des caractéristiques du persona. Son objectif est d'initier immédiatement une démarche réflexive chez le concepteur. En le confrontant aux biais potentiels contenus dans ses propres intentions initiales, l'outil l'invite à adopter une posture critique dès l'amont du processus de création et à potentiellement affiner ou nuancer son objectif avant de poursuivre. Cela illustre le premier mécanisme par lequel PersonaGenAI tente de favoriser l'atténuation des biais cognitifs.

4.2. PHASE 2 : IDENTITE DE BASE & COCREATION VISUELLE



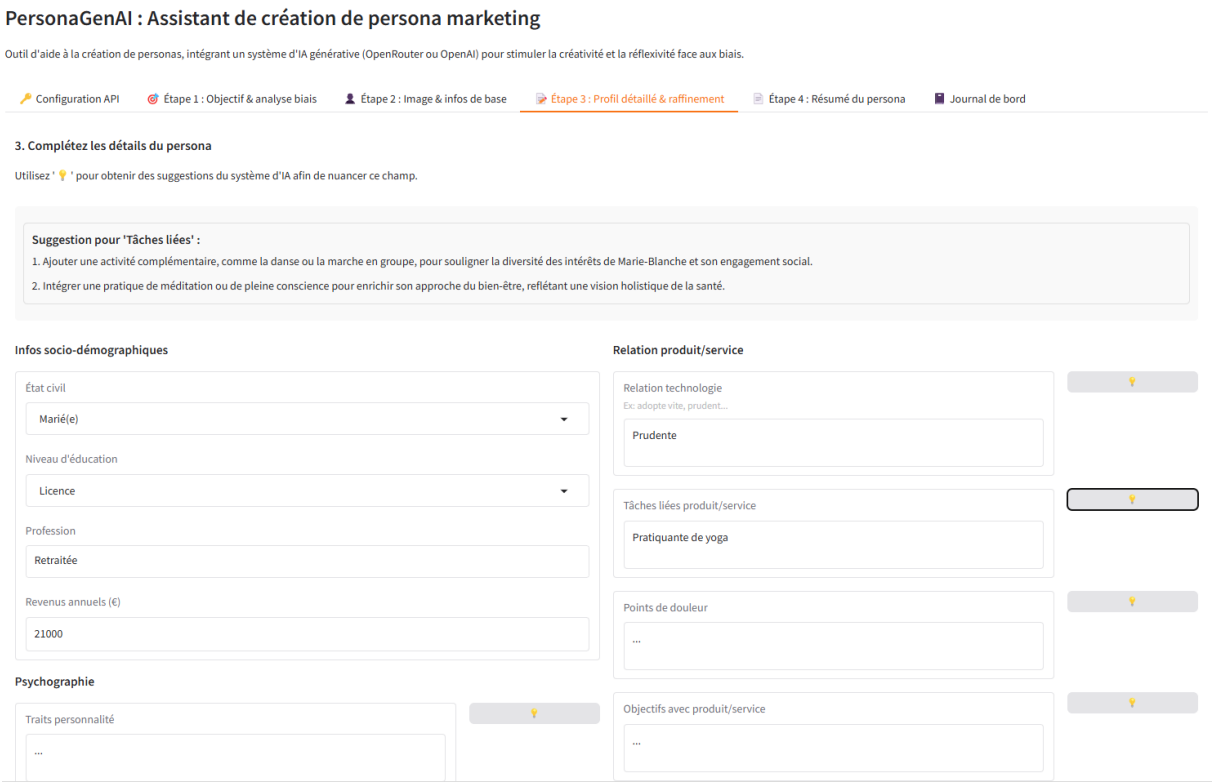
Commentaire : Cette figure présente l'interface de l'étape 2 de PersonaGenAI, « Image & infos de base », dédiée à la cocréation visuelle du persona. L'utilisateur a renseigné les informations démographiques fondamentales (ici, Prénom : Marie-Blanche, Nom : Dubois, Âge : environ 68 ans via le curseur, Genre : Femme). L'activation de la fonction *generate_persona_image*, qui utilise un modèle de génération d'images comme DALL-E 3, a produit le portrait affiché sur la partie droite.

Conformément aux enjeux soulevés dans notre revue de la littérature concernant les biais algorithmiques dans la génération d'images (section 2.2, citant par exemple Zhou et al., 2024), un message d'avertissement est systématiquement affiché sous l'image générée (partiellement visible en bas à droite). Il rappelle à l'utilisateur que « les systèmes d'IA générative peuvent reproduire ou amplifier » les stéréotypes présents dans leurs données d'entraînement. Il est

d'ailleurs intéressant de noter, avec une pointe d'ironie critique, que l'image initialement proposée pour « Marie-Blanche Dubois », 68 ans (Figure 2), représente une femme caucasienne, aux cheveux argentés impeccablement coiffés, souriante mais à l'expression neutre et posée, dans un cadre sobre. Cette représentation, bien que plausible, pourrait aisément être interprétée comme l'incarnation d'une certaine norme idéalisée de la « senior active », potentiellement issue des visions dominantes véhiculées par les données d'entraînement ou reflétant implicitement la perspective des concepteurs du modèle. Cela illustre concrètement la nécessité de la vigilance critique annoncée. Cette alerte et cette observation visent à inciter à la vigilance critique face à la représentation visuelle proposée par la machine.

Au-delà de l'alerte, l'interface offre des moyens d'intervention au concepteur. La section « Détails visuels [optionnel] » (visible mais repliée sur cette capture) permet à l'utilisateur de spécifier ou de modifier de nombreuses caractéristiques (teint de peau, coiffure, couleur des yeux, expression faciale, style vestimentaire, etc.) et de régénérer l'image. Cette fonctionnalité est essentielle : elle positionne l'humain non pas comme un simple récepteur passif de l'image générée, mais comme un acteur capable d'ajuster, de corriger ou de nuancer activement la représentation pour, par exemple, éviter un stéréotype perçu ou mieux refléter la diversité envisagée pour le persona.

4.3. PHASE 3 : ENRICHISSEMENT & RAFFINEMENT ASSISTE



Commentaire : La Figure 3 illustre le fonctionnement de l'étape 3 de PersonaGenAI, « Profil détaillé & raffinement ». Cette phase permet au concepteur d'ajouter des informations détaillées sur le persona dans différentes catégories (socio-démographiques, psychographiques, relation au produit/service, etc.). Pour chaque champ de saisie, une icône « 💡 » est disponible, permettant à l'utilisateur de solliciter une suggestion spécifique de la part du système d'IA pour enrichir ou nuancer l'information saisie, via la fonction *refine_persona_details*.

Dans l'exemple présenté, l'utilisateur a rempli plusieurs champs, notamment « Tâches liées produit/service » avec la mention « Pratiquante de yoga ». Après avoir cliqué sur l'icône « 💡 » associée à ce champ, le système a généré des suggestions affichées dans le bandeau supérieur. Ces suggestions visent à aller au-delà de l'information initiale :

1. Elles proposent d'ajouter une activité complémentaire (danse, marche en groupe) pour souligner la diversité des intérêts potentiels de « Marie-Blanche » et son possible engagement social, évitant de la réduire à une seule activité.

2. Elles suggèrent d'intégrer une autre pratique (méditation, pleine conscience) pour enrichir l'approche du bien-être et refléter une vision plus holistique de la santé, potentiellement en lien avec l'objectif initial de « maintenir une vie saine et sociale ».

Ces propositions ne sont pas conçues pour remplacer l'entrée de l'utilisateur mais pour agir comme des stimuli réflexifs. Elles invitent le concepteur à :

- **Complexifier le profil :** Envisager des facettes additionnelles ou complémentaires du persona.
- **Éviter la simplification excessive :** Ne pas s'arrêter à une caractéristique unique ou potentiellement stéréotypée.
- **Relier les attributs entre eux :** S'interroger sur la cohérence et la richesse du profil global au regard des objectifs marketing.

L'utilisateur conserve l'entière maîtrise : il peut choisir d'ignorer ces suggestions, de les intégrer telles quelles, ou de s'en inspirer pour modifier son texte différemment. Cette fonctionnalité incarne l'approche de cocréation où le système d'IA n'est pas une source de vérité, mais un outil qui, par ses propositions contextuelles, peut aider l'humain à approfondir sa réflexion, à identifier des angles morts potentiels et à nuancer la représentation finale du persona.

4.4. ANALYSE DU POTENTIEL REFLEXIF ET PEDAGOGIQUE ILLUSTRE

Ce guide illustre comment PersonaGenAI est conçu pour fonctionner non pas comme un générateur automatique de personas, mais comme un *catalyseur de réflexivité* pour le concepteur. Les fonctionnalités clés – l'analyse initiale des biais sur l'objectif (*analyze_biases*), les avertissements et options de personnalisation pour l'image (*generate_persona_image*), et les suggestions de raffinement contextuelles (*refine_persona_details*) – agissent comme des

points de friction critiques ou des *invites à la délibération*. Elles sont conçues pour interrompre un flux de travail potentiellement routinier ou biaisé, en invitant l'utilisateur à :

- Questionner ses propres hypothèses initiales (biais cognitifs).
- Examiner de manière critique les propositions générées par le système d'IA.
- Complexifier et nuancer activement la représentation de l'utilisateur final.
- Maintenir l'agentivité humaine au centre du processus de conception.

Sur le plan pédagogique, cette interaction *simulée* montre comment l'outil pourrait être utilisé dans un cadre de formation. En confrontant directement les apprenants à des biais potentiels (les leurs et ceux de l'outil d'IA) et en leur offrant des moyens d'y répondre de manière interactive, PersonaGenAI vise à développer une compétence essentielle pour les futurs managers et concepteurs : la capacité à collaborer de manière critique et constructive avec les systèmes d'IA générative, en reconnaissant à la fois leur potentiel et leurs limites.

5. DISCUSSIONS

L'approche hybride homme-IA et l'artefact PersonaGenAI, tels que présentés à travers notre démonstration guidée, soulèvent plusieurs points de discussion. Ces réflexions concernent non seulement le processus de conception de personas mais aussi des questions relatives au management et à l'intégration de l'intelligence artificielle dans les organisations dans un contexte de démocratisation rapide de ces technologies.

En premier lieu, l'interaction médiatisée par PersonaGenAI illustre la gestion de tensions inhérentes à l'intégration de l'IA dans les pratiques de travail. Ces tensions peuvent être analysées comme des paradoxes managériaux. Comme le soulignent Raisch & Krakowski (2021), les managers sont confrontés au « paradoxe automatisation-augmentation », devant trouver le juste équilibre entre déléguer des tâches à l'IA et augmenter les capacités humaines

tout en maintenant le contrôle. L'outil tente de concilier l'efficacité et le potentiel créatif associés à l'IA générative avec la nécessité de rigueur méthodologique, de considération éthique et de nuance apportée par l'intervention humaine critique. Dans ce cadre, le manager n'est plus face à une alternative binaire entre approche humaine ou automatisée, mais doit plutôt développer des capacités d'orchestration de cette collaboration homme-machine, visant un équilibre dynamique. Cette orchestration suggère le développement de nouvelles compétences managériales, relevant potentiellement du « Centaure » où le jugement humain reste central malgré l'assistance de l'IA.

En deuxième lieu, la capacité à générer des représentations d'utilisateurs plus nuancées et moins stéréotypées, potentiellement facilitée par les mécanismes critiques de l'outil, ne constitue pas seulement une amélioration tactique. Elle peut renforcer la capacité de l'organisation à interpréter la complexité et la diversité de ses marchés. Des personas plus précis et approfondis sont susceptibles de conduire à une meilleure identification des besoins, à une innovation plus ciblée et, par conséquent, à une adaptation stratégique informée dans des environnements évolutifs. La qualité des représentations internes des clients peut ainsi constituer un levier pour l'analyse stratégique.

En troisième lieu, l'utilisation de l'IA générative dans les processus organisationnels comporte des risques identifiés (éthiques, réputationnels, de marché), notamment liés à la perpétuation ou l'amplification des biais. Adopter une démarche critique et réflexive, soutenue par un artefact comme PersonaGenAI, peut être interprété comme une forme de gestion anticipative de ces risques. Au-delà de la simple mise en conformité, une telle démarche, si elle est intégrée dans les pratiques, peut signaler un engagement envers une utilisation responsable de l'IA, ce qui pourrait influencer la perception des parties prenantes. Cette approche s'inscrit dans une démarche plus large d'« IA responsable » ou d'« IA digne de confiance » qui nécessite une

gouvernance claire (Batool et al., 2023) et, idéalement, une intégration de l'éthique dès la conception (« *Ethics by Design* », issue de la littérature professionnelle). Toutefois, la facilité avec laquelle des outils similaires à PersonaGenAI peuvent être développés (Python, Gradio, APIs) soulève la question du « Shadow IT » ou de l'informatique fantôme : comment les organisations peuvent-elles encadrer l'usage d'outils d'IA générative développés et utilisés en dehors des canaux officiels, potentiellement sans supervision adéquate des biais ou des risques associés ? La prolifération de ces applications « maison » pourrait devenir un enjeu de gouvernance majeur, allant à l'encontre des efforts pour établir des cadres de responsabilité clairs. Ce défi est d'autant plus prégnant que ces outils reposent sur des modèles d'IA générative (LLM) fondamentalement probabilistes, marquant un changement de paradigme par rapport au développement logiciel traditionnel déterministe. L'imprévisibilité inhérente et la nature généraliste de ces technologies rendent les approches classiques par cas d'usage ou « points de douleur » métier potentiellement insuffisantes pour anticiper tous les détournements possibles. En effet, même un outil conçu pour une tâche spécifique comme la création de personas (PersonaGenAI) pourrait être vulnérable au « prompt hacking » : un utilisateur malveillant pourrait tenter d'exploiter les capacités sous-jacentes du LLM pour générer des contenus inappropriés, offensants ou nuisibles (par exemple, des stéréotypes extrêmes, des discours haineux déguisés), portant ainsi atteinte à la réputation de l'organisation qui déploie ou cautionne l'outil. La question n'est plus seulement de savoir si l'outil peut être utilisé correctement, mais comment gérer le risque qu'il puisse être utilisé de manière malveillante ou non anticipée, un défi pour lequel les intégrateurs et les entreprises cherchent encore des solutions robustes.

Enfin, cette approche suggère un rôle pour le manager qui n'est pas celui d'un simple utilisateur passif de technologie, mais plutôt celui d'un décideur dont les capacités sont potentiellement

augmentées par l'IA. L'intelligence artificielle fournit des analyses préliminaires ou des suggestions, mais l'interprétation finale, l'arbitrage éthique et l'ajustement contextuel demeurent des prérogatives humaines. Cela rejoint les principes de l'IA centrée sur l'humain (HCAI), qui insistent sur la primauté du contrôle et de la responsabilité humaine. Cela suggère l'importance du développement de compétences managériales spécifiques axées sur le jugement critique, l'évaluation des *outputs* algorithmiques et la capacité à interagir de manière éclairée avec les systèmes d'IA. Cependant, un défi considérable émerge : celui de l'explicabilité et de la responsabilité lorsque les managers utilisent des outils dont ils ne maîtrisent pas nécessairement les mécanismes internes. Même si le code de PersonaGenAI est accessible, sa compréhension peut rester hors de portée pour des non-spécialistes. Comment un manager peut-il alors justifier une décision basée sur une suggestion de l'outil s'il ne peut en expliquer le fondement ? Ce déficit potentiel de compétence technique chez les utilisateurs finaux pose des questions sur la formation nécessaire, visant une forme de « littératie algorithmique », mais aussi sur la conception même de ces outils : doivent-ils intégrer des mécanismes d'explicabilité adaptés aux managers, ou faut-il repenser la collaboration entre experts techniques et utilisateurs métier ? La question de l'ouverture du code (« *open source* ») trouve ici ses limites si elle ne s'accompagne pas d'une forme de littératie algorithmique chez les managers.

Concernant la créativité et la réflexivité, l'approche combinant humain et IA (parfois qualifiée de « *cyborg* », Dell'Acqua et al., 2023) suggère un potentiel. Notre exemple d'utilisation commenté a illustré comment les suggestions générées par l'IA peuvent contribuer à générer des idées ou à explorer de nouvelles dimensions lors de la définition d'un persona. Cependant, cette créativité potentiellement augmentée est conditionnée par le maintien d'une posture réflexive de la part de l'utilisateur. Il existe un risque de dépendance cognitive ou de confiance excessive envers les propositions de la machine (le fameux « *automation bias* » déjà évoqué),

ce qui pourrait avoir pour conséquence paradoxale d'appauvrir la diversité des perspectives ou de renforcer certains biais si les suggestions ne sont pas examinées de manière critique. La valeur ajoutée semble donc résider dans la capacité du concepteur à instaurer une interaction où l'outil d'IA est utilisé comme un outil de questionnement critique, dont les propositions sont systématiquement évaluées et intégrées au jugement humain. Un accompagnement pédagogique et une compréhension minimale des principes de fonctionnement de l'IA apparaissent donc importants pour cultiver cette interaction.

Il convient également de considérer les limites actuelles de notre proposition. L'artefact PersonaGenAI a été présenté via une démonstration guidée et n'a pas fait l'objet d'une évaluation empirique rigoureuse. La qualité de la détection des biais et la pertinence des suggestions dépendent des capacités et des biais potentiels du modèle de langage sous-jacent. De plus, le risque de « technosolutionnisme », c'est-à-dire l'attente que l'outil résolve à lui seul le problème complexe des biais, mérite d'être considéré. Une application mécanique des suggestions, sans discernement, pourrait conduire à des personas apparemment moins biaisés mais potentiellement aseptisés ou déconnectés des réalités sociales. Se pose alors la question de l'équilibre entre la description fidèle d'une situation existante (qui peut inclure des inégalités) et la promotion active de l'équité dans la représentation. Cet arbitrage éthique dépasse les capacités techniques de l'outil et relève de la responsabilité du concepteur humain, soulevant à nouveau des enjeux de gouvernance sur les finalités assignées à ces systèmes. Enfin, une analyse plus large pourrait intégrer les impacts environnementaux et sociaux associés au déploiement de ces technologies. Les pistes de recherche futures pourraient donc s'orienter, dans un premier temps, vers des études empiriques qualitatives (observations d'usage, entretiens) pour comprendre comment les concepteurs interagissent avec l'outil et mobilisent (ou non) une posture réflexive. Des études comparatives pourraient ensuite être envisagées pour

évaluer plus formellement les effets de cette approche, en examinant notamment les variations d'usage et d'impact selon les contextes organisationnels et les profils de compétence des utilisateurs.

Des recherches futures pourraient évaluer plus formellement l'efficacité de cet artefact comme intervention, en analysant précisément les mécanismes (M) par lesquels l'interaction avec l'outil (I) dans le contexte (C) de la création de personas conduit à des résultats (O) en termes de réduction des biais et de développement de la réflexivité, s'inscrivant ainsi dans un cadre explicatif de type CIMO.

6. APPORTS POUR L'ENSEIGNEMENT DU MANAGEMENT DANS UN CONTEXTE TECHNOLOGIQUE

6.1. POTENTIEL ET DEFIS PEDAGOGIQUES DE L'IA GENERATIVE

L'application présentée peut trouver sa place dans des cursus de marketing, de systèmes d'information ou d'innovation. L'utilisation pédagogique des applications d'IA générative comme ChatGPT fait l'objet d'une attention croissante, avec des bénéfices potentiels identifiés tels que le gain de temps dans la création de matériel pédagogique, le feedback personnalisé pour les étudiants ou la stimulation de la créativité (Kasneci et al., 2023). Par la pratique (même simulée ou guidée), elle contribuerait à sensibiliser les apprenants aux biais cognitifs et algorithmiques. Cependant, des défis importants subsistent : risque d'hallucinations factuelles, problèmes de plagiat (Cotton et al., 2023) et nécessité d'adapter les méthodes d'évaluation. Surtout, elle offre un terrain d'expérimentation pour développer des compétences managériales clés identifiées précédemment (Section 5) : la pensée critique face aux *outputs* de l'IA, la capacité de collaboration homme-machine constructive, et le jugement éthique appliqué à des processus de conception assistés par technologie. L'objectif est de former des futurs managers

capables non seulement d'utiliser les outils d'IA, mais aussi d'en comprendre les enjeux, d'en déjouer les pièges et d'en orienter l'usage de manière responsable (Goel et al., 2023).

6.2. UN DISPOSITIF D'APPRENTISSAGE TRANSVERSAL

L'enseignement d'un management responsable engage une réflexion éthique et une attention active à la diversité. Les personas, à la fois outil de conception et miroir de nos représentations sociales (Beeghly, 2015), invitent à ce questionnement. L'insertion d'un dispositif comme PersonaGenAI dans un parcours pédagogique ne vise pas seulement à enseigner une technique, mais à ancrer la pratique de l'éthique de l'IA. Il sensibilise les futurs managers à leur propre perception et aux biais potentiels de la machine, et les pousse à interroger la finalité de leurs décisions, en les plaçant en position d'acteurs critiques qui doivent interpréter, arbitrer et assumer la responsabilité des résultats, même lorsqu'ils sont co-produits avec un système d'IA. Cela répond à la nécessité de former des managers dotés d'une littératie algorithmique et d'une conscience éthique adaptées aux enjeux de l'intégration responsable de l'IA dans les organisations. L'ancrage dans une démarche de DSR attentive à ses implications éthiques (Myers & Venable, 2014) est ici essentiel pour guider les développements futurs.

7. CONCLUSION

Dans ce texte, nous avons exposé un prototype (PersonaGenAI) conçu pour accompagner la création de personas selon une approche hybride, associant intelligence humaine et IA générative. Notre contribution principale est d'avoir proposé et illustré, via une démonstration guidée détaillée, une démarche méthodologique et un artefact visant à favoriser l'atténuation des biais cognitifs et algorithmiques. L'originalité de l'approche réside dans le fait qu'elle cherche à atteindre cet objectif non pas par une automatisation accrue, mais au contraire par la

stimulation de la réflexivité critique du concepteur humain, médiatisée par des interactions spécifiques avec l'outil. Ces éléments nous permettent de répondre à nos questions de recherche de la manière suivante :

- **Concernant notre première sous-question (SQ1) sur les mécanismes des biais :** Notre analyse et la démonstration guidée ont mis en évidence comment les biais cognitifs du concepteur (ex : hypothèses initiales dans l'objectif) peuvent interagir avec les biais potentiels de l'IA (ex : stéréotypes dans les images ou textes générés). L'outil PersonaGenAI, par ses interventions (analyse d'objectif, avertissements), vise précisément à *rendre visibles* ces points de friction et ces risques de convergence biaisée au sein du processus de cocréation.
- **Pour répondre à la deuxième sous-question (SQ2) sur l'incitation à l'interaction critique :** Nous avons montré comment l'architecture même de PersonaGenAI et ses fonctionnalités clés (l'analyse structurée des biais de l'objectif, les suggestions de raffinement contextuelles, les options de personnalisation visuelle accompagnées d'alertes) sont *conçues intentionnellement* comme des « invites à la délibération » ou des « points de friction critiques ». Elles ne fournissent pas de solution clé en main mais cherchent à engager l'utilisateur dans un dialogue réflexif avec ses propres idées et avec les productions de la machine.
- **Enfin, s'agissant des implications managériales et stratégiques (SQ3) :** Notre discussion (Section 5) a exploré comment une telle approche dépasse le cadre opérationnel de la création de personas. Elle touche à la capacité des managers à naviguer les paradoxes de l'IA, à améliorer le « *sensing* » organisationnel, à gérer les risques liés à l'outil d'IA de manière proactive, et à développer les compétences nécessaires pour une prise de décision

augmentée mais critique. L'intégration responsable de l'IA devient ainsi un enjeu de management stratégique.

Globalement, ces éléments nous permettent de proposer une réponse à notre question principale : l'approche hybride incarnée par PersonaGenAI peut potentiellement transformer la conception de personas non pas en la rendant plus automatique, mais en la reconfigurant comme un processus de co-construction médiatisé, réflexif et critique. La transformation réside dans l'accent mis sur la vigilance humaine et le dialogue homme-machine, ouvrant la voie à des représentations potentiellement plus justes et à une intégration plus responsable et stratégique de l'IA. Ce travail demeure exploratoire et illustratif, mais il ouvre des pistes pour de nouvelles pratiques pédagogiques et managériales, tout en soulignant des défis stratégiques majeurs liés à la gouvernance de l'IA, à l'évolution des compétences et à l'explicabilité des systèmes dans les organisations.

Notre travail s'inscrit dans une démarche de Recherche en Sciences de la Conception, en proposant et démontrant un artefact destiné à nourrir l'apprentissage et la réflexion critique. L'ambition est de former des managers capables d'employer l'IA de façon créative et critique, conscients de l'existence de multiples biais et compétents pour les interroger et les gérer dans leur pratique (Vandangeon-Derumez & Saives, 2022).

REFERENCES BIBLIOGRAPHIQUES

Abid, A., Farooqi, M., & Zou, J. (2021). *Persistent Anti-Muslim Bias in Large Language Models* (No. arXiv:2101.05783). arXiv. https://doi.org/10.48550/arXiv.2101.05783

Babaei, Golnoosh, David Banks, Costanza Bosone, Paolo Giudici, et Yunhong Shan. « Is ChatGPT More Biased Than You? » *Harvard Data Science Review* 6, nº 3 (31 juillet 2024). https://doi.org/10.1162/99608f92.2781452d.

Batool, A., Zowghi, D., & Bano, M. (2023). *Responsible AI Governance : A Systematic Literature Review* (No. arXiv:2401.10896). arXiv. https://doi.org/10.48550/arXiv.2401.10896

Baumard, P. (2019). Quand l'intelligence artificielle théorisera les organisations. *Revue française de gestion*, *285*(8), 135-159. https://doi.org/10.3166/rfg.2020.00409

Beeghly, E. (2015). What is a Stereotype? What is Stereotyping? *Hypatia*, *30*(4), 675-691. https://doi.org/10.1111/hypa.12170

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots : Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. https://doi.org/10.1145/3442188.3445922

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). *Multimodal datasets : Misogyny, pornography, and malignant stereotypes* (No. arXiv:2110.01963). arXiv. https://doi.org/10.48550/arXiv.2110.01963

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (No. arXiv:1607.06520). arXiv. https://doi.org/10.48550/arXiv.1607.06520

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. « On the Opportunities and Risks of Foundation Models ». arXiv, 12 juillet 2022. https://doi.org/10.48550/arXiv.2108.07258.

Chapman, C. N., & Milham, R. P. (2006). The Personas' New Clothes : Methodological and Practical Arguments against a Popular Method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(5), 634-636. https://doi.org/10.1177/154193120605000503

Chung, S. J. (2024). Revolutionizing Persona Design with Generative AI : Insights from Experts. *The International Journal of Design Management and Professional Practice*, *18*(2), 109-124. https://doi.org/10.18848/2325-162X/CGP/v18i02/109-124

Cooper, A. (1999). The Inmates are Running the Asylum. In U. Arend, E. Eberleh, & K. Pitschke (Éds.), *Software-Ergonomie '99 : Design von Informationswelten* (p. 17-17). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-99786-9_1

Cotton, D. R. E., Cotton ,Peter A., & and Shipway, J. R. (2023). Chatting and cheating : Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the Jagged Technological Frontier : Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* (SSRN Scholarly Paper No. 4573321). Social Science Research Network. https://doi.org/10.2139/ssrn.4573321

Denyer, D., Tranfield, D., & Van Aken, J. E. (2008). Developing Design Propositions through Research Synthesis. Organization Studies, 29(3), 393-413. https://doi.org/10.1177/0170840607088020

Fazelpour, S., & Danks, D. (2021). Algorithmic bias : Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760. https://doi.org/10.1111/phc3.12760

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4443189

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias : A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121-127. https://doi.org/10.1136/amiajnl-2011-000089

Goel, T., Shaer, O., Delcourt, C., Gu, Q., & Cooper, A. (2023). Preparing Future Designers for Human-AI Collaboration in Persona Creation. *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, 1-14. https://doi.org/10.1145/3596671.3598574

Hevner, A., R, A., March, S., T, S., Park, Park, J., Ram, & Sudha. (2004). Design Science in Information Systems Research. *Management Information Systems Quarterly*, *28*, 75.

Is ChatGPT More Biased Than You ? · Issue 6.3, Summer 2024. (s. d.). Consulté 7 avril 2025, à l'adresse https://hdsr.mitpress.mit.edu/pub/qh3dbdm9/release/2

Johnson, G. M. (2021). Algorithmic bias : On the implicit biases of social technology. *Synthese*, *198*(10), 9941-9961. https://doi.org/10.1007/s11229-020-02696-y

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kasparov, G. (2017). *Deep Thinking : Where Machine Intelligence Ends and Human Creativity Begins*. PublicAffairs.

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories : Analysis of ChatGPT. *Central European Management Journal*, *31*(1), 3-13. https://doi.org/10.1108/CEMJ-02-2023-0091

Lambert, S., & Gentelet, K. (2022, août 8). *Voici pourquoi l'intelligence artificielle ne peut être considérée comme un simple outil*. The Conversation. http://theconversation.com/voici-pourquoi-lintelligence-artificielle-ne-peut-etre-consideree-comme-un-simple-outil-186014

Myers, M. D., & Venable, J. R. (2014). A set of ethical principles for design science research in information systems. *Information & Management*, *51*(6), 801-809. https://doi.org/10.1016/j.im.2014.01.002

Nielsen, L. (2019). *Personas—User Focused Design*. Springer.

Orlikowski, W. J., & Scott, S. V. (2023). The Digital Undertow and Institutional Displacement : A Sociomaterial Approach. *Organization Theory*, *4*(2), 26317877231180898. https://doi.org/10.1177/26317877231180898

Pareschi, R. (2024). Beyond Human and Machine : An Architecture and Methodology Guideline for Centaurian Design. *Sci*, *6*(4), Article 4. https://doi.org/10.3390/sci6040071

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45-77.

Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management : The Automation–Augmentation Paradox. Academy of Management Review, 46(1), 192-210. https://doi.org/10.5465/amr.2018.0072

Salminen, J., Jansen, B. J., & Soongyo, J. (2019). Detecting demographic bias in automatically generated personas : 2019 CHI Conference on Human Factors in Computing Systems, CHI EA 2019. *CHI EA 2019 - Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290607.3313034

Turner, P., & Turner, S. (2011). Is stereotyping inevitable when designing with personas? *Design Studies*, *32*(1), 30-44. https://doi.org/10.1016/j.destud.2010.06.002

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty : Heuristics and Biases. *Science (New York, N.Y.)*, *185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Vandangeon-Derumez, I., & Saives, A.-L. (2022). L'enseignement créatif et critique du management : En partenariat avec le groupe thématique MACCA (Méthodes et Approches Créatives et Critiques de l'Apprentissage et de la formation au Management) de l'AIMS. *Finance Contrôle Stratégie, NS-13*. https://doi.org/10.4000/fcs.10133

Wyer, S., & Black, S. (2025). Algorithmic bias : Sexualized violence against women in GPT-3 models. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00641-0

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). *Bias in Generative AI* (No. arXiv:2403.02726). arXiv. https://doi.org/10.48550/arXiv.2403.02726