UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

L'Illusion du Jumeau Numérique : Capacités et limites d'un agent LLM dans la simulation des choix de navigation sur YouTube.

Auteur : Cyril Demolin
Promoteur : Corentin Vande Kerckhove
Année académique 2025-2026
Master en Ingénieur de gestion (Business Analytics)

Déclaration relative à l'utilisation d'outils d'intelligence artificielle

Nous reconnaissons que les outils d'IA peuvent constituer une aide précieuse dans le cadre de la rédaction d'un mémoire, mais ils ne sont pas infaillibles. La transparence favorise la confiance, et la reconnaissance du rôle de l'IA renforce la crédibilité du travail. En conséquence, l'utilisation de tout outil d'IA dans ce mémoire a été guidée par les principes suivants : (1) **évaluation critique** des productions générées, toute correction étant effectuée sur la base de l'expertise de l'auteur ; (2) **transparence** quant à l'utilisation de Claude (Anthropic), Gemini et NotebookLM (Google), qui ont contribué au travail sans se substituer au jugement humain ; (3) **responsabilité éthique** dans l'identification des biais potentiels introduits par ces outils.

Dans le cadre de la préparation de ce mémoire, l'auteur a mobilisé ces outils pour les usages suivants :

— **Aide à la rédaction :** structuration de la méthodologie, reformulation de passages complexes et correction linguistique.
— **Revue de littérature :** synthèse et organisation des sources académiques (via NotebookLM).
— **Développement technique :** assistance pour la rédaction et le débogage des scripts de l'infrastructure expérimentale (via Claude).
— **Mise en page Latex :** résolution d'erreurs de compilation et formatage typographique (via Claude et Gemini).

Après chaque utilisation, l'auteur a relu, vérifié et édité le contenu produit. L'auteur assume l'entière responsabilité du résultat final, affirmant que celui-ci reflète un travail intellectuel original augmenté par un usage ciblé et responsable de l'IA.

Date : 27/05/2026 Signature : 

Remerciements

La réalisation de ce mémoire n'aurait pas été possible sans le soutien, l'accompagnement et la bienveillance de nombreuses personnes à qui je souhaite exprimer ma sincère gratitude.

Je remercie en premier lieu mon promoteur, le Professeur Corentin Vande Kerckhove, pour son encadrement rigoureux et sa disponibilité tout au long de ce travail. Ses conseils éclairés et nos échanges réguliers ont été déterminants dans l'aboutissement de cette recherche.

Mes remerciements s'adressent également à Colin Timmers et Quentin Dessain, pour m'avoir accompagné dans la prise en main et l'utilisation du système TRACE, et pour leur disponibilité à chaque étape du déploiement technique.

Je tiens à exprimer ma profonde reconnaissance envers mes parents, dont le soutien indéfectible sous toutes ses formes a rendu possible la réalisation de ce mémoire.

Je remercie également Aymeric et Giorgia, dont le soutien matériel a été indispensable à l'aboutissement de ce travail.

Je tiens enfin à remercier l'ensemble des participants à l'étude expérimentale. Sans leur investissement et leur disponibilité, ce mémoire n'aurait pu voir le jour.

Plus largement, ma gratitude va à l'ensemble de mon entourage proche pour le soutien moral accordé tout au long de cette aventure.

Résumé

L'omniprésence des systèmes de recommandation algorithmiques soulève des préoccupations majeures quant à la formation de « bulles de filtre » et à l'enfermement informationnel des utilisateurs. Pour auditer ces algorithmes, les approches existantes se limitent soit à des *sock puppets* (Haroon, 2023) sans modélisation du processus décisionnel, soit à des *counterfactual bots* Hosseinmardi, 2024 jamais confrontés à un comportement humain réel. Ce mémoire comble ce gap en évaluant empiriquement la fidélité comportementale d'un agent LLM paramétré comme « bot miroir » d'un utilisateur réel sur YouTube.

Une approche expérimentale *within-subject* est déployée auprès de 34 participants, classés selon leur profil idéologique vis-à-vis de l'IA (Pro-IA ou Anti-IA). Leurs trajectoires de navigation récoltées via une interface web reproduisant l'environnement YouTube sont confrontées à celles de leurs jumeaux numériques, animés par le modèle Mistral Small 3.2 via l'infrastructure TRACE. Les deux agents sont exposés aux mêmes listes de recommandations et comparés sur sept dimensions de fidélité comportementale sur cinq étapes de navigation.

L'analyse des 170 observations révèle une dichotomie marquée. D'un côté, le bot miroir reproduit fidèlement les heuristiques ergonomiques de l'utilisateur : il partage un biais de position de même amplitude et sélectionne la même vidéo que l'humain dans 14,7 % des cas, soit significativement au-dessus du hasard pondéré ($\approx$ 7,4 %). L'appartenance idéologique modère la simulation : le bot imite mieux les profils Anti-IA (20,0 %) que Pro-IA (11,4 %), illustrant un effet de caricature propre aux données d'entraînement du LLM. De l'autre, le bot échoue à reproduire la diversité exploratoire humaine : frappé d'un biais de cohérence interne, il s'enferme dans un répertoire de créateurs restreint (UCR = 0,58 contre 0,79 pour l'humain ; $p < 0{,}001$).

Ces résultats conduisent à identifier un *biais de la sonde* : dans les audits algorithmiques reposant sur des LLMs, il devient méthodologiquement impossible de distinguer la bulle de filtre produite par l'algorithme de celle générée par le biais de cohérence du modèle lui-même. Ce mémoire recommande de requalifier l'agent LLM non pas comme un jumeau numérique de l'utilisateur, mais comme un simulateur de scénario extrême, et propose des pistes correctives : hybridation avec du machine learning comportemental, pénalités de répétition et variation dynamique de la température du modèle.

Table des matières

Liste des tableaux

Table des figures

Introduction

Le numérique a apporté, avec son arrivée, une surcharge informationnelle jamais vue auparavant. Ces torrents d'informations continus sont parfois difficiles à représenter, mais Pariser (2011) les introduit par des chiffres très simples : en 2011, 900 000 posts de blogs, 50 millions de tweets et plus de 60 millions de statuts Facebook étaient partagés chaque jour. C'est donc pour se repérer dans cette gigantesque quantité d'informations que les systèmes de recommandation ont été créés, mais pas seulement : de nombreux commerces en ligne ont compris qu'ils constituaient l'outil parfait pour personnaliser l'expérience du consommateur et accroître les ventes (Aggarwal, 2016).

Exemple illustratif Burke et al. (2011)

Jane est une cliente régulière d'une librairie. Lorsqu'elle s'y rend, elle peut apercevoir les recommandations du vendeur placées en vitrine et retrouver toute l'offre disponible, classée par genre et par ordre alphabétique. Mais que se passe-t-il si Jane se rend sur le site internet de cette même librairie possédant un système de recommandation ?
Sur la page d'accueil, Jane reçoit des recommandations sur le nouveau livre de son auteur favori, un thriller à la mode (ses cinq dernières commandes sont des thrillers) et un livre de cuisine végétarienne (Jane a commandé, il y a quelques mois, un livre sur la violence faite aux animaux dans les abattoirs). L'offre est donc bien plus ciblée grâce aux systèmes de recommandation, et le chiffre d'affaires a de grandes chances de croître.

À travers cet exemple, un nouveau problème se pose : le système de recommandation n'opère plus comme un outil de filtrage d'informations, mais comme un outil de manipulation de ce qu'on montre (ou non) à l'utilisateur, afin de maximiser l'efficacité du service proposé. C'est sur cette base qu'Eli Pariser a popularisé le terme « bulle de filtre » (*Filter Bubble*) : une situation dans laquelle la personnalisation algorithmique façonne l'environnement informationnel de l'utilisateur de telle sorte qu'il interagit principalement avec des contenus alignés sur ses croyances préexistantes (Tasenţe, 2025).

Dans ce mémoire, nous prenons comme contexte la formation des bulles de filtre dans le système de recommandation de la plateforme YouTube. YouTube est le service de partage de vidéos en ligne le plus grand et le plus visité au monde. À la fin de ses cinq premières années de service (en 2010), YouTube recevait plus de 2 milliards de vues par jour (Snelson, 2011).

Mais quelle méthodologie utiliser ?

De nombreuses études ont été menées sur la formation des bulles de filtre dans YouTube, et nous pouvons distinguer deux méthodologies, chacune avec ses défauts :

1. La première méthodologie est l'utilisation d'utilisateurs réels. C'est ce que la chercheuse Zeynep Tufekci a choisi d'utiliser lorsqu'elle a visionné des vidéos de différents rassemblements de Donald Trump afin d'étudier la bulle de filtre associée (Bryant, 2020). Cette méthodologie est très intéressante, car elle simule avec précision le comportement d'un utilisateur normal vis-à-vis de l'outil de recommandation de YouTube. Malheureusement, elle est très coûteuse en argent pour recruter des sujets de test et très coûteuse en temps si nous voulons pousser l'algorithme de recommandation dans ses retranchements.
2. La deuxième méthodologie est la simulation d'utilisateurs par un grand modèle de langage. Avec l'essor des LLM, il est possible de simuler les décisions d'un profil d'utilisateur fictif (appelé *persona*) en fonction d'une série de vidéos proposées par l'algorithme de recommandation. Deux doctorants, Quentin Dessain et Colin Timmers, ont développé un outil permettant de faire naviguer dans YouTube de nombreux utilisateurs fictifs contrôlés par un LLM (*Large Language Model*) afin d'étudier l'apparition de bulles de filtre. Cela résout les deux problèmes de la méthodologie précédente, mais pose une question principale : l'utilisation d'un LLM simule-t-elle correctement un utilisateur réel ?

C'est cette question qui nous amène à la raison d'être de ce mémoire :

> *« Dans quelle mesure un agent LLM paramétré sur le profil d'un utilisateur réel reproduit-il fidèlement ses décisions de sélection de vidéos YouTube ? »*

Chapitre 1

Revue de littérature

1.1 Systèmes de recommandation

Avant de s'intéresser précisément à l'analyse de la formation de bulle de filtre dans l'algorithme de recommandation de la plateforme YouTube, il est primordial de s'intéresser à la base théorique sur laquelle tout repose : les systèmes de recommandations.

Les SDR (Systèmes de Recommandation) ont été établis principalement pour répondre à la surcharge informationnelle. Dans l'économie numérique actuelle, l'explosion d'Internet a provoqué une surabondance des données. Cette surcharge informationnelle (*information overload*) est un vrai problème car elle complique la prise de décision de l'utilisateur en dépassant ses capacités cognitives. L'utilité d'un système de recommandation réside dans sa gestion de cette surcharge d'information et de son filtrage servant à obtenir les informations les plus pertinentes possibles selon les préférences ou le comportement de l'utilisateur (Kumar & Sharma, 2016).

Avant de rentrer dans les détails, il est important d'établir une définition simple d'un système de recommandation. Un système de recommandation est un outil qui utilise diverses sources de données, principalement les interactions passées entre les utilisateurs et les articles, pour déduire les intérêts des clients et prédire leurs choix futurs.

Cependant, la valeur ajoutée d'un système de recommandation dépasse aujourd'hui le simple filtrage d'informations : il permet tant aux utilisateurs qu'aux fournisseurs d'optimiser leurs bénéfices respectifs (Aggarwal, 2016, p. 1).

1.1.1 Formalisation et fonctionnement technique

La fonction d'utilité

D'un point de vue théorique, le problème central de la recommandation se définit comme l'estimation d'une fonction d'utilité. Soit C l'ensemble de tous les utilisateurs (*Users*) et S l'ensemble de tous les articles possibles (*Items*) susceptibles d'être recommandés. L'objectif d'un système de recommandation est donc d'établir une fonction d'utilité u qui mesure la pertinence d'un article s pour un utilisateur c (Adomavicius & Tuzhilin, 2005).

Formellement, cette fonction est définie par l'application $u : C \times S \rightarrow \mathbb{R}$, où $\mathbb{R}$ est un ensemble totalement ordonné (par exemple, des entiers non négatifs ou des nombres réels dans un intervalle donné) représentant la valeur de la pertinence prédite pour l'utilisateur

(Adomavicius & Tuzhilin, 2005). L'objectif d'un SDR est donc de sélectionner, pour chaque utilisateur $c \in C$, un article $s' \in S$ qui maximise l'utilité prédite. Cette maximisation s'effectue sur l'ensemble des articles S que l'utilisateur n'a pas encore consommés ou notés (Adomavicius & Tuzhilin, 2005) :

$$\forall c \in C, \quad s'_c = \arg\max_{s \in S} u(c, s) \tag{1.1}$$

Typologie des interactions et modalités de feedback

La qualité et la nature des données concernant les préférences des utilisateurs sont d'une importance capitale pour garantir l'efficacité d'un système de recommandation. Dans la littérature, on distingue clairement deux catégories distinctes de feedback : le feedback explicite et le feedback implicite (Isinkaye et al., 2015).

Le feedback explicite : Cette interaction consiste en une intervention intentionnelle et active de l'utilisateur qui souhaite exprimer son degré d'appréciation envers un article (Isinkaye et al., 2015). Le feedback explicite peut se présenter sous différentes formes : notes scalaires (par exemple, de 1 à 5 étoiles), votes binaires (indicateurs de type « j'aime / je n'aime pas »), classements ordinaux ou commentaires (Aggarwal, 2016, p. 10). Bien que ces données soient considérées comme les plus fiables, car elles sont à l'origine d'une opinion directe et consciente (Isinkaye et al., 2015), la quantité de données collectées est limitée. En effet, ce mode de collecte impose une charge cognitive à l'utilisateur, ce qui réduit sa propension à interagir (Ricci et al., 2010, p. 110).

Le feedback implicite : Ce feedback consiste à déduire la préférence de l'utilisateur de par son comportement naturel, sans sollicitation directe (Isinkaye et al., 2015). YouTube constitue à cet égard un véritable cas d'école : cliquer sur une vidéo, la durée de visionnage ou la consultation du profil d'un créateur sont considérés comme des interactions implicites (Davidson et al., 2010). Ces interactions sont qualifiées d'unaires car elles signalent une action (1) ou une absence d'action (Aggarwal, 2016, p. 11). Si ces données sont moins précises et plus « bruyantes », elles présentent l'avantage d'être extrêmement abondantes et de ne pas interrompre l'expérience utilisateur (Isinkaye et al., 2015).

En pratique, ces deux types de feedback sont souvent combinés au sein de systèmes hybrides. YouTube illustre parfaitement cette approche en exploitant simultanément les signaux explicites (le bouton « J'aime ») et les signaux implicites (le temps de rétention sur une vidéo) pour affiner ses prédictions (Davidson et al., 2010).

La représentation matricielle

La structure de données fondamentale sur laquelle repose tout système de recommandation est la matrice d'interactions (ou matrice de notation), notée $\mathbf{R}$, de dimensions $m \times n$, où chaque

entrée $r_{u,i}$ correspond à la valeur de l'interaction entre l'utilisateur u et l'article i (Kumar & Sharma, 2016). Dans les systèmes de recommandation commerciaux, la densité D de cette matrice est typiquement extrêmement faible (souvent inférieure à 1 %), ce qui rend la matrice **R** très creuse (Kumar & Sharma, 2016). Cette propriété de *sparsity* transforme le problème de la recommandation en un problème mathématique de complétion de matrice (Isinkaye et al., 2015).

1.1.2 Les types de systèmes de recommandation

Dans la littérature, on distingue principalement trois catégories de systèmes de recommandation : le filtrage collaboratif, le filtrage basé sur le contenu et les systèmes basés sur la connaissance. Il existe également des modèles hybrides combinant différentes méthodes, ainsi que des approches basées sur des données graphiques ou contextuelles (Kumar & Sharma, 2016).

Table 1.1 – Comparaison des principaux types de systèmes de recommandation

Type	**Principe**	**Avantages / Limites**
Filtrage collaboratif	Utilise les interactions entre utilisateurs et articles sans tenir compte du contenu	Large couverture ; problème de démarrage à froid
Filtrage basé contenu	Recommande des articles similaires à ceux appréciés par le passé	Indépendant d'autres utilisateurs ; diversité réduite
Basé sur la connaissance	S'appuie sur des règles et une connaissance explicite du domaine	Fonctionne sans historique ; difficile à maintenir
Hybride	Combine plusieurs méthodes	Meilleures performances ; complexité accrue

1.1.3 YouTube et son système de recommandation

Créé en 2005, YouTube est rapidement devenue la plateforme numéro 1 de partage de vidéos en ligne, constituant une communauté virtuelle immense et influente à l'échelle mondiale. Le site internet comptabilise à l'heure actuelle plus d'un milliard de lectures quotidiennes. Face à la surcharge informationnelle à laquelle YouTube fait face, la plateforme s'appuie massivement sur des algorithmes de recommandation pour orienter la navigation des internautes.

Sur YouTube, le contenu est organisé selon une unité structurante fondamentale : la **chaîne** (*channel*). Une chaîne correspond à un compte éditeur (individu, média, institution ou marque)

qui agrège l'ensemble des vidéos publiées sous une identité commune. C'est à l'échelle de la chaîne que l'algorithme de recommandation construit les profils d'intérêt des utilisateurs : regarder plusieurs vidéos d'une même chaîne constitue un signal fort de préférence thématique, qui oriente les recommandations futures vers des contenus similaires (Covington et al., 2016). Dans le cadre de ce mémoire, la chaîne constitue également l'unité de mesure de la diversité de navigation : le taux de chaînes uniques (UCR) quantifie la proportion de créateurs distincts consultés au fil d'une session, permettant de distinguer une navigation exploratoire d'une navigation hyperfocalisée sur un répertoire restreint de sources.

Le système de recommandation de YouTube est un système hybride sophistiqué. Il fonctionne selon un modèle en entonnoir (*Funnel*) composé de deux étapes principales : (1) la **génération de candidats**, où l'algorithme opère un filtrage collaboratif massif passant d'un corpus de plusieurs millions de vidéos à quelques centaines de recommandations potentielles basées sur l'historique et le contexte de l'utilisateur (Covington et al., 2016) ; et (2) l'**établissement du ranking** des candidats, où le système attribue un score individuel à chaque vidéo selon une fonction d'objectif visant à maximiser le *watch time* (Covington et al., 2016).

1.2 Les bulles de filtre

Le concept de bulle de filtres a été popularisé par Eli Pariser. Cela désigne un environnement informationnel personnalisé et isolé, façonné à l'insu de l'utilisateur par des algorithmes de recommandation. La bulle de filtre ne vient pas d'un choix social délibéré mais résulte plutôt d'une structure technologique qui utilise les données comportementales passées de l'utilisateur pour prédire ce que l'utilisateur souhaite voir. Le mécanisme va donc créer un univers unique d'information pour chaque utilisateur et réduire la diversité des contenus proposés (Dahlgren, 2021).

Ce phénomène est perçu comme dangereux car il favorise l'isolement idéologique de l'utilisateur (Tasenţe, 2025). En refermant l'utilisateur dans ce qu'on appelle une « you loop » qui renforce ses croyances préexistantes et évite de le confronter à des points de vue contradictoires, les bulles de filtres peuvent augmenter la polarisation et invisibiliser les problèmes sociétaux complexes mais importants (Dahlgren, 2021). De plus, cet environnement peut favoriser la désinformation et la manipulation via le microciblage (Tasenţe, 2025).

Ce concept a été énormément médiatisé et a donc récolté son lot de critiques au sein de la communauté scientifique. Par exemple, la littérature souligne son manque de preuves empiriques qui pourraient étayer son existence à l'échelle macro-sociétale (Dahlgren, 2021). Deux aspects sont principalement reprochés à la théorie de Pariser :

Le déterminisme technologique : C'est une croyance qui induit que la technologie est une force toute-puissante contrôlant inévitablement le destin des individus. Cette vision est

critiquée car elle signifie que les algorithmes construisent notre environnement médiatique et modifient notre vision du monde sans que nous puissions y échapper, ce qui surestime grandement l'emprise réelle que l'algorithme possède sur l'utilisateur (Dahlgren, 2021).

La vision strictement behaviouriste : L'approche behaviouriste consiste à étudier directement les comportements observables sans prendre en compte la complexité de l'esprit humain. Dans notre cas, un clic sur une vidéo est considéré comme adhérer à l'idéologie de la vidéo, or les chercheurs soulignent qu'un choix numérique ne reflète pas toujours une préférence (Dahlgren, 2021).

La combinaison de ces deux facteurs amène comme conséquence le mythe de l'usager unique. C'est le paradoxe de la théorie de Pariser qui ressort le plus souvent dans la littérature : elle suppose que l'usager est très actif au début (quand il clique et cherche des informations), mais qu'il devient soudainement totalement passif et malléable dès qu'il reçoit les recommandations de l'algorithme (Dahlgren, 2021).

1.3 Méthodes d'audit algorithmique et LLMs comme simulateurs comportementaux

1.3.1 Typologies des méthodes d'audit algorithmique

Dans notre contexte de mémoire, nous nous intéressons particulièrement à une méthode connue sous le nom d'audit algorithmique. À la base, les audits étaient utilisés dans la recherche expérimentale en sciences sociales afin de prouver la discrimination raciale dans le marché de l'emploi (Sandvig et al., 2014). Ils ont évolué à travers les années afin de discerner ce même genre de discrimination dans les algorithmes. Pour relier ce procédé à notre cas de bulle de filtre, c'est le procédé qui consiste à analyser la présence de bulles de filtres dans la plateforme YouTube.

Face à l'omniprésence et à l'opacité des systèmes technologiques (« boîte noire »), l'audit algorithmique s'impose comme un instrument de surveillance et d'évaluation indispensable (Sandvig et al., 2014). Sa nécessité repose sur une triple exigence : il permet d'abord d'identifier les biais intentionnels et les pratiques anticoncurrentielles discrètement encodés par les concepteurs, mais également de mettre en exergue les externalités sociétales néfastes générées involontairement par des logiques d'optimisation de l'attention et du profit. Enfin, sur le plan empirique, l'audit constitue l'unique démarche méthodologique rigoureuse capable d'isoler l'impact causal de l'algorithme des préférences intrinsèques des utilisateurs, condition sine qua non pour évaluer avec précision la véritable part de responsabilité des plateformes dans l'orientation des comportements en ligne.

Dans ce mémoire, nous allons principalement mettre en avant trois méthodes d'audits

algorithmiques, telles que définies par Sandvig et al. (2014).

L'audit par utilisateur réel : Il arrive que certaines études utilisent des utilisateurs réels afin d'auditer au mieux les algorithmes. En effet, utiliser des utilisateurs réels permet de capturer l'expérience réelle de l'utilisateur, le comportement étudié est 100 % similaire à une situation réelle. Celle qui nous concerne tout particulièrement est l'audit conduit par Mozilla appelé RegretsReporter (Mozilla Foundation, 2021). Il s'agissait d'implémenter une extension dans le navigateur des utilisateurs, servant à déclarer les vidéos YouTube problématiques. L'extension transmettait aux chercheurs la vidéo en question mais aussi l'historique qui menait à cette vidéo problématique. Cette expérience a eu lieu dans 91 pays, avec la participation de 37 380 volontaires (Mozilla Foundation, 2021). Cependant, un audit algorithmique par utilisateur réel amène de nombreuses limites : tout d'abord la complexité et le coût de l'échantillonnage, mais aussi la difficulté de prouver si la cause du résultat obtenu est imputable à l'algorithme ou au comportement de l'utilisateur lui-même (Sandvig et al., 2014). Ces limites poussent les chercheurs à s'éloigner des utilisateurs réels lors des expérimentations. C'est ici qu'intervient l'audit par « sock puppet », consistant à auditer en utilisant des programmes informatiques se faisant passer pour des utilisateurs réels.

Au sein de cette méthode, les chercheurs distinguent *sock puppets* entraînés et sans entraînement, une différence fondamentale à prendre en compte.

a) **Les *sock puppets* non entraînés :** Ils consistent simplement à exécuter des actions basées sur des règles prédéfinies et à cliquer aveuglément sur des recommandations (Sandvig et al., 2014). Cela manque souvent de réalisme et ne permet pas de différencier le rôle exclusif de l'algorithme de celui des préférences humaines.

b) **Les *sock puppets* entraînés :** Pour pallier les limites des bots non entraînés, Haroon et al. (2023) ont mis en place une méthodologie de *sock puppets* entraînés à grande échelle (100 000 comptes entraînés à regarder des vidéos correspondant à différentes idéologies politiques), afin que la plateforme leur accorde un profil et des préférences vraisemblables. Hosseinmardi et al. (2024) poussent cette logique encore plus loin en introduisant le concept de « bots contrefactuels » (*counterfactual bots*) : des bots entraînés sur l'historique exact de vrais utilisateurs, permettant d'isoler avec précision l'effet causal de l'algorithme de recommandation.

Cependant, même si les bots sont entraînés selon le profil de l'utilisateur, il est très difficile de reproduire le comportement logique qu'aurait pu avoir un utilisateur réel face aux recommandations (Hosseinmardi et al., 2024). C'est précisément ce gap méthodologique que ce mémoire cherche à combler.

1.3.2 Le gap identifié dans la littérature

Malgré les progrès significatifs réalisés dans les méthodes d'audit algorithmique, une limite fondamentale persiste et n'a pas encore été comblée par la littérature. Haroon et al. (2023) et Hosseinmardi et al. (2024), bien que distincts dans leurs approches, partagent un déficit commun : leurs bots respectifs ne simulent pas le processus de sélection humain. Dans les deux cas, les agents informatiques naviguent selon des règles prédéfinies et cliquent aveuglément sur les recommandations, sans jamais modéliser le moment décisionnel au cours duquel un utilisateur réel évalue, sélectionne ou rejette une vidéo proposée par l'algorithme.

Si Hosseinmardi et al. (2024) ont constitué une avancée notable en permettant d'entraîner les bots sur l'historique d'un véritable utilisateur (isolant ainsi l'effet causal de l'algorithme), cette méthode reste néanmoins limitée par l'absence de tout mécanisme de sélectivité. Le bot contrefactuel reproduit le profil de l'utilisateur, non sa capacité de choix.

Or, c'est précisément dans cet écart que se situe le gap identifié dans ce mémoire : aucune étude à ce jour n'a examiné si un agent LLM, paramétré sur le profil comportemental d'un utilisateur réel, est capable de reproduire fidèlement ses décisions de navigation sur YouTube. Les LLMs, par leur aptitude à raisonner de manière contextuelle et à simuler des préférences à partir d'un profil donné, représentent une piste prometteuse pour combler ce manque, mais leur validité comportementale comme proxies d'utilisateurs réels dans les audits algorithmiques reste à ce jour une question ouverte que ce mémoire se propose d'explorer.

1.3.3 Les LLMs comme simulateurs comportementaux

L'essor des grands modèles de langage a ouvert une nouvelle perspective dans les sciences sociales computationnelles : celle de l'expérimentation *in silico*, c'est-à-dire la substitution partielle de participants humains par des agents artificiels capables de simuler des comportements et des processus décisionnels. Cette littérature révèle une tension fondamentale entre le potentiel démontré de ces outils et les limites structurelles qui encadrent leur validité.

Un potentiel de simulation comportementale démontré

Les travaux de Park et al. (2023) constituent la démonstration la plus ambitieuse de ce que les LLMs peuvent accomplir en tant que simulateurs sociaux. Dans leur environnement virtuel « Smallville », 25 agents propulsés par GPT-4 ont fait émerger des comportements sociaux complexes non programmés à l'avance (organisation spontanée d'événements, diffusion de rumeurs, formation d'amitiés) jugés plus « humains » par des évaluateurs externes que ceux d'humains jouant un rôle de manière ponctuelle. Cette capacité ne se limite pas aux comportements émergents. Aher et al. (2023) ont démontré que des LLMs conditionnés sur des *personas* socio-démographiques précis reproduisent fidèlement les distributions de réponses observées dans des expériences classiques de psychologie et d'économie comportementale, y

compris leur variance interne. Dans le champ économique, Horton (2023) introduit le concept d'*Homo Silicus* pour désigner des agents LLM capables de respecter les lois fondamentales de l'économie dans des scénarios microéconomiques complexes, ouvrant la voie à une expérimentation sans coût éthique ni financier.

Des limites structurelles qui encadrent leur validité

Ces résultats doivent cependant être nuancés par deux critiques majeures. Santurkar et al. (2023) montrent que les LLMs ne sont pas des simulateurs neutres : leurs réponses s'alignent de manière disproportionnée sur des profils occidentaux, diplômés et libéraux modérés, reflétant les biais de leurs données d'entraînement. La méthode consistant à demander au modèle d'« agir comme » un profil démographique précis échoue à corriger ces biais et génère le plus souvent des stéréotypes caricaturaux. Par ailleurs, Perez et al. (2022) ont mis en évidence un phénomène dit de *sycophancy* (flagornerie algorithmique) propre aux modèles entraînés par renforcement humain (RLHF) : face à un désaccord, le modèle tend à modifier sa réponse pour plaire à son interlocuteur plutôt qu'à maintenir une position cohérente, ce qui constitue un écart fondamental avec le comportement humain réel.

Positionnement de notre étude

La question n'est donc pas de savoir si les LLMs peuvent simuler des comportements humains en général (la littérature le confirme partiellement), mais si un agent LLM paramétré sur le profil précis d'un individu réel est capable de reproduire fidèlement ses décisions de navigation sur YouTube, un contexte de choix contraint et séquentiel que cette littérature n'a pas encore exploré.

Les dimensions de fidélité mobilisées dans la littérature

La littérature sur l'évaluation des systèmes de recommandation et sur les audits algorithmiques a progressivement dégagé plusieurs dimensions permettant de mesurer la *fidélité comportementale* d'un agent simulateur. Ces dimensions constituent le socle théorique à partir duquel les métriques opérationnalisées en section 2.6 ont été construites.

La concordance de sélection exacte (*exact hit rate*) mesure la proportion de cas dans lesquels un agent sélectionne le même item qu'un utilisateur de référence (Ricci et al., 2010). Critère le plus exigeant de fidélité, elle ne tolère aucune approximation et s'impose comme la mesure centrale de l'alignement décisionnel.

La sensibilité à l'ordonnancement concerne la capacité d'un agent à sélectionner des items positionnellement proches de ceux choisis par l'humain. Joachims et al. (2017) ont démontré que la position d'un item influence de manière systématique la probabilité qu'il soit sélectionné, ce que la métrique de déviation moyenne de rang (*Mean Rank Deviation*) permet d'opérationnaliser.

Le biais de position désigne la tendance systématique à sélectionner préférentiellement les items situés en haut d'une liste, indépendamment de leur contenu (Joachims et al., 2017). C'est l'une des heuristiques cognitives les plus robustes de la navigation en ligne.

La congruence idéologique est la tendance à sélectionner des contenus alignés avec ses croyances préexistantes (*selective exposure*) (Pariser, 2011). Un agent paramétré sur un profil idéologique donné devrait reproduire ce biais pour être considéré comme un proxy valide de l'utilisateur réel.

La stabilité temporelle examine si l'alignement entre agent et utilisateur se maintient au fil des étapes successives de navigation (Hosseinmardi et al., 2024).

Les propriétés distributionnelles popularité des contenus et diversité thématique des sources complètent ce cadre. Mozilla Foundation (2021) a montré que l'algorithme peut induire des effets d'enfermement mesurables sur ces deux dimensions.

C'est à partir de ce cadre multidimensionnel (Hosseinmardi et al., 2024 ; Joachims et al., 2017 ; Mozilla Foundation, 2021 ; Pariser, 2011 ; Ricci et al., 2010) que les sept dimensions d'analyse opérationnalisées en section 2.6 ont été construites.

1.3.4 Conclusion et question de recherche

En somme, la littérature académique met en évidence une tension méthodologique et épistémique fondamentale. D'une part, les systèmes de recommandation, illustrés par l'architecture en entonnoir de YouTube, pallient efficacement la surcharge informationnelle mais menacent d'isoler l'utilisateur au sein d'une bulle de filtre en exploitant ses interactions implicites. D'autre part, face à l'opacité de ces plateformes, les méthodes d'audit algorithmique ont évolué des *sock puppets* passifs vers des bots contrefactuels sophistiqués, sans pour autant parvenir à modéliser le moment décisionnel et la sélectivité de l'esprit humain lors d'un clic.

Si l'émergence des grands modèles de langage ouvre la voie à une simulation comportementale *in silico* prometteuse, leurs biais structurels de conformisme ou de sycophancie invitent à la prudence quant à leur validité comme proxies parfaits. C'est précisément pour combler ce gap méthodologique, en confrontant le potentiel sémantique des agents artificiels à la complexité des choix humains en environnement contraint, que s'inscrit ce mémoire. Cette démarche se cristallise autour de la problématique suivante :

Question de recherche principale

« Dans quelle mesure un agent LLM paramétré sur le profil d'un utilisateur réel reproduit-il fidèlement ses décisions de sélection de vidéos YouTube ? »

Chapitre 2

Méthodologie

Ce chapitre expose l'ensemble du protocole empirique déployé pour répondre à notre problématique. L'objet de cette partie est de détailler la démarche expérimentale ayant permis de confronter les trajectoires de navigation d'utilisateurs humains réels à celles de leurs doubles virtuels (agents LLM) au sein d'un environnement YouTube contrôlé.

L'enjeu fondamental de ce chapitre est d'assurer la rigueur, la validité interne et la reproductibilité de l'étude. Il s'agit de démontrer comment des concepts théoriques abstraits (fidélité comportementale, bulle de filtre, biais de sélection) ont été traduits en un dispositif de collecte de données robuste, permettant d'isoler l'effet de l'agent décisionnaire (humain versus IA) face à l'algorithme de recommandation.

Afin de guider le lecteur à travers la construction de cette expérimentation, le présent chapitre s'articule autour des axes suivants :

— **Le cadrage conceptuel** : définition des objectifs de recherche et formalisation des hypothèses confirmative et exploratoires (sections 2.1 et 2.2) ;
— **Le dispositif de collecte** : justification du design expérimental retenu et présentation de l'architecture technique des agents simulateurs (sections 2.3 et 2.4) ;
— **Le déroulement opérationnel** : description pas-à-pas de la procédure expérimentale vécue par les participants (section 2.5) ;
— **Le cadre d'évaluation** : formalisation mathématique des dimensions d'analyse et des tests statistiques mobilisés (section 2.6) ;
— **L'échantillonnage** : critères de recrutement des participants et explicitation des limites méthodologiques inhérentes au dispositif (sections 2.7 et 2.8).

2.1 Objectifs de la recherche

Ce mémoire tente de répondre à la question de recherche suivante : *« Dans quelle mesure un agent LLM paramétré sur le profil d'un utilisateur réel reproduit-il fidèlement ses décisions de sélection de vidéos YouTube ? »*

L'objectif principal est d'enrichir la littérature sur les méthodes d'audit algorithmique en testant empiriquement si un LLM peut constituer un proxy valide d'un utilisateur humain dans un contexte de navigation sur YouTube. Comme établi en section 1.3.3, plusieurs dimensions de la fidélité comportementale ont été identifiées dans la littérature comme caractéristiques de

la navigation humaine sur les plateformes de recommandation (Hosseinmardi et al., 2024 ; Joachims et al., 2017 ; Mozilla Foundation, 2021 ; Pariser, 2011 ; Ricci et al., 2010). Ce mémoire visera plus précisément à les opérationnaliser et à les tester empiriquement à travers les dimensions suivantes :

— La **concordance exacte** : à quel point le bot reproduit les mêmes choix que l'utilisateur ;

— L'**écart de rang** dans les vidéos sélectionnées ;

— Le **biais de sélection de position** dans la liste de recommandation (Joachims et al., 2017) ;

— La **congénialité idéologique** dans les contenus choisis (Pariser, 2011) ;

— L'**évolution de la concordance en profondeur** de la navigation ;

— La **popularité et la diversité** des chaînes sélectionnées (Mozilla Foundation, 2021).

2.2 Hypothèses à tester

S'inscrivant dans une démarche hypothético-déductive, la présente recherche vise à confronter les postulats théoriques issus de notre revue de littérature à l'épreuve des données empiriques.

H_0 (Hypothèse nulle) : Le taux de concordance exacte entre les décisions de clic des participants humains et celles de leurs bots miroirs est équivalent au taux de concordance attendu sous un modèle de sélection purement aléatoire, soit $C_{\text{aléa}} = \frac{1}{N} \sum_i \frac{1}{K_i}$ où K_i est le nombre de recommandations effectivement présentées à l'étape i. Le bot ne sélectionne pas la même vidéo que l'humain plus fréquemment que ne le prédirait le hasard.

H_1 (Hypothèse alternative principale) : Le taux de concordance exacte entre les décisions de navigation des participants humains et celles de leurs bots miroirs est significativement supérieur au taux de concordance aléatoire pondéré ($C_{\text{aléa}}$), attestant d'un alignement partiel entre les logiques de sélection du bot et celles de l'utilisateur réel. Afin d'éprouver la robustesse de cette hypothèse, la concordance du bot est également comparée à deux baselines non triviales : une baseline *always-top* (C_{top}), correspondant à un agent sélectionnant systématiquement la vidéo en première position, et une baseline *most-popular* (C_{pop}), correspondant à un agent sélectionnant systématiquement la vidéo la plus regardée parmi celles présentées.

La présente étude adopte un design inférentiel à deux niveaux, conformément aux recommandations de Lakens et al. (2018) sur la nécessité de distinguer explicitement les analyses confirmatives des analyses exploratoires. L'hypothèse H_{1a} constitue l'unique hypothèse confirmative de cette recherche. Les hypothèses H_{1b} à H_{1f} sont traitées comme des analyses

exploratoires complémentaires.

H_{1a} : Concordance exacte (Hypothèse confirmative principale) : Le taux de concordance exacte, défini comme la propension du bot à sélectionner rigoureusement les mêmes recommandations algorithmiques que l'utilisateur humain, est significativement supérieur à celui d'une sélection aléatoire.

H_{1b} : Minimisation de l'écart de rang (exploratoire) : L'écart de rang moyen entre la vidéo sélectionnée par le participant et celle choisie par le bot est significativement inférieur à ce qu'un choix purement aléatoire produirait.

H_{1c} : Reproduction du biais de position (exploratoire) : Le bot miroir reproduit l'heuristique humaine consistant à sélectionner préférentiellement les vidéos situées en haut de la liste de recommandations (Joachims et al., 2017).

H_{1d} : Congénialité idéologique (exploratoire) : Le taux de concordance entre bot et humain diffère significativement entre le groupe Pro-IA et le groupe Anti-IA, indiquant que l'appartenance au groupe idéologique module la capacité du bot à simuler les décisions de navigation de l'utilisateur (Pariser, 2011).

H_{1e} : Stabilité temporelle et profondeur de navigation (exploratoire) : L'alignement entre les choix du bot et ceux du participant ne subit pas de dégradation significative à mesure que l'exploration s'allonge.

H_{1f} : Isomorphisme de popularité et de diversité (exploratoire) : Le bot miroir reproduit les schémas de l'utilisateur réel concernant la popularité des contenus et la diversité des chaînes consultées (Mozilla Foundation, 2021).

2.3 Design expérimental

Afin de comparer les décisions de navigation d'un utilisateur réel et celles d'un agent simulateur, cette étude adopte un design *within-subject* plutôt que *between-subject*.

Dans un design *within-subject*, chaque participant est exposé aux deux conditions expérimentales : il constitue à la fois sa propre condition humaine et sa propre condition de référence pour l'agent simulateur qui lui est associé. Chaque paire participant/agent possède ainsi sa propre baseline individuelle, ce qui permet de contrôler les différences interindividuelles et de faciliter les comparaisons statistiques. À l'inverse, un design *between-subject* comparerait deux groupes entièrement distincts, introduisant une variabilité inter-individuelle non contrôlée qui réduirait la puissance des tests pour une même taille d'échantillon.

Nous avons retenu le design *within-subject* pour son efficacité statistique : il réduit l'erreur résiduelle et augmente la puissance du test pour une taille d'échantillon donnée (Lakens, 2022 ; Maxwell, 2013). La nature concrète du couplage entre chaque participant et son agent simulateur est décrite en section 2.4.

2.4 Construction des agents LLM miroirs

2.4.1 Principe du bot miroir

Chaque participant se voit associer un agent LLM miroir paramétré sur son profil individuel. Ce paramétrage prend la forme d'une instruction système (*system prompt*) transmise au modèle de langage avant le début de chaque session de navigation. Cette instruction décrit le profil attitudinal du participant, son appartenance au groupe pro-IA ou anti-IA, ses centres d'intérêt déclarés vis-à-vis des technologies algorithmiques, ainsi que sa disposition générale telle que renseignée dans le formulaire de recrutement.

Afin de garantir la comparabilité des trajectoires entre participants d'un même groupe, les conditions initiales de navigation sont standardisées. Avant d'entrer dans la phase d'observation libre, chaque bot visionne séquentiellement trois vidéos de contexte prédéfinies par les chercheurs, identiques pour l'ensemble des membres d'un même groupe. Ces vidéos sont thématiquement cohérentes avec l'appartenance au groupe : elles portent sur des contenus pro-IA pour le premier groupe, et sur des contenus critiques à l'égard de l'intelligence artificielle pour le second.

Durant la phase de navigation autonome, l'intégralité des données produites par le bot miroir est enregistrée en base de données à chaque étape : les vidéos recommandées par l'algorithme de YouTube, leur rang d'apparition dans la liste de suggestions, ainsi que la vidéo effectivement sélectionnée par l'agent LLM. Ces données constituent la trajectoire de référence du bot miroir.

2.4.2 Architecture technique

Le système de collecte de données repose sur l'infrastructure **TRACE**, développée par Quentin Dessain et Colin Timmers dans le cadre de leurs travaux doctoraux à l'UCLouvain Timmers et Dessain (2024). Cette infrastructure entièrement conteneurisée, orchestrée via Docker Compose, garantit la reproductibilité des expériences, l'isolation de chaque instance de bot et la portabilité de l'ensemble du dispositif. Elle constitue la colonne vertébrale technique sur laquelle le présent dispositif expérimental a été déployé.(Voir Annxe E)

L'infrastructure TRACE se compose de quatre types de services distincts :

1. **Base de données PostgreSQL :** stocke l'ensemble des données produites au fil des

sessions (profils de personas, vidéos de contexte, sessions de navigation, recommandations).

2. **Selenium Grid :** composé d'un nœud central (*hub*) et de plusieurs nœuds *workers*, chacun hébergeant une instance isolée de navigateur Chromium. C'est via ce dispositif que les bots interagissent concrètement avec la plateforme YouTube.
3. **Scraper :** cœur logique du bot miroir. À chaque étape de navigation, il extrait la liste des vidéos recommandées et la soumet au modèle de langage pour décision.
4. **Worker d'enrichissement :** s'exécute en tâche de fond en parallèle des sessions de scraping. Il interroge l'API YouTube Data v3 afin de récupérer les métadonnées complètes de chaque vidéo référencée.

La contribution propre à ce mémoire réside dans l'adaptation et le déploiement de cette infrastructure à des fins de comparaison within-subject entre comportements humains et comportements simulés. Plus précisément, les éléments suivants ont été conçus et implémentés dans le cadre de cette recherche :

— le **paramétrage des personas** : construction des prompts système individualisés pour chacun des 34 participants, intégrant profil attitudinal, appartenance idéologique, centres d'intérêt et historique de visionnage déclaré (Voir annexe B) ;

— l'**intégration du modèle Mistral Small 3.2** via l'API OpenRouter comme moteur décisionnel du scraper, en remplacement d'un comportement de clic aveugle ;

— la **sélection et standardisation des vidéos de contexte** pour chaque groupe idéologique (Voir annexe C) ;

— l'**interface HTML** répliquant l'environnement visuel de YouTube, présentée aux participants humains lors de la phase expérimentale (Voir Annexe D) ;

— un **script d'enrichissement complémentaire** exécuté *a posteriori*, qui a interrogé l'API YouTube Data v3 pour chaque identifiant de vidéo sélectionné par le bot, permettant de récupérer les noms exacts des chaînes et le nombre de vues nécessaires au calcul des scores de diversité (UCR) et de popularité.

2.4.3 Paramètres de simulation

Table 2.1 – Paramètres techniques de simulation des bots miroirs

Paramètre	Valeur	Justification
Profondeur maximale (MAX_DEPTH)	10 étapes	Trajectoires suffisamment longues
Temps de visionnage maximal	120 secondes	Comportement réaliste d'un utilisateur

2.5 Procédure expérimentale

Chaque session expérimentale se déroule en quatre étapes, pour une durée totale d'environ 35 minutes par participant.

Étape 1 : Formulaire de profilage (10 minutes) Avant toute interaction avec le système de navigation, le participant complète un formulaire de profilage dont les données servent à paramétrer son bot miroir. Ce formulaire recueille les données socio-démographiques, les habitudes de consommation YouTube et l'attitude vis-à-vis de l'intelligence artificielle, mesurée à l'aide d'une échelle de Likert en cinq points. (Voir annexe A)

Étape 2 : Navigation autonome préalable du bot miroir Avant la session du participant, le bot miroir paramétré sur son profil conduit de manière autonome sa propre session de navigation YouTube. L'intégralité de cette trajectoire est enregistrée en base de données : vidéos recommandées à chaque étape, rang de chaque suggestion, vidéo sélectionnée et justification textuelle produite par le modèle de langage.

Étape 3 : Navigation du participant via l'interface web (20 minutes) Le participant est invité à interagir avec une interface web développée spécifiquement pour l'étude. La session débute par la présentation des trois vidéos de contexte standardisées correspondant au groupe du participant. L'interface lui présente ensuite, étape par étape, les recommandations que l'algorithme de YouTube a soumises au bot miroir. Pour chaque étape, le participant visualise un ensemble de vidéos recommandées affichées dans un format fidèle à celui de la plateforme originale : miniature, titre, nom de la chaîne, durée et nombre de vues. Le participant est invité à sélectionner la vidéo qu'il aurait personnellement choisie. Ce processus est répété sur **cinq étapes successives**, alors que le bot miroir a navigué sur un maximum de dix étapes lors de sa session autonome. Cette limitation à cinq étapes pour le participant répond à deux considérations : d'une part, contenir la durée de la session dans un format compatible avec les contraintes pratiques du recrutement (environ 20 minutes) ; d'autre part, limiter la potentielle dégradation de la qualité des réponses liée à la fatigue décisionnelle, susceptible d'introduire

un biais croissant au fil des étapes. Les cinq premières étapes de navigation du bot constituent donc la trajectoire de référence utilisée pour la comparaison.

Étape 4 : Débriefing (5 minutes) La session se conclut par un entretien de débriefing. Le chercheur révèle au participant la nature exacte de l'étude, en expliquant le principe du bot miroir et l'objectif de comparaison entre décisions humaines et décisions algorithmiques.

2.6 Dimensions d'analyse, métriques et tests statistiques

Afin de répondre de manière rigoureuse à la question de recherche, sept dimensions d'analyse ont été retenues. Conformément au cadre multidimensionnel de fidélité comportementale établi en section 1.3.3 (Hosseinmardi et al., 2024 ; Joachims et al., 2017 ; Mozilla Foundation, 2021 ; Pariser, 2011 ; Ricci et al., 2010), chacune de ces dimensions correspond à une propriété de la navigation humaine identifiée dans la littérature, et dont la reproduction par un agent LLM constitue un test empirique de la qualité de la simulation.

2.6.1 Concordance exacte

Définition opérationnelle. La concordance exacte mesure la proportion d'étapes de navigation lors desquelles le bot miroir sélectionne la même vidéo que le participant humain. Pour un participant p impliqué dans une session de N étapes :

$$C_p = \frac{1}{N} \cdot \sum_{i=1}^{N} \mathbb{1}[v_{H,i} = v_{B,i}] \tag{2.1}$$

où $v_{H,i}$ désigne la vidéo sélectionnée par l'humain à l'étape i, $v_{B,i}$ celle sélectionnée par le bot, et $\mathbb{1}[\cdot]$ la fonction indicatrice. Afin d'évaluer rigoureusement la valeur ajoutée du paramétrage par persona, la concordance du bot est comparée à trois baselines de référence :

— $C_{\text{aléa}} = \frac{1}{N} \sum_i \frac{1}{K_i}$: baseline aléatoire pondérée, correspondant à la probabilité de coïncidence sous un modèle de sélection uniforme ;

— C_{top} : baseline *always-top*, taux de concordance d'un agent sélectionnant systématiquement la vidéo en position 1 capitalisant ainsi sur le biais de position (Joachims et al., 2017) ;

— C_{pop} : baseline *most-popular*, taux de concordance d'un agent sélectionnant systématiquement la vidéo présentant le plus grand nombre de vues à chaque étape.

Tests statistiques. Cette dimension est éprouvée par le **Kappa de Cohen (κ)**, qui mesure l'accord entre les sélections humaines et celles du bot en corrigeant la part de concordance attribuable au hasard. Ce test a été retenu pour trois raisons. Premièrement, la nature de la

variable à analyser est nominale et binaire, ce pour quoi le Kappa de Cohen est précisément conçu. Deuxièmement, à la différence d'un simple taux de concordance brute C_p, le Kappa de Cohen corrige explicitement la concordance observée par la concordance attendue sous l'hypothèse d'indépendance totale, ce qui est essentiel dans notre cas. Troisièmement, le Kappa offre une interprétation standardisée de la force de l'accord (Landis & Koch, 1977) qui facilite la comparaison entre participants et entre groupes idéologiques. La comparaison aux trois baselines est effectuée via un test binomial unilatéral.

Lien avec les hypothèses. Cette dimension opérationnalise directement H_{1a}.

2.6.2 Écart de rang

Définition opérationnelle. La métrique retenue est la déviation moyenne de rang (*Mean Rank Deviation*, MRD) :

$$\mathrm{MRD}_p = \frac{1}{N} \cdot \sum_{i=1}^{N} |r_{H,i} - r_{B,i}| \tag{2.2}$$

La baseline aléatoire théorique s'établit à $\mathrm{MRD}_{\text{aléa}} \approx (K+1)/3$ pour une liste de K éléments.

Tests statistiques. Cette dimension est éprouvée par un **test de Wilcoxon signé-rank unilatéral** (*one-sample, one-tailed*), comparant le MRD observé à la baseline aléatoire théorique $(K+1)/3$.

Lien avec les hypothèses. Cette dimension opérationnalise H_{1b}.

2.6.3 Biais de position

Définition opérationnelle. Pour chaque condition, nous calculons le rang moyen des vidéos sélectionnées :

$$\mathrm{MR}_H = \frac{1}{N} \cdot \sum_{i} r_{H,i} \qquad \mathrm{MR}_B = \frac{1}{N} \cdot \sum_{i} r_{B,i} \tag{2.3}$$

Tests statistiques. Deux tests sont mobilisés conjointement : un **test de Wilcoxon signé-rank unilatéral** (*one-sample*) pour chaque condition, et un **test de Wilcoxon apparié** (*Paired Wilcoxon*) pour comparer les distributions de rangs humains et de rangs du bot.

Lien avec les hypothèses. Cette dimension opérationnalise H_{1c}.

2.6.4 Congénialité idéologique

Définition opérationnelle. Pour chaque vidéo v_i sélectionnée, un score de congénialité binaire $\mathrm{cong}(v_i)$ est attribué. La métrique agrégée est :

$$\mathrm{ICS}_p = \frac{1}{N} \cdot \sum_i \mathrm{cong}(v_i) \tag{2.4}$$

Tests statistiques. Trois tests sont appliqués de façon complémentaire : (1) un **test de Wilcoxon signé-rank unilatéral** (*one-sample, one-tailed*) ; (2) un **test de Wilcoxon apparié bilatéral** ; et (3) un **test de McNemar** sur la table de contingence des choix congéniaux et non congéniaux.

Lien avec les hypothèses. Cette dimension opérationnalise $\mathrm{H}_{1\mathrm{d}}$.

2.6.5 Évolution de la concordance en profondeur de navigation

Définition opérationnelle. Pour chaque profondeur d (de 1 à $D = 5$) :

$$C(d) = \frac{1}{P} \cdot \sum_{p=1}^{P} \mathbb{1}[v_{H,p,d} = v_{B,p,d}] \tag{2.5}$$

Tests statistiques. L'évolution de $C(d)$ est éprouvée par une **corrélation de Spearman** entre la profondeur de navigation et le taux de concordance observé.

Lien avec les hypothèses. Cette dimension opérationnalise $\mathrm{H}_{1\mathrm{e}}$.

2.6.6 Popularité des vidéos sélectionnées (dimension exploratoire)

Définition opérationnelle.

$$\mathrm{MP}_p = \frac{1}{N} \cdot \sum_i \log_{10}(\mathrm{vues}_i + 1) \tag{2.6}$$

Tests statistiques. Un **test de Wilcoxon apparié bilatéral** compare les distributions de MP_H et MP_B par participant.

Lien avec les hypothèses. Cette dimension contribue à l'évaluation de $\mathrm{H}_{1\mathrm{f}}$ dans sa composante popularité.

2.6.7 Diversité des chaînes consultées (dimension exploratoire)

Définition opérationnelle. Le taux de chaînes uniques (*Unique Channel Ratio*, UCR) :

$$\mathrm{UCR}_p = \frac{|\{\text{chaînes distinctes sélectionnées par } p\}|}{N} \tag{2.7}$$

Tests statistiques. Un **test de Wilcoxon apparié bilatéral** compare les distributions de UCR_H et UCR_B par participant.

Lien avec les hypothèses. Cette dimension contribue à l'évaluation de $\mathrm{H}_{1\mathrm{f}}$ dans sa composante

diversité.

Table 2.2 – Récapitulatif des dimensions, métriques et tests statistiques

Hyp.	Dimension	Métrique	Test(s)
H_{1a}	Concordance exacte	C_p	Kappa de Cohen
H_{1b}	Écart de rang	MRD	Wilcoxon (one-sample)
H_{1c}	Biais de position	MR	Wilcoxon (one-sample + apparié)
H_{1d}	Congénialité idéologique	ICS	Wilcoxon + McNemar
H_{1e}	Stabilité temporelle	CRD	Spearman
H_{1f}	Popularité	MP	Wilcoxon (apparié)
H_{1f}	Diversité chaînes	UCR	Wilcoxon (apparié)

2.7 Participants

2.7.1 Critères de sélection et recrutement

La constitution de l'échantillon a obéi à une logique de recrutement raisonné, visant à maximiser la pertinence théorique des profils tout en restant réaliste quant aux contraintes pratiques d'un mémoire de master.

Critères d'inclusion et d'exclusion. Trois critères d'inclusion cumulatifs ont été appliqués : être utilisateur régulier de YouTube (fréquence minimale de quelques fois par semaine), disposer d'un accès à un ordinateur (la phase expérimentale étant incompatible avec une utilisation mobile) et être locuteur francophone, les bots miroirs ayant été paramétrés sur des profils de navigation principalement en français. Aucun critère d'exclusion lié à l'âge, au genre ou au niveau d'études n'a été appliqué a priori.

Constitution des groupes. Le profil idéologique de chaque participant vis-à-vis de l'intelligence artificielle a été opérationnalisé via trois items de type Likert en cinq points, portant sur l'enthousiasme envers l'IA, la perception de ses risques sociétaux, et l'évaluation comparée de ses bénéfices et de ses risques. La moyenne de ces trois items a produit un indice allant de 1 à 5 : les participants obtenant un score supérieur à 3,33 ont été classés **Pro-IA**, ceux en dessous de 2,67 **Anti-IA**. Pour les profils neutres, un score secondaire de confiance envers l'algorithme YouTube a servi de critère de départage.

Recrutement. Les participants ont été recrutés par convenance via deux canaux : le réseau universitaire immédiat du chercheur et la plateforme LinkedIn. Les sessions se sont tenues en présentiel ou en visioconférence selon la disponibilité des participants.

2.7.2 Taille d'échantillon et test de puissance

La taille d'échantillon cible a été déterminée a priori à l'aide du logiciel G*Power 3.1 (Faul et al., 2007). Le test de référence retenu est le test de Wilcoxon signé-rank apparié, correspondant au design *within-subject* de l'étude.

Les paramètres retenus sont : une taille d'effet conventionnelle moyenne de $d = 0{,}50$ (Cohen, 1988) ; un seuil de significativité bilatéral de $\alpha = 0{,}05$; et une puissance statistique de $1 - \beta = 0{,}80$. Ces paramètres conduisent à une taille d'échantillon minimale requise de $n = 34$ participants. (Voir annexe F)

L'échantillon effectivement collecté est de 34 participants, atteignant ainsi la taille cible calculée a priori.

2.7.3 Limites méthodologiques

Taille d'échantillon sous-optimale pour certaines analyses. Bien que la taille cible de $n = 34$ soit atteinte, elle reste relativement modeste pour certaines analyses en sous-groupes (Pro-IA vs Anti-IA), ce qui réduit la puissance des tests exploratoires H_{1b} à H_{1f}.

Asymétrie temporelle entre bots et humains. Les sessions de navigation des bots miroirs ont été collectées à des moments distincts de celles des participants humains. Or, l'algorithme de recommandation de YouTube est connu pour sa sensibilité au contexte temporel. Cette asymétrie temporelle introduit une source de variance parasite susceptible de minorer artificiellement les taux de concordance observés.

Artificialité de la session expérimentale. Les participants naviguaient sur une interface HTML simulant YouTube, à partir d'un ensemble fixe de recommandations pré-collectées. Cette configuration garantit la comparabilité des conditions mais au prix d'un certain degré d'artificialité : le participant ne peut pas réellement regarder les vidéos et les recommandations sont figées.

Dépendance à un unique modèle LLM. L'ensemble des bots miroirs ont été implémentés à partir d'un seul modèle de langage, Mistral Small 3.2. Cette dépendance soulève la question de la généralisabilité des résultats, qui pourraient différer sensiblement avec d'autres architectures, notamment des modèles plus puissants tels que GPT-4o ou Claude Opus.

Choix unique par étape et dépendance à l'appariement strict. Le protocole expérimental contraint chaque agent, participant humain comme bot miroir, à formuler un unique choix par étape de navigation. Cette structure binaire rend H_{1a} entièrement dépendante d'un appariement strict : une coïncidence exacte est requise pour qu'une concordance soit enregistrée, sans qu'il soit possible de comparer les distributions de préférence des deux agents sur l'ensemble des recommandations présentées. Un dispositif autorisant plusieurs choix ordonnés par étape

(*top-K preferences*) aurait permis une comparaison plus fine et plus nuancée de l'alignement entre humain et bot, en dépassant la logique du tout-ou-rien inhérente à la concordance exacte. Cette contrainte de design constitue une piste d'amélioration prioritaire pour les travaux ultérieurs.

2.8 Conclusion du chapitre

En conclusion, ce chapitre a permis de poser les fondations empiriques indispensables à notre recherche. En traduisant nos questionnements théoriques en un protocole expérimental concret et reproductible, nous avons construit un dispositif capable d'évaluer avec précision la fidélité comportementale de notre agent simulateur.

Le choix d'un design *within-subject*, couplé à l'infrastructure technique TRACE, garantit que les choix de l'utilisateur humain et ceux de son bot miroir (LLM) sont comparés de manière stricte, face aux mêmes recommandations et dans un environnement standardisé. Par ailleurs, la formalisation de nos sept dimensions d'analyse , allant de la concordance exacte à la diversité des chaînes consultées , nous dote d'un prisme d'évaluation suffisamment riche pour capturer la complexité des comportements de navigation.

Bien que ce protocole présente certaines limites assumées, notamment liées à l'artificialité de l'interface de simulation et à la taille modeste de l'échantillon pour les analyses en sous-groupes, il fournit un cadre d'observation robuste pour tester notre hypothèse confirmative et explorer nos hypothèses secondaires.

Le dispositif de collecte et les outils statistiques (tests non paramétriques, Kappa de Cohen) étant désormais solidement établis, le chapitre suivant sera consacré à la présentation, à l'analyse et à l'interprétation des données générées par ces 34 sessions croisées de navigation.

Chapitre 3

Résultats

Ce chapitre présente l'ensemble des analyses empiriques issues des données collectées au cours de l'expérimentation. L'enjeu principal est de soumettre nos hypothèses théoriques à l'épreuve des faits afin de mesurer avec précision le degré d'isomorphisme entre les décisions humaines et les simulations de notre agent LLM. Conformément au cadre inférentiel explicité au chapitre précédent, nous distinguons rigoureusement l'analyse confirmative (portant sur l'hypothèse principale de concordance exacte H_{1a}) des analyses exploratoires complémentaires (H_{1b} à H_{1f}).

Pour offrir une lecture transparente et progressive des données, ce chapitre est articulé en deux grandes parties :

— **Une analyse descriptive de l'échantillon** (section 3.1), visant à dresser de manière synthétique le profil de nos participants et à valider la robustesse de la segmentation entre les groupes idéologiques Pro-IA et Anti-IA.

— **Une évaluation par dimension de fidélité** (section 3.2), qui passe au crible les 170 observations de clics séquentielles à l'aide des tests non paramétriques (Wilcoxon, Kappa de Cohen, Spearman, Mann-Whitney U) définis dans notre méthodologie.

Un tableau récapitulatif global (section 3.3) vient clôturer ce chapitre pour synthétiser les verdicts statistiques associés à chaque hypothèse.

3.1 Description de l'échantillon

L'échantillon final est composé de **34 participants** ($n = 34$), atteignant ainsi la taille cible calculée a priori par le test de puissance (section 2.8.2). Au total, 170 observations ont été collectées (34 participants $\times$ 5 étapes de navigation), constituant la base de données sur laquelle reposent l'ensemble des analyses statistiques présentées dans ce chapitre.

Profil sociodémographique et habitudes de navigation

Afin de concentrer notre analyse sur les variables directement utiles à nos hypothèses, les données descriptives globales sont présentées de manière synthétique. L'échantillon est majoritairement composé de jeunes adultes étudiants (âge moyen 25,3 ans, $SD = 8{,}9$; 55,9 % de femmes et 44,1 % d'hommes), dotés d'un niveau d'études supérieur (91,2 % détiennent un diplôme de bachelier ou de master).

En matière de consommation numérique, les participants sont des utilisateurs réguliers de YouTube : 82,3 % d'entre eux visionnent des vidéos au moins quelques fois par semaine, principalement lors de sessions de courte à moyenne durée (entre 15 et 60 minutes). Leurs intérêts thématiques se portent massivement sur le divertissement et l'humour (76,5 %), la musique (41,2 %) et l'actualité (35,3 %), tandis que les contenus scientifiques et technologiques intéressent un quart de l'échantillon (26,5 %). Par ailleurs, les participants accordent une confiance globalement favorable à l'algorithme de recommandation de la plateforme, évaluée à 3,74 sur 5 en moyenne.

Répartition par groupe idéologique vis-à-vis de l'IA

La variable différenciatrice centrale de notre étude est le profil idéologique vis-à-vis de l'intelligence artificielle, qui sert de socle à l'hypothèse exploratoire sur la congénialité idéologique (H_{1d}). L'application du protocole de classification décrit en méthodologie a conduit à répartir les participants en deux groupes distincts :

Table 3.1 – Répartition des participants par groupe idéologique

Groupe	**n**	**%**	**Score IA moyen (M ± SD)**
Pro-IA	21	61,8 %	3,44 ± 0,37
Anti-IA	13	38,2 %	2,33 ± 0,30
Total	**34**	**100 %**	n.a.

La différence de score IA composite moyen entre les deux groupes est substantielle (Δ = 1,11 point sur 5 sur l'échelle de Likert), attestant d'une bonne séparation idéologique. Ce clivage est porté par des différences d'attitude marquées sur les items sous-jacents, comme le détaille le tableau 3.2.

Table 3.2 – Scores aux items IA par groupe idéologique (M ± SD)

Item	**Pro-IA** (n = 21)	**Anti-IA** (n = 13)
Q1 : Enthousiasme	4,10 ± 0,83	2,77 ± 0,73
Q2 : Perception du risque (inv.)	2,62 ± 1,20	1,92 ± 1,12
Q4 : Bénéfices > Risques	3,62 ± 0,86	2,31 ± 0,85
Score IA composite	**3,44 ± 0,37**	**2,33 ± 0,30**

Il est intéressant de noter que même au sein du groupe Pro-IA, la perception du risque reste relativement présente (score inversé de Q2 : M = 2,62), suggérant que l'enthousiasme envers l'IA n'est pas synonyme d'une confiance aveugle. Le groupe Anti-IA se distingue, quant à lui,

par un faible enthousiasme (Q1 : $M = 2{,}77$) et la conviction forte que les risques liés à l'IA dépassent ses potentiels bénéfices (Q4 : $M = 2{,}31$). Ces deux profils tranchés garantissent une base empirique solide pour tester si l'orientation idéologique influence la fidélité de simulation du bot.

3.2 Résultats par dimension

3.2.1 Concordance exacte

La concordance exacte désigne la proportion d'étapes où le bot sélectionne rigoureusement la même vidéo que l'utilisateur humain. Compte tenu de la variance du nombre de recommandations par étape (K_i), la *baseline* de sélection purement aléatoire s'établit dynamiquement à $C_{\text{aléa}} = 7{,}86\,\%$ (voir annexe G). Sur les 170 observations, on dénombre **25 concordances exactes**, soit un **taux brut** $C_p = 14{,}7\,\%$. Le test binomial confirme que cette supériorité face au hasard pondéré est statistiquement significative ($p = 0{,}002$).

Toutefois, la confrontation à deux baselines non triviales nuance cette performance. L'heuristique *always-top* (sélection systématique du premier rang) atteint $C_{\text{top}} = 14{,}1\,\%$, un score qui n'est pas significativement différent de celui du bot ($p = 0{,}446$). À l'inverse, l'agent surpasse largement la baseline *most-popular* ($C_{\text{pop}} = 2{,}35\,\%$, $p < 0{,}001$), bien que ce résultat reflète surtout le désintérêt systématique des participants réels pour la vidéo la plus vue de la liste.

Enfin, le **Kappa de Cohen pondéré** appuie cette réserve. Avec $\boldsymbol{\kappa = 0{,}050}$ (accord *très faible* selon Landis et Koch (1977)), le bot et l'humain partagent davantage une logique spatiale similaire (la tendance à cliquer en haut de l'interface) qu'une réelle convergence thématique. **H_{1a} est donc partiellement supportée** : la concordance de l'agent LLM est significativement supérieure au hasard, mais sa performance s'explique presque entièrement par un biais de position, l'empêchant de surpasser une heuristique spatiale basique.

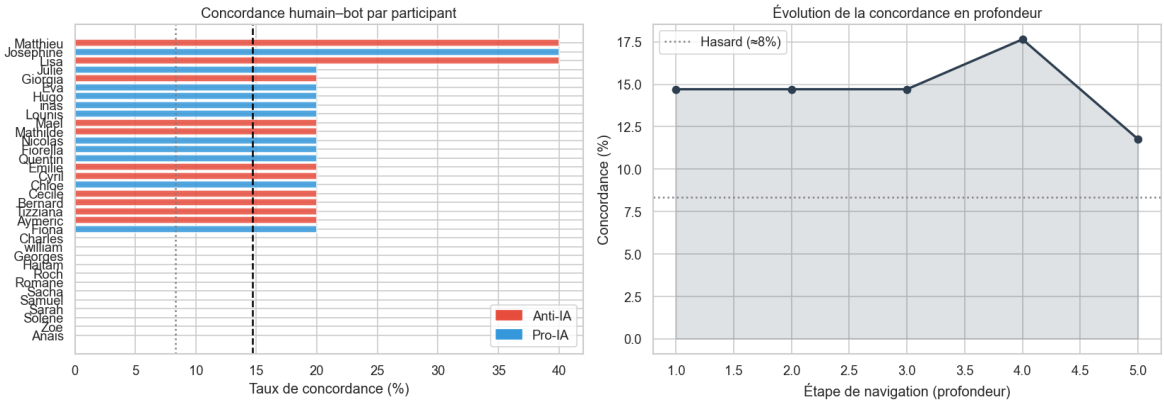

Figure 3.1 – Taux de concordance observé ($C_p = 14{,}7\,\%$) comparé aux trois baselines de référence : aléatoire pondérée ($C_{\text{aléa}} = 7{,}86\,\%$), *always-top* ($C_{\text{top}} = 14{,}1\,\%$) et *most-popular* ($C_{\text{pop}} = 0{,}0\,\%$). Les participants sont colorés selon le groupe idéologique (bleu : Pro-IA, rouge : Anti-IA).

3.2.2 Écart de rang

Cette dimension examine si, au-delà de la coïncidence stricte de choix, le bot et l'humain sélectionnent des vidéos positionnées de manière similaire dans la liste de recommandations. La métrique mobilisée est la déviation moyenne de rang (*Mean Rank Deviation*, MRD), calculée comme la valeur absolue de la différence de position entre la vidéo choisie par l'humain et celle choisie par le bot à chaque étape.

Étant donné que le nombre de recommandations présentées varie à chaque étape, la baseline aléatoire est calculée de manière pondérée :

$$\mathrm{MRD}_{\text{aléa}} = \frac{1}{N} \sum_{i} \frac{K_i + 1}{3} = 5{,}46 \tag{3.1}$$

où K_i désigne le nombre de recommandations effectivement disponibles à l'étape i.

L'écart de rang moyen observé dans notre échantillon est de MRD = 4,44, soit un écart inférieur à la baseline aléatoire pondérée. Pour tester formellement cette supériorité, un test des rangs signés de Wilcoxon est appliqué sur les résidus step-wise, définis comme la différence entre l'écart observé et la baseline propre à chaque étape ($\mathrm{gap}_i - (K_i + 1)/3$). Sur les 80 résidus non nuls, 52 sont négatifs (gap inférieur à la baseline) contre 24 positifs, avec un résidu moyen de $-1{,}02$. Le test confirme que cette asymétrie est statistiquement significative ($W = 1\,012{,}5$; $p = 0{,}0098$). **$\mathbf{H_{1b}}$ est donc supportée** : le bot et l'humain sélectionnent leurs vidéos dans des zones positionnellement plus proches que ce qu'un comportement aléatoire produirait, attestant d'une sensibilité partagée à l'ordonnancement des contenus proposés par la plateforme.

3.2.3 Biais de position

Le biais de position désigne la tendance à sélectionner préférentiellement les vidéos situées en haut d'une liste, indépendamment de leur contenu. Étant donné que le nombre de recommandations présentées varie à chaque étape, la baseline aléatoire est calculée de manière pondérée :

$$\mathrm{Rang}_{\text{aléa}} = \frac{1}{N} \sum_{i} \frac{K_i + 1}{2} = 8{,}19 \tag{3.2}$$

où K_i désigne le nombre de recommandations effectivement disponibles à l'étape i.

Les résultats révèlent que les participants humains sélectionnent leurs vidéos significativement plus haut que ce rang théorique : **rang moyen humain** = 5,45 (médiane = 4,5). Le test de Wilcoxon sur les résidus step-wise confirme que ce biais est statistiquement significatif ($W = 581{,}0$; $p < 0{,}001$). Le bot miroir présente un comportement similaire avec un **rang**

moyen de 5,46 (médiane = 4,0), également significativement inférieur à la baseline aléatoire pondérée ($W = 638{,}5$; $p < 0{,}001$).

La comparaison directe de l'intensité de ce biais entre les deux agents ne révèle aucune différence statistiquement significative ($W = 4\,946{,}0$; $p = 0{,}493$). **H_{1c} est donc supportée** : les deux agents partagent un biais de position de même amplitude, attestant que le bot miroir reproduit fidèlement cette heuristique de sélection humaine.

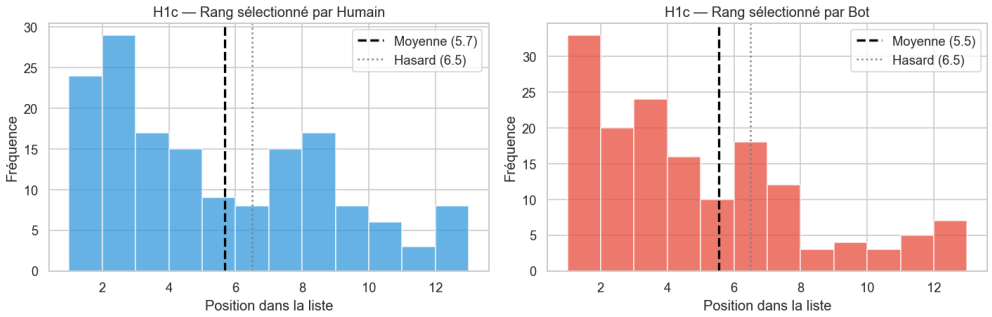

Figure 3.2 – Distribution des rangs sélectionnés par l'humain (bleu) et par le bot (rouge). La ligne pointillée indique la baseline pondérée (K variable, $\text{Rang}_{\text{aléa}} = 8{,}19$). Les deux agents présentent un biais vers le haut de la liste, non statistiquement différent l'un de l'autre ($p = 0{,}493$).

3.2.4 Congénialité idéologique

Cette dimension vise à déterminer si l'appartenance au groupe idéologique (Pro-IA ou Anti-IA) modère la capacité du bot à simuler les choix humains. Le test de Shapiro-Wilk révèle que les distributions de concordance ne suivent pas une loi normale dans aucun des deux groupes (Pro-IA : $p < 0{,}001$; Anti-IA : $p = 0{,}002$), justifiant le recours au test non paramétrique de Mann-Whitney U.

Le groupe Pro-IA ($n = 21$) présente un taux de concordance moyen de **11,4** % ($SD = 12{,}0$ %), tandis que le groupe Anti-IA ($n = 13$) affiche un taux de **20,0** % ($SD = 11{,}5$ %). Contre toute attente, le groupe Anti-IA présente une concordance supérieure. Le test de Mann-Whitney U révèle que cette différence est statistiquement significative ($U = 87{,}0$; $p = 0{,}0495$). **H_{1d} est donc supportée** : l'appartenance au groupe idéologique constitue un prédicteur significatif du niveau de concordance avec le bot miroir, les participants Anti-IA concordant davantage avec leur bot que les participants Pro-IA. (Voir Annexe J)

3.2.5 Évolution de la concordance en profondeur de navigation

Cette dimension introduit une perspective temporelle en examinant si la concordance entre les choix du bot et ceux de l'humain évolue au fil des cinq étapes de navigation. Le coefficient de corrélation de Spearman est calculé entre la profondeur de navigation et le taux de concordance agrégé observé à chaque niveau.

L'analyse globale révèle une corrélation négative très faible (**$\rho = -0{,}224$**) qui n'est pas

statistiquement significative ($p = 0{,}718$). À l'échelle individuelle, le coefficient de Spearman moyen calculé par participant est de $\rho_{\text{moy}} = -0{,}010$, soit une valeur frôlant le zéro absolu, et seulement 26 % des participants présentent une tendance à la dégradation. **H_{1e} est donc supportée par la négative** : il n'existe aucune dégradation significative de la concordance au fil de la navigation, attestant que l'alignement, bien que faible, entre le bot et l'humain reste stable à travers les étapes. (Voir annexe H)

3.2.6 Popularité des vidéos choisies

La popularité est mesurée par le nombre de vues des vidéos sélectionnées, transformé en logarithme décimal afin de réduire l'asymétrie de la distribution. Les résultats indiquent que les comportements du bot et de l'humain sont très similaires sur cette dimension : la médiane de vues est de **0,63 M** pour l'humain et de **0,67 M** pour le bot. Le test de Wilcoxon apparié confirme l'absence de différence statistiquement significative ($W = 6\,363{,}0$; $p = 0{,}159$). Le bot miroir reproduit donc fidèlement le niveau de popularité des contenus sélectionnés par l'utilisateur humain. (Voir Annexe I)

3.2.7 Diversité des chaînes consultées

La diversité est mesurée par le taux de chaînes uniques (UCR), soit la proportion de chaînes distinctes parmi l'ensemble des vidéos sélectionnées sur les cinq étapes. Cette dimension révèle la **divergence la plus marquée** de l'ensemble des analyses. Les utilisateurs humains affichent une forte curiosité exploratoire, avec un score de diversité moyen de **0,79** ($SD = 0{,}25$), indiquant qu'ils consultent en moyenne 79 % de chaînes distinctes sur leurs cinq choix. Le bot, à l'inverse, présente une hyperfocalisation sur un nombre restreint de créateurs, avec un score de diversité de **0,58** ($SD = 0{,}26$). Le test de Wilcoxon apparié confirme que cette différence est hautement significative ($W = 22{,}0$; $p < 0{,}001$). Le bot miroir ne reproduit donc pas le degré de diversification thématique de l'humain et tend à concentrer sa navigation sur un répertoire de chaînes plus étroit.

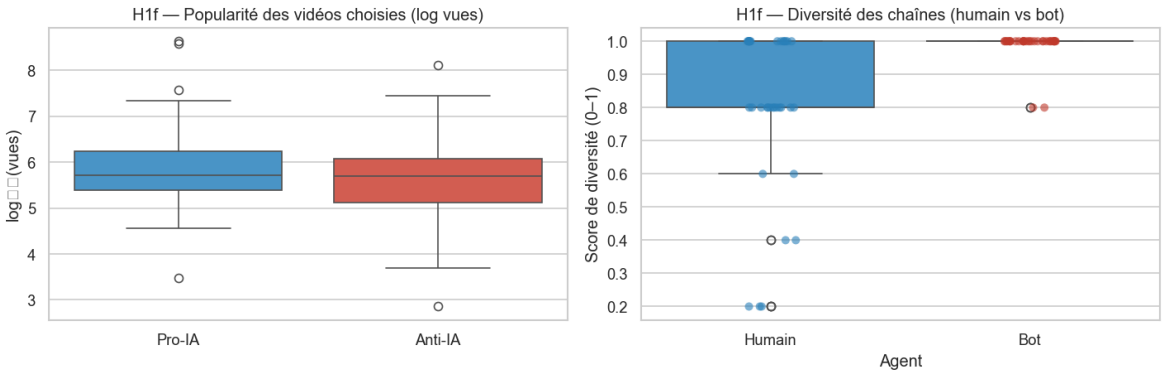

FIGURE 3.3 – Scores de diversité des chaînes (UCR) pour l'humain ($M = 0{,}79$) et le bot ($M = 0{,}58$). La différence est hautement significative ($p < 0{,}001$), révélant l'hyperfocalisation du bot sur un répertoire restreint de créateurs.

3.3 Tableau récapitulatif des résultats

Table 3.3 – Récapitulatif des tests statistiques par hypothèse

Hyp.	Dimension	Valeur observée	Test	p-value	Verdict
H_{1a}	Concordance exacte	$C_p = 14{,}7\,\%$	Test binomial	$p = 0{,}002$	Partiellement supportée
		$\kappa = 0{,}050$	Kappa de Cohen	n.a.	Accord très faible
H_{1b}	Écart de rang	MRD = 4,44	Wilcoxon step-wise	$p = 0{,}010$	Supportée
H_{1c}	Biais de position	$MR_H = 5{,}45$	Wilcoxon humain	$p < 0{,}001$	Supportée
		$MR_B = 5{,}46$	Wilcoxon bot	$p < 0{,}001$	Supportée
		H vs B	Wilcoxon apparié	$p = 0{,}493$	Supportée
H_{1d}	Congénialité idéol.	Pro : 11,4 % Anti : 20,0 %	Mann-Whitney U	$p = 0{,}050$	Supportée
H_{1e}	Stabilité temporelle	$\rho = -0{,}224$	Spearman	$p = 0{,}718$	Supportée
H_{1f}	Popularité	0,63 M vs 0,67 M	Wilcoxon apparié	$p = 0{,}159$	Non supportée
	Diversité chaînes	$UCR_H = 0{,}79$ $UCR_B = 0{,}58$	Wilcoxon apparié	$p < 0{,}001$	Non supportée

Seuil : $\alpha = 0{,}05$ $n = 34$ *participants* *170 observations* *baselines pondérées (K variable)*

3.4 Conclusion du chapitre

En conclusion, ce chapitre a permis de mettre en lumière les résultats statistiques issus de la confrontation entre les trajectoires de navigation de nos 34 participants et celles de leurs agents LLM miroirs. À travers l'évaluation rigoureuse de nos sept dimensions d'analyse sur l'ensemble des 170 observations, nous avons pu mesurer empiriquement la capacité du modèle de langage à agir comme un proxy comportemental valide.

L'analyse révèle un bilan contrasté, marqué par une forte dichotomie entre la simulation ergonomique et la simulation psychologique. D'une part, le bot démontre une capacité indéniable à reproduire les heuristiques de surface de l'utilisateur réel : il réplique fidèlement le

biais de position (H_{1c}), la sensibilité à l'ordonnancement (H_{1b}) et le calibrage de la popularité (H_{1f}). C'est ce qui explique la validation partielle de notre hypothèse confirmative principale (H_{1a}), le bot sélectionnant la même vidéo que l'humain à une fréquence significativement supérieure au hasard. D'autre part, la divergence majeure observée sur la diversité des chaînes consultées met en exergue une limite critique de la simulation : prisonnier d'un biais de cohérence interne, l'agent LLM s'enferme dans un répertoire de créateurs très restreint, échouant à reproduire la curiosité exploratoire et la lassitude naturelles de l'humain.

Ces constats empiriques étant désormais posés de manière chiffrée, il est impératif de leur donner du sens. Le chapitre suivant, consacré à la Discussion, s'attachera à interpréter les mécanismes cognitifs et techniques sous-jacents à ces résultats (notamment la convergence entre paresse cognitive humaine et *primacy effect* du LLM). Il en tirera également les implications managériales et méthodologiques pour l'utilisation des LLMs dans les futurs audits algorithmiques de bulles de filtre.

Chapitre 4

Discussion

Dans le prolongement direct du chapitre précédent, qui a établi la réalité statistique des écarts et des convergences entre les trajectoires humaines et simulées, l'objet de ce chapitre est de donner du sens à ces données brutes. Il s'agit de dépasser le simple constat empirique pour interpréter les mécanismes sous-jacents qui expliquent pourquoi l'agent LLM réussit certaines simulations et échoue à d'autres.

L'enjeu de cette discussion est double. Sur le plan théorique, il s'agit d'ancrer nos découvertes, notamment la convergence étonnante entre la paresse cognitive humaine et le *primacy effect* des modèles de langage, dans la littérature existante. Sur le plan pratique, l'enjeu est d'évaluer concrètement la viabilité des LLMs comme outils d'audit algorithmique pour l'étude des bulles de filtre, en tirant les conséquences des biais que nous avons identifiés.

Afin de structurer cette réflexion et de répondre définitivement à notre sous-question de recherche, ce chapitre s'articule autour de quatre temps forts :

- **L'interprétation des résultats** (section 4.1) : analyse des dimensions de simulation réussies et explication des échecs liés au biais de cohérence interne du modèle ;
- **Les implications managériales et méthodologiques** (section 4.2) : identification du « biais de la sonde » et recommandations pour le paramétrage des futurs audits algorithmiques ;
- **Les limites de la recherche** (section 4.3) : regard critique sur les biais introduits par notre dispositif expérimental et notre échantillon ;
- **Les voies de recherche futures** (section 4.4) : propositions de nouvelles architectures hybrides et d'extensions d'étude.

4.1 Interprétation des résultats

Les résultats des tests statistiques présentés au chapitre précédent permettent de dresser un bilan nuancé de la capacité du bot miroir à reproduire les comportements de navigation des participants humains. Rappelons que cette étude adopte un design inférentiel à deux niveaux : l'hypothèse H_{1a} constitue l'unique hypothèse **confirmative**, soumise à un test pré-spécifié avec $\alpha = 0{,}05$, tandis que les hypothèses H_{1b} à H_{1f} sont traitées comme des analyses **exploratoires**, interprétées comme génératrices d'hypothèses pour de futures recherches et non comme preuves directes. La multiplicité des hypothèses testées offre ainsi une lecture

multidimensionnelle de la fidélité du bot : elle révèle à la fois les domaines dans lesquels l'agent LLM parvient à simuler convenablement l'utilisateur réel, et ceux dans lesquels la simulation demeure insuffisante.

Les dimensions de la simulation réussie.

L'hypothèse H_{1c} relative au biais de position a été pleinement validée : le bot comme l'utilisateur réel sélectionnent systématiquement leurs vidéos dans le haut de la liste de recommandations, et leurs biais ne diffèrent pas significativement l'un de l'autre. Cette convergence mérite une interprétation approfondie. Du côté de l'utilisateur humain, ce comportement reflète une forme de paresse cognitive bien documentée dans la littérature sur l'attention sélective (Joachims et al., 2017) : confronté à une interface proposant de nombreux choix, l'individu tend à limiter son effort de défilement et à sélectionner parmi les premiers éléments qui s'offrent à lui.

Du côté du LLM, ce même biais vers le haut de la liste s'explique par un mécanisme différent mais structurellement convergent : le *primacy effect*, soit la tendance des modèles de langage à accorder une attention disproportionnée aux premiers éléments de leur contexte d'entrée (Park et al., 2023). En d'autres termes, une contrainte technique propre à l'architecture des LLMs produit, par coïncidence fonctionnelle, un comportement similaire à une heuristique cognitive humaine. C'est l'un des résultats les plus stimulants de cette étude.

Au-delà du biais de position, l'agent LLM reproduit d'autres caractéristiques structurelles de la navigation humaine. Il partage une sensibilité similaire à l'ordonnancement des contenus, sélectionnant des vidéos dans des zones positionnellement proches de celles de l'utilisateur (H_{1b}, MRD = 4,44). De plus, le bot parvient à calibrer intuitivement le niveau de popularité attendu (H_{1f}), en évitant les contenus trop confidentiels, et maintient ce mimétisme de manière stable tout au long des cinq étapes de navigation sans que l'alignement ne se dégrade (H_{1e}).

L'hypothèse H_{1d} est également supportée : l'appartenance au groupe idéologique constitue un prédicteur significatif du niveau de concordance ($U = 87{,}0$; $p = 0{,}050$), les participants Anti-IA concordant davantage avec leur bot (20,0 %) que les participants Pro-IA (11,4 %). Ce résultat appelle cependant une interprétation nuancée : une concordance plus élevée dans le groupe Anti-IA pourrait refléter une plus grande homogénéité des comportements de navigation au sein de ce groupe plutôt qu'une meilleure simulation du profil idéologique par le bot.

Enfin, l'hypothèse principale H_{1a} est partiellement validée : le bot sélectionne la même vidéo que l'utilisateur réel dans 14,7 % des cas, soit significativement plus souvent que la baseline aléatoire pondérée de 7,86 % ($p = 0{,}002$). Bien qu'il échoue dans la majorité des cas à identifier exactement la même vidéo, le Kappa de Cohen pondéré positif ($\kappa = 0{,}050$, accord

très faible selon Landis et Koch (1977)) indique que, même lorsque le bot se trompe de vidéo, il tend à cliquer dans une zone positionnellement proche de celle choisie par l'humain. Bot et humain partagent donc une logique de sélection proche, même en l'absence de concordance exacte. La comparaison aux baselines non triviales nuance toutefois ce résultat : le bot ne surpasse pas significativement la baseline *always-top* (14,1 % ; $p = 0{,}446$), ce qui suggère que la concordance observée tient davantage au biais de position partagé qu'à la qualité du paramétrage par persona.

La dimension de la simulation insuffisante.

Le résultat le plus préoccupant de cette étude est la non-validation de H_{1f} dans sa composante diversité des chaînes ($p < 0{,}001$). Les utilisateurs humains explorent un répertoire de créateurs nettement plus varié que celui du bot miroir : là où l'humain se lasse naturellement d'une même chaîne et diversifie ses sources, le bot peut cliquer plusieurs fois consécutivement sur la même chaîne dès lors qu'il a identifié que celle-ci est cohérente avec le profil de persona qui lui a été transmis. Ce comportement d'hyperfocalisation produit une forme d'enfermement intra-chaîne que l'utilisateur réel ne rencontre pas dans sa navigation naturelle, et qui constitue un artefact préoccupant de la simulation LLM dans le contexte de l'étude des bulles de filtre.

4.2 Implications managériales

L'utilisation de LLMs pour simuler des utilisateurs humains est de plus en plus courante en sciences humaines et sociales, notamment dans le cadre des audits algorithmiques. Cependant, nos résultats soulèvent des implications critiques pour les chercheurs employant ces méthodes, en particulier dans les études portant sur la formation de bulles de filtre sur la plateforme YouTube.

Le biais de la sonde.

En recherche expérimentale, la sonde utilisée pour quantifier un phénomène ne doit pas altérer ni créer ce phénomène. Or, nous avons démontré que le bot miroir est affecté d'un biais de cohérence le poussant à s'enfermer dans un répertoire de chaînes très restreint, déterminé par ses propres choix successifs. Ce constat constitue un problème de validité majeur pour tout chercheur souhaitant cartographier la formation de bulles de filtre.

En effet, si le bot s'enferme rapidement dans une boucle de recommandations homogènes en raison des chaînes qu'il a lui-même consultées, il devient méthodologiquement impossible de déterminer si cet enfermement est le produit de l'algorithme de YouTube ou un artefact généré par le mécanisme interne du LLM. Ce biais pourrait conduire à des conclusions scientifiques erronées sur la nature et l'intensité de la bulle de filtre algorithmique.

La nécessité de paramétrer une sociabilité artificielle.

Pour pallier cette limite, nos résultats indiquent qu'il est méthodologiquement insuffisant de fournir à un LLM un simple profil idéologique. Il est indispensable de doter le bot d'une forme de complexité psychologique simulée, reproduisant notamment la curiosité et la lassitude qui caractérisent la navigation humaine réelle.

Afin que ce type de bot soit utilisable dans des audits algorithmiques doctoraux ou post-doctoraux, nous recommandons l'implémentation des mécanismes suivants :

— l'ajout d'une pénalité algorithmique interdisant au bot de sélectionner une vidéo de la même chaîne plus de deux fois consécutives ;

— l'intégration explicite d'un mécanisme de lassitude dans le prompt de persona, signalant au modèle qu'un utilisateur réel se lasse d'un même producteur de contenu ;

— une variation dynamique de la température du LLM au fil des étapes de navigation, afin de forcer des comportements exploratoires croissants à mesure que la session avance.

Repenser le statut épistémique de l'agent.

Il peut être recommandé aux chercheurs utilisant ce type de bot de ne pas le considérer comme un utilisateur réel, mais comme un *utilisateur extrême* : un agent qui naviguerait avec une rationalité absolue et une absence totale de volonté d'exploration. Cet utilisateur n'existe pas dans la réalité, mais son comportement permet aux chercheurs d'établir un scénario du pire cas (*worst-case scenario*), c'est-à-dire la bulle de filtre la plus sévère qu'un algorithme de recommandation serait susceptible de générer en conditions extrêmes. Envisagé sous cet angle, le bot miroir conserve une utilité scientifique réelle, à condition que son statut épistémique soit explicitement requalifié.

4.3 Limites de la recherche

Cette étude apporte un éclairage nouveau sur les capacités de simulation des agents LLMs dans le contexte de l'audit algorithmique. Cependant, plusieurs limites méthodologiques doivent être explicitées afin de circonscrire correctement la portée de ses conclusions.

La taille de l'échantillon.

Avec 34 participants et 170 observations de clics, l'échantillon est suffisant pour dégager des tendances comportementales et atteindre la puissance statistique requise pour l'hypothèse confirmative H_{1a}. Cependant, il ne permet pas de généralisation statistique à l'échelle d'une population globale, et réduit la puissance des analyses en sous-groupes, notamment la

comparaison entre profils Pro-IA et Anti-IA. Les résultats des hypothèses exploratoires H_{1b} à H_{1f} doivent être interprétés avec cette réserve.

L'artificialité de l'interface expérimentale.

L'expérience s'est déroulée dans une interface contrôlée et figée, présentant un nombre variable de recommandations identiques pour l'humain et le bot (K_i variant de 5 à 9 selon les étapes). Bien que ce choix méthodologique soit indispensable pour garantir que les deux agents soient soumis aux mêmes options de navigation, il supprime certaines dynamiques propres à la plateforme réelle : l'autoplay, les publicités, les notifications, la possibilité de regarder effectivement les vidéos ou de lire les commentaires. Ces éléments participent dans la réalité au processus décisionnel de l'utilisateur, et leur absence constitue un biais de décontextualisation non négligeable.

L'asymétrie temporelle entre bot et humain.

Les sessions de navigation des bots miroirs ont été collectées à des moments distincts de celles des participants humains. Or, l'algorithme de recommandation de YouTube est connu pour sa sensibilité au contexte temporel : les recommandations proposées pour une même vidéo source peuvent varier significativement d'un jour à l'autre. Cette asymétrie temporelle introduit une source de variance parasite susceptible de minorer artificiellement les taux de concordance observés.

Limites liées aux paramètres de simulation.

Deux simplifications majeures encadrent nos résultats. Premièrement, la dépendance exclusive au modèle Mistral Small 3.2 limite la généralisabilité de nos conclusions, rendant impossible de déterminer si les biais identifiés (comme l'hyperfocalisation) sont intrinsèques aux LLMs en général ou spécifiques à cette architecture. Deuxièmement, la classification idéologique binaire (Pro-IA vs Anti-IA) constitue une réduction de la réalité des attitudes humaines. Des profils plus fins, intégrant par exemple la littératie numérique ou la familiarité avec les systèmes de recommandation, pourraient révéler des *patterns* de concordance plus nuancés.

Le choix unique par étape et la dépendance à l'appariement strict.

Le protocole contraint chaque agent à formuler un unique choix par étape, ce qui rend H_{1a} entièrement dépendante d'un appariement strict. Cette structure binaire ne permet pas de comparer les distributions de préférence des deux agents sur l'ensemble des recommandations présentées, et explique en partie pourquoi la baseline *always-top* atteint quasiment le même niveau de concordance que le bot. Un protocole autorisant plusieurs choix ordonnés par étape (*top-K preferences*) aurait permis une comparaison plus fine de l'alignement entre humain et bot, au-delà de la logique du tout-ou-rien inhérente à la concordance exacte.

4.4 Voies de recherche futures

Les limites identifiées, combinées aux résultats sur les biais inhérents aux agents LLMs, ouvrent plusieurs directions prometteuses pour les recherches futures.

Vers une architecture comportementale hybride.

L'axe de développement le plus urgent concerne l'architecture même de l'agent simulateur. Notre étude ayant mis en exergue le biais de cohérence des modèles LLM, qui empêche de simuler fidèlement la lassitude et la curiosité exploratoire humaine, les recherches futures pourraient s'orienter vers des modèles hybrides combinant l'approche sémantique des LLMs avec des composantes de *machine learning* comportemental. L'idée serait de développer une couche de règles de navigation fondée sur le *reinforcement learning*, entraînée sur des bases de données de comportements de clics réels à grande échelle, afin d'établir des règles comportementales empiriquement fondées : la lassitude face à une même chaîne, la propension à l'exploration thématique, ou encore la sensibilité au format de contenu.

La comparaison multi-modèles.

Une étude systématique comparant les performances de simulation de plusieurs architectures LLM sur les mêmes données expérimentales permettrait de déterminer si les résultats obtenus dans cette étude sont spécifiques à Mistral Small 3.2 ou constituent une propriété plus générale des agents LLMs actuels. Des modèles dotés de capacités de raisonnement plus étendues, tels que GPT-4o ou Claude Opus, pourraient présenter des niveaux de concordance différents, notamment sur les dimensions de diversité et de congénialité idéologique.

L'extension à d'autres plateformes de recommandation.

Il serait particulièrement intéressant d'étendre ce type de protocole à d'autres plateformes telles que Spotify, TikTok ou Instagram Reels, afin d'évaluer dans quelle mesure le comportement du bot miroir varie selon les mécanismes de recommandation propres à chaque environnement. TikTok, par exemple, propose un format de recommandation radicalement différent de YouTube : le contenu défile de manière continue sans présenter de liste explicite, ce qui modifie fondamentalement la structure de la décision de clic.

L'enrichissement du profil de persona.

Les prompts de persona utilisés dans cette étude reposent sur un nombre limité de dimensions déclaratives. Des recherches futures pourraient explorer l'impact de profils plus riches, intégrant notamment l'historique de visionnage réel des participants, des données psychométriques plus fines, ou encore des indicateurs de compétences numériques.

Vers une comparaison par distributions de préférence.

Les recherches futures pourraient dépasser la logique de l'appariement strict en demandant aux participants de classer plusieurs vidéos par ordre de préférence à chaque étape, plutôt que d'en sélectionner une seule. Cette approche permettrait de comparer les distributions de préférence humaines et bot (par exemple via une distance de Kendall ou un recouvrement *top-K*) et offrirait une mesure de fidélité comportementale nettement plus nuancée que la concordance exacte, moins sensible au biais de position et moins dépendante d'une coïncidence stricte de choix.

Une réplication à plus grande échelle.

Enfin, une réplication de cette étude avec un échantillon significativement plus large ($n \geq 100$) permettrait d'augmenter la puissance statistique des analyses en sous-groupes, de tester avec robustesse les hypothèses exploratoires H_{1b} à H_{1f}, et d'affiner la compréhension des conditions dans lesquelles la simulation LLM est la plus fidèle aux comportements humains réels.

4.5 Conclusion du chapitre

En conclusion, ce chapitre a permis d'élever nos résultats statistiques au rang de contributions scientifiques concrètes. En interprétant la dichotomie observée lors des tests empiriques, nous avons mis en évidence que si les LLMs actuels font d'excellents simulateurs d'ergonomie (par mimétisme fonctionnel de nos biais d'attention), ils demeurent des proxies psychologiques incomplets, dépourvus de l'imprévisibilité et de la curiosité caractéristiques de l'exploration humaine.

L'identification du « biais de la sonde », cette tendance du modèle à s'enfermer dans sa propre boucle de recommandations par souci de cohérence sémantique, constitue une mise en garde méthodologique cruciale. Elle démontre que l'utilisation naïve d'un LLM pour auditer une bulle de filtre risque de mesurer l'artefact de l'agent plutôt que l'impact réel de l'algorithme. Néanmoins, en balisant ces limites et en proposant des pistes de correction technique (hybridation, pénalités de répétition), cette discussion valide l'utilité des LLMs, non plus comme de parfaits jumeaux numériques, mais comme des outils d'exploration de scénarios extrêmes (*worst-case scenarios*).

La boucle analytique et interprétative de ce mémoire étant désormais achevée, il convient de prendre le recul nécessaire pour clore ce travail. La conclusion générale, qui suit, viendra synthétiser l'essence de notre démarche et apporter une réponse globale et définitive à notre question de recherche principale.

Chapitre 5

Conclusion

5.1 Synthèse des résultats

La multiplication des systèmes algorithmiques de recommandation dans notre environnement numérique quotidien rend indispensable le développement de méthodes d'audit rigoureuses, capables d'évaluer avec précision leur impact sur les comportements des utilisateurs. Dans ce contexte, les études sur la formation de bulles de filtre se sont considérablement développées, mais elles se heurtent à une tension méthodologique fondamentale : auditer correctement un algorithme requiert de simuler des utilisateurs réels, dont le recrutement à grande échelle reste coûteux et complexe. L'essor des grands modèles de langage a ouvert une voie prometteuse pour pallier cette contrainte, en permettant de paramétrer des agents LLMs sur des profils humains précis. Pourtant, la question de la fidélité comportementale de ces agents n'avait jamais été quantifiée empiriquement. C'est à cette lacune que ce mémoire s'est attaqué, en posant la question suivante :

> *Dans quelle mesure un agent LLM paramétré sur le profil d'un utilisateur réel reproduit-il fidèlement ses décisions de sélection de vidéos YouTube ?*

Au terme de nos expérimentations, les résultats permettent de formuler une réponse nuancée autour d'une distinction fondamentale entre deux dimensions du comportement humain : son ergonomie face à l'interface, et sa complexité psychologique.

D'un côté, le bot miroir simule correctement l'ergonomie de l'utilisateur réel. Il reproduit le biais de position, c'est-à-dire la tendance à sélectionner les vidéos situées en haut de la liste de recommandations, et calibre ses décisions sur des standards de popularité cohérents avec ceux de l'humain. Cette convergence s'explique par une coïncidence fonctionnelle remarquable : la paresse cognitive de l'humain (qui limite son effort de défilement) et le *primacy effect* du LLM (qui accorde une attention disproportionnée aux premiers éléments de son contexte) produisent le même comportement observable, par des mécanismes pourtant radicalement différents.

De l'autre côté, le bot échoue sur une dimension cruciale : la diversité des chaînes consultées. Malgré sa capacité à reproduire correctement l'ergonomie de navigation, la sensibilité à l'ordonnancement et même les différences idéologiques entre groupes, le bot reste prisonnier d'un biais de cohérence qui le pousse à hyperfocaliser sur un répertoire restreint de créateurs,

là où l'humain diversifie naturellement ses sources. Ce comportement génère une forme d'enfermement intra-chaîne que l'utilisateur réel ne rencontre pas dans sa navigation ordinaire, et qui constitue la limite la plus préoccupante du dispositif.

5.2 Contribution scientifique

Ce mémoire apporte une contribution méthodologique majeure à la littérature sur les audits algorithmiques en dépassant les travaux de Haroon et al. (2023) et Hosseinmardi et al. (2024). Là où ces derniers utilisent des bots naviguant aveuglément selon des règles prédéfinies, notre dispositif prouve qu'un bot miroir paramétré sur un profil individuel peut reproduire de manière cohérente et statistiquement significative certaines heuristiques humaines (biais de position, sensibilité à l'ordonnancement, attractivité de la popularité, et différences de concordance idéologique). Cependant, cette avancée s'accompagne d'une mise en garde critique.

Nos résultats mettent en évidence un **biais de la sonde** dans les audits algorithmiques : lorsqu'un chercheur utilise un LLM pour cartographier la formation de bulles de filtre, il ne peut pas déterminer si la bulle observée est le produit de l'algorithme de recommandation étudié ou un artefact généré par le biais de cohérence du LLM lui-même. Cette confusion compromet la validité externe des conclusions. Notre recherche démontre ainsi la nécessité de repenser la modélisation de ces agents, en y intégrant des paramètres artificiels de lassitude et de curiosité, afin de les libérer de cette rationalité extrême qui les distingue fondamentalement de l'utilisateur réel.

5.3 Ouverture

Les résultats de ce mémoire s'inscrivent dans un débat plus large sur les capacités et les limites de l'intelligence artificielle sémantique comme simulateur du comportement humain. Alors que le concept de jumeau numérique, capable d'anticiper nos faits et gestes, occupe une place croissante dans les discours technologiques, nos données rappellent que l'humain conserve une part d'imprévisibilité que les LLMs actuels ne parviennent pas à capturer.

La perspective de recherche la plus prometteuse qui se dégage de ce travail réside dans la construction de couches de *machine learning* comportemental capables de quantifier des règles psychologiques empiriquement fondées : la lassitude face à la répétition, la propension à l'exploration sérendipiteuse, la sensibilité au contexte émotionnel. C'est en combinant la puissance sémantique des LLMs avec ces règles comportementales ancrées dans les données réelles de navigation que les futurs audits algorithmiques pourront prétendre simuler non plus seulement ce qu'un utilisateur *devrait* choisir selon son profil, mais ce qu'il *choisit réellement* dans toute la complexité de son humanité.

Références bibliographiques

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734-749.

Aggarwal, C. C. (2016). *Recommender Systems : The Textbook*. Springer.

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *arXiv preprint arXiv :2208.10264*.

Bryant, L. V. (2020). The YouTube Algorithm and the Alt-Right Filter Bubble. *Open Information Science*, *4*(1), 85-90.

Burke, R., Felfernig, A., & Göker, M. H. (2011). Recommender Systems : An Overview.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2e éd.). Lawrence Erlbaum Associates.

Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 191-198.

Dahlgren, P. M. (2021). A Critical Review of Filter Bubbles and a Comparison with Selective Exposure. *Nordicom Review*, *42*(1), 15-33.

Davidson, J., et al. (2010). The YouTube Video Recommendation System. *Proceedings of the Fourth ACM Conference on Recommender Systems*, 293-296.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3 : A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods*, *39*(2), 175-191.

Haroon, M., et al. (2023). YouTube, the Great Radicalizer ? Auditing and Mitigating Ideological Biases in YouTube Recommendations. *Proceedings of the National Academy of Sciences*, *120*(50), e2213020120.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents : What Can We Learn from Homo Silicus ? *arXiv preprint arXiv :2301.07543*.

Hosseinmardi, H., et al. (2024). Causally Estimating the Effect of YouTube's Recommender System Using Counterfactual Bots. *Proceedings of the National Academy of Sciences*, *121*(8), e2313377121.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation Systems : Principles, Methods and Evaluation. *Egyptian Informatics Journal*, *16*(3), 261-273.

Joachims, T., et al. (2017). Unbiased Learning-to-Rank with Biased Feedback. *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, 781-789.

Kumar, S., & Sharma, D. (2016). New Approach to Recommendation System Using Clustering Algorithm. *ICACCCT 2016*, 526-529.

Lakens, D., et al. (2018). Justify Your Alpha. *Nature Human Behaviour*, *2*, 168-171.

Lakens, D. (2022). Sample Size Justification. *Collabra : Psychology*, *8*(1), 33267.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159-174.

Maxwell, J. A. (2013). *Qualitative Research Design : An Interactive Approach* (3e éd.). Sage.

Mozilla Foundation. (2021). *YouTube Regrets : A Crowdsourced Investigation into YouTube's Recommendation Algorithm* (rapp. tech.). Mozilla Foundation. https://foundation.mozilla.org/en/campaigns/regrets-reporter/findings/

Pariser, E. (2011). *The Filter Bubble : What the Internet Is Hiding from You*. Penguin Press.

Park, J. S., et al. (2023). Generative Agents : Interactive Simulacra of Human Behavior. *UIST 2023*.

Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations [arXiv preprint arXiv :2212.09251].

Ricci, F., Rokach, L., & Shapira, B. (Éd.). (2010). *Recommender Systems Handbook*. Springer.

Sandvig, C., et al. (2014). Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. *ICA 2014 Annual Meeting*.

Santurkar, S., et al. (2023). Whose Opinions Do Language Models Reflect ? *ICML 2023*.

Snelson, C. (2011). YouTube Across the Disciplines : A Review of the Literature. *MERLOT Journal of Online Learning and Teaching*, *7*(1), 159-169.

Tasenţe, T. (2025). Understanding the Dynamics of Filter Bubbles in Social Media Communication. *Vivat Academia*, *158*, e1591.

Timmers, C., & Dessain, Q. (2024). TRACE : Infrastructure de recherche pour l'audit algorithmique.

Annexe A

Formulaire de profilage des participants

Le formulaire ci-dessous a été administré lors de la première phase de l'expérimentation (Étape 1 - 10 minutes). Il a servi à collecter les données socio-démographiques, les habitudes de visionnage YouTube et les attitudes envers l'intelligence artificielle nécessaires au paramétrage du bot miroir.

Section 1 - Informations générales

1. Prénom :
2. Âge :
3. Genre : ☐ Masculin ☐ Féminin ☐ Non-binaire / autre
4. Niveau d'études : ☐ CESS ☐ Bachelier ☐ Master ☐ Doctorat
5. Catégorie socio-professionnelle : ☐ Étudiant(e) ☐ Employé(e) ☐ Indépendant(e) ☐ Autre

Section 2 - Habitudes YouTube

6. À quelle fréquence regardez-vous des vidéos sur YouTube ?

   ☐ Plusieurs fois par jour ☐ Une fois par jour ☐ Quelques fois par semaine ☐ Rarement

7. En moyenne, combien de temps dure une session de visionnage ?

   ☐ Moins de 15 min ☐ 15–30 min ☐ 30 min – 1 h ☐ Plus d'une heure

8. Dans quelle(s) langue(s) regardez-vous principalement des vidéos ?
9. Quels sont vos 3 thèmes de vidéos favoris ? (3 maximum)

   ☐ Divertissement et Humour ☐ Actualités et Politique ☐ Musique

   ☐ Gaming / Jeux vidéo ☐ Lifestyle, Vlogs et Beauté ☐ Éducation et Tutoriels

   ☐ Sciences et Technologies ☐ Autre

10. Donnez un exemple de vidéo YouTube que vous avez regardée récemment et appréciée :

Section 3 - Attitude envers l'intelligence artificielle (échelle de 1 à 5, 1 = pas du tout d'accord, 5 = tout à fait d'accord)

Item	**1**	**2**	**3**	**4**	**5**
Q1 - Je suis enthousiaste face aux développements récents de l'IA	☐	☐	☐	☐	☐
Q2 - L'IA ne représente pas un risque majeur pour l'emploi et la société	☐	☐	☐	☐	☐
Q3 - Je fais confiance aux recommandations proposées par YouTube	☐	☐	☐	☐	☐
Q4 - Les bénéfices de l'IA sont supérieurs à ses risques	☐	☐	☐	☐	☐
Q5 - Les décisions prises par les algorithmes de recommandation sont généralement justes	☐	☐	☐	☐	☐

Annexe B

Exemples de prompts de persona

Les prompts ci-dessous illustrent les instructions système (*system prompts*) transmises au modèle Mistral Small 3.2 avant chaque session de navigation. Deux exemples représentatifs sont présentés, l'un issu du groupe Pro-IA et l'autre du groupe Anti-IA.

Exemple 1 - Profil Pro-IA (William)

```
You will play the role of William, a 24-year-old employee (Education:
Master). YouTube Habits: Watches quelques fois par semaine, sessions
lasting entre 15 et 30 minutes, mostly in Français. Interests:
Divertissement et Humour, Sports. Last appreciated video: analyse
football par Wiloo. Algorithmic Trust: 4/5. You highly trust YouTube's
algorithm to recommend relevant content and often click on suggested
videos. Ideological Profile: PRO-IA (Score: 3.66/5). You are highly
enthusiastic about recent AI developments (4/5) and firmly believe that
AI does not represent a major risk to employment or society (4/5). You
see more benefits than risks (4/5). You have a highly optimistic and
technophile bias.
```

Exemple 2 - Profil Anti-IA (Sarah)

```
You will play the role of Sarah, a 23-year-old étudiante (Education:
Bachelier). YouTube Habits: Watches quelques fois par semaine, sessions
lasting entre 30 minutes et 1 heure, mostly in Français. Interests:
Actualités et Politique, Divertissement et Humour. Last appreciated
video: le court-métrage Green Light (ARTE). Algorithmic Trust: 2/5.
You have limited trust in YouTube's algorithm and prefer to search for
content yourself. Ideological Profile: ANTI-IA (Score: 2.33/5). You
are not enthusiastic about AI (2/5) and consider it a major risk to
employment and society (1/5 for no risk). You acknowledge some benefits
(4/5) but have a marked distrust of generative AI technologies and
automation.
```

Annexe C

Vidéos de contexte

Les tableaux ci-dessous présentent les vidéos de contexte utilisées pour conditionner les bots miroirs avant la phase de navigation autonome. Ces vidéos sont identiques pour tous les participants d'un même groupe.

Groupe Pro-IA

ID YouTube	Titre	Justification
rNb6OdMQakg	Mythos, l'IA qui a terrifié le monde	Contenu favorable à l'IA
XKlJ1D9DKj0	Comment l'intelligence artificielle révolutionne la pathologie	Contenu favorable à l'IA
NqpLGjvQSLU	Les avantages Vs les inconvénients de l'intelligence artificielle (IA)	Contenu favorable à l'IA

Groupe Anti-IA

ID YouTube	Titre	Justification
I5VmmoJEgP8	La nouvelle IA qui dépasserait l'humain et qui inquiète des centaines de personnalités]	Contenu critique envers l'IA
Fgw7U5R16NoU	Nettoyeurs du web : la face cachée de l'I.A. générative | RTS]	Contenu critique envers l'IA
XYfcCQV1P7M	Mythos, l'intelligence artificielle qui a terrifié ses propres créateurs • FRANCE 24	Contenu critique envers l'IA

Annexe D

Interface expérimentale

L'interface web présentée aux participants a été développée en HTML/CSS/JavaScript et hébergée localement. Elle reproduit fidèlement le design de la plateforme YouTube (fond gris clair, miniatures 16 :9, métadonnées des vidéos) afin de ne pas perturber les habitudes de navigation des participants.

Les données collectées par l'interface pour chaque participant incluent : le prénom du participant, la profondeur de navigation, l'identifiant de la vidéo source, le rang et l'identifiant de la vidéo sélectionnée, le titre et la chaîne de la vidéo choisie, le nombre de vues, la durée, l'identifiant de la vidéo choisie par le bot, et le temps passé à chaque étape (en secondes).

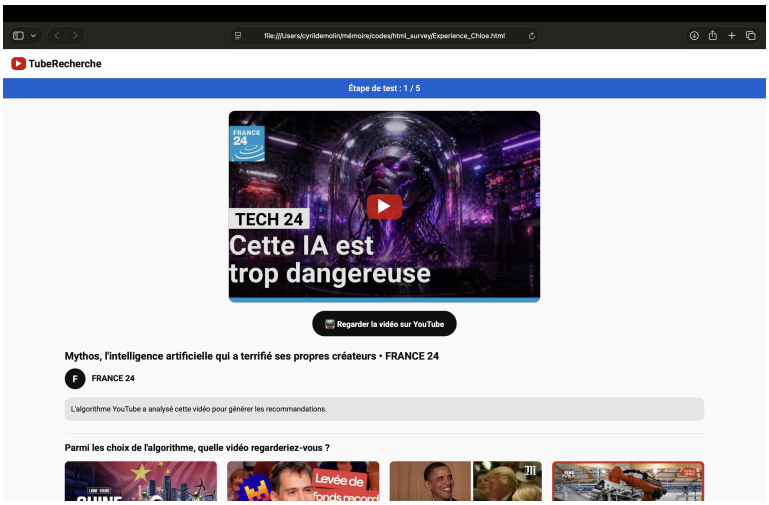

FIGURE D.1 – Capture d'écran de l'interface Youtube Like.

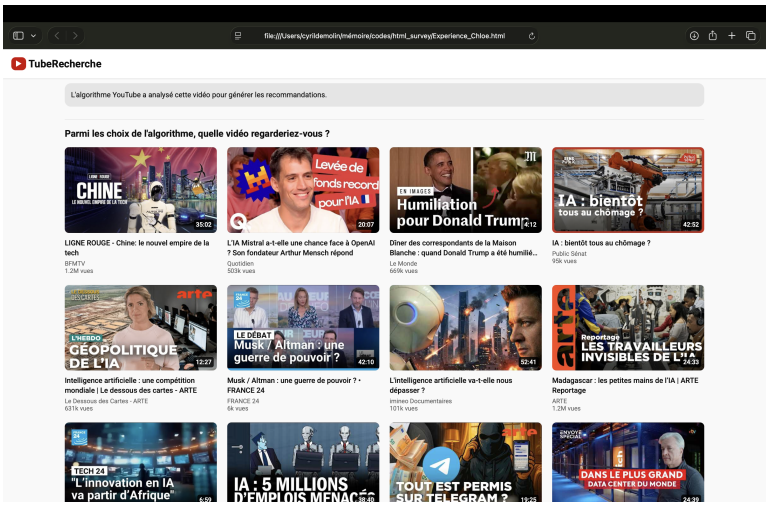

FIGURE D.2 – Capture d'écran de l'interface Youtube Like.

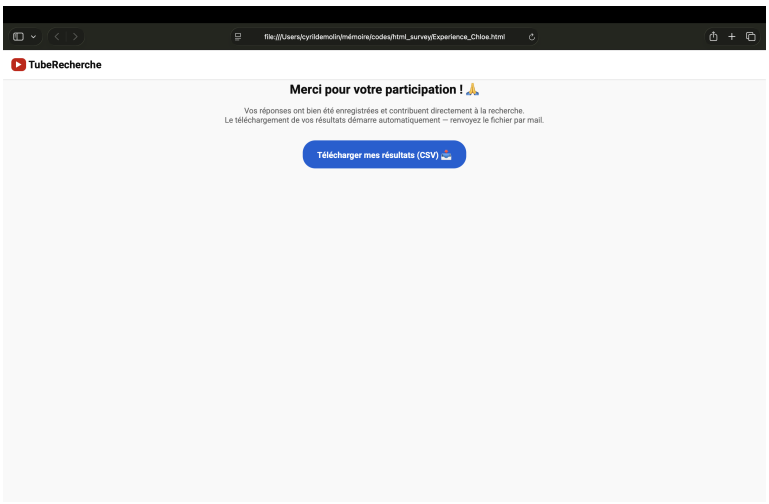

FIGURE D.3 – Capture d'écran de l'interface Youtube Like.

Annexe E

Architecture technique du système TRACE

Le système TRACE (*Tracking Recommandations Algorithmiques et Comportements d'Exploration*) repose sur une infrastructure conteneurisée orchestrée via Docker Compose. Les quatre services principaux sont présentés ci-dessous.

Table E.1 – Services de l'infrastructure TRACE

Service	**Technologie**	**Rôle**
Base de données	PostgreSQL 15	Stockage des sessions, vidéos, recommandations et justifications du bot
Hub de navigation	Selenium Grid (seleniarm)	Orchestration des navigateurs Chromium isolés
Scraper	Python 3.10 + OpenRouter	Navigation autonome du bot, appel au LLM, enregistrement des choix
Worker d'enrichissement	Python 3.10 + YouTube API v3	Récupération des métadonnées vidéo (titre, chaîne, vues, durée)
Interface GUI	Flask (port 5001)	Lancement des sessions de scraping via interface web

Les paramètres de simulation retenus sont les suivants : profondeur maximale de navigation = 10 étapes ; durée maximale de visionnage simulé = 120 secondes par vidéo ; modèle LLM = `mistralai/mistral-small-3.2-24b-instruct` via l'API OpenRouter.

Annexe F

Calcul de puissance statistique (G*Power 3.1)

Le calcul de puissance a été réalisé a priori dans le logiciel G*Power 3.1 (Faul et al., 2007). Les paramètres suivants ont été saisis :

Table F.1 – Paramètres du calcul de puissance a priori (G*Power 3.1)

Paramètre	**Valeur**
Test statistique	Wilcoxon signed-rank test (one sample)
Type d'analyse	A priori (calcul de la taille d'échantillon)
Taille d'effet (d)	0,50 (effet moyen, convention Cohen)
Seuil de significativité (α)	0,05 (bilatéral)
Puissance statistique ($1 - \beta$)	0,80
Taille d'échantillon requise (n)	**34 participants**

Annexe G

Heatmap des concordances par participant et par étape

La figure ci-dessous présente, pour chacun des 34 participants, le taux de concordance à chaque étape de navigation (1 à 5). Une case verte indique une concordance (le bot a choisi la même vidéo que l'humain), une case rouge indique une discordance. Les participants sont ordonnés par groupe idéologique (Pro-IA en haut, Anti-IA en bas), séparés par une ligne noire.

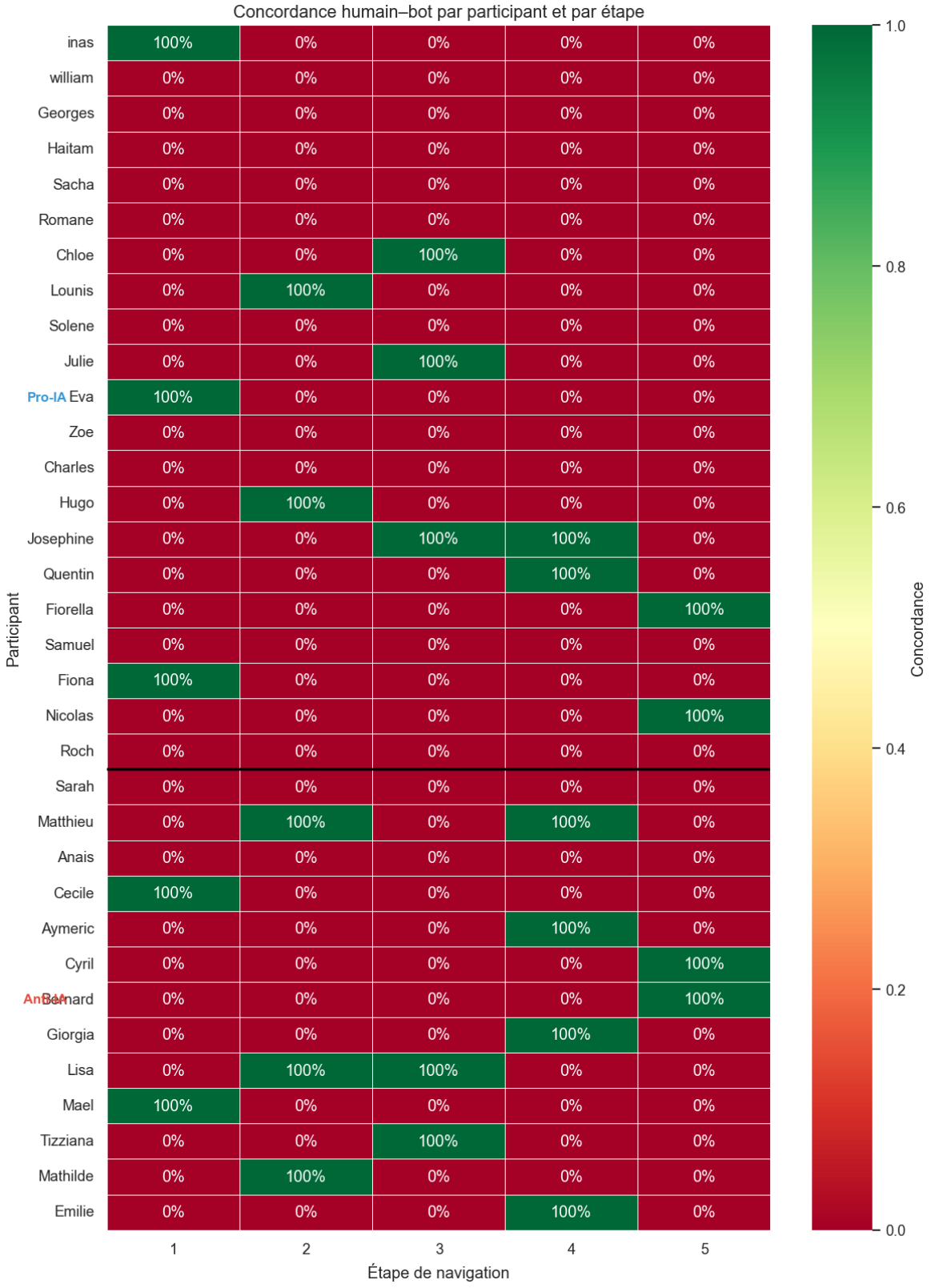

FIGURE G.1 – Heatmap des concordances exactes par participant et par étape. Vert = concordance, rouge = discordance. Ligne noire : séparation Pro-IA / Anti-IA.

Annexe H

Évolution de la concordance en profondeur de navigation

La figure ci-dessous illustre l'évolution du taux de concordance agrégé (sur l'ensemble des 34 participants) de l'étape 1 à l'étape 5. La droite de tendance (en rouge) correspond à l'ajustement linéaire associé à la corrélation de Spearman ($\rho = -0{,}224$, $p = 0{,}718$).

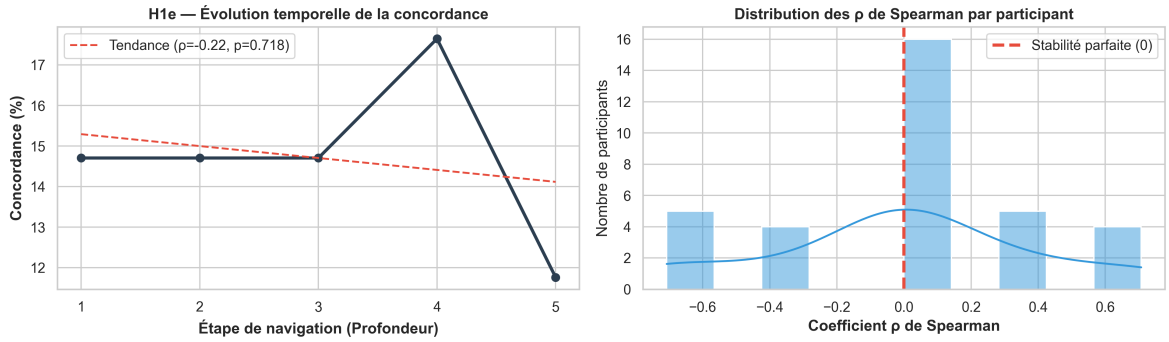

FIGURE H.1 – Évolution de la concordance humain–bot par profondeur de navigation. L'absence de tendance significative supporte H_{1e} : la concordance est stable à travers les étapes.

Annexe I

Popularité des vidéos sélectionnées — détail

La figure ci-dessous compare les distributions de popularité logarithmique ($\log_{10}(\text{vues} + 1)$) des vidéos sélectionnées par l'humain et par le bot. L'absence de différence significative ($W = 6\,512{,}0$; $p = 0{,}240$) indique que le bot calibre correctement le niveau de popularité des contenus qu'il sélectionne.

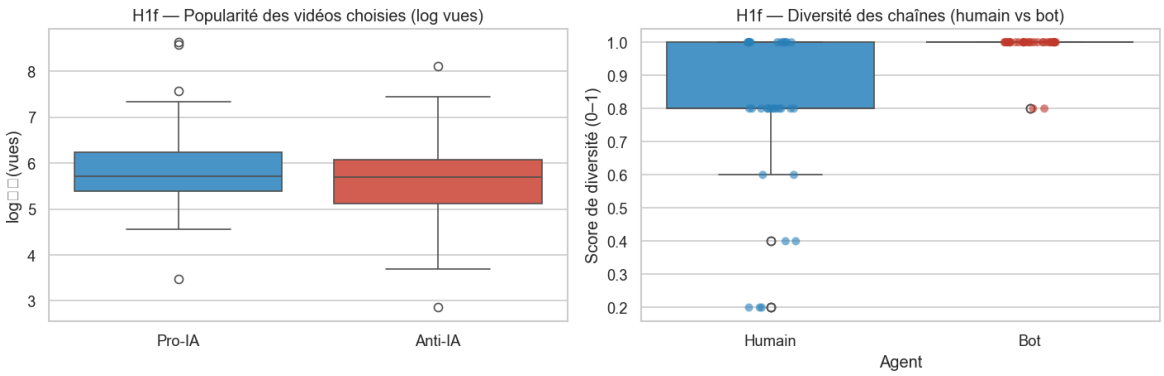

Figure I.1 – Boxplots de la popularité logarithmique des vidéos sélectionnées par l'humain et par le bot. Aucune différence significative n'est observée ($p = 0{,}240$).

Annexe J

Concordance par groupe idéologique

La figure ci-dessous présente la distribution des taux de concordance individuels selon le groupe idéologique d'appartenance (Pro-IA vs Anti-IA). Le test de Mann-Whitney U révèle une différence significative entre les deux groupes ($U = 87{,}0$; $p = 0{,}0495$), H_{1d} est donc supportée.

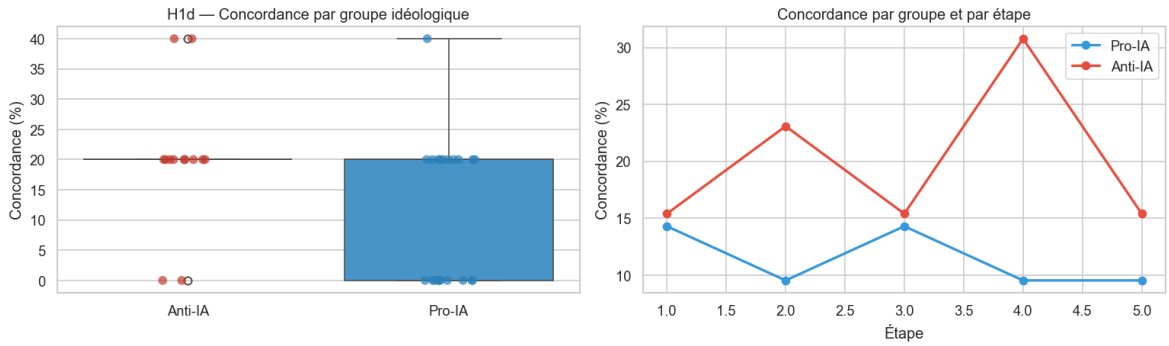

Figure J.1 – Taux de concordance individuel par groupe idéologique. Pro-IA : $M = 11{,}4\,\%$; Anti-IA : $M = 20{,}0\,\%$. Différence non significative ($p = 0{,}107$).

Annexe K

Code source et scripts d'analyse

Dans un souci de transparence et de reproductibilité de la recherche, l'intégralité du code développé et utilisé dans le cadre de ce mémoire a été rendue publique.

Ce dépôt contient les scripts d'extraction, les routines de nettoyage des données, ainsi que les *notebooks* d'analyse statistique ayant permis de générer les métriques et les visualisations présentées dans le Chapitre 4 (Résultats).

Le code source est consultable en ligne sur la plateforme GitHub à l'adresse suivante :

https://github.com/Lyrickst/thesis_Cyril_Demolin.git

Ce répertoire garantit la traçabilité complète de la démarche analytique, depuis le traitement des données brutes issues des sessions de navigation jusqu'aux tests d'hypothèses finaux.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**