Éthique du couplage Nudges-IA : enjeux épistémologiques et sociotechniques des Nudges IA dans les interactions humain-IA

Ethics of Nudges-AI coupling: Epistemological and Sociotechnical Issues of Nudges-AI in Human-AI Interactions

< Béa ARRUABARRENA [1] > < Anna NESVIJEVSKAIA [2] >

1. Laboratoire DICEN CNAM
beatrice.arruabarrena@lecnam.net

2. Laboratoire DICEN CNAM
anna.nesvijevskaia@lecnam.net

DOI : 10.25965/interfaces-numeriques.5208

< RÉSUMÉ >

Avec l'essor de l'économie comportementale, les Nudges se sont largement répandus dans divers aspects de la société et du numérique. Récemment, les progrès combinés des technologies de l'IA et des connaissances scientifiques, notamment en psychologie, en sciences humaines et sociales et en neurosciences, ont donné naissance à un nouveau phénomène appelé « l'Hypernudging ». Ce dernier se distingue par un changement d'échelle dans sa capacité à agir sur les décisions des individus. L'objectif de cette recherche, menée selon une approche socio-anthropologique, est d'avoir une meilleure compréhension des questions éthiques que soulève ce couplage des Nudges et de l'IA dans les interactions humain-IA. Dans cette perspective, nous avons réalisé une revue de littérature approfondie sur le sujet et mené une enquête exploratoire auprès d'un échantillon composé d'étudiants et de professionnels évoluant dans différents secteurs d'activités tels que la banque, la finance, le marketing ou les médias ainsi qu'auprès de deux agences de communication

spécialistes des Nudges. Il s'agissait d'examiner les enjeux épistémologiques et sociotechniques de l'Hypernudging et de les confronter au développement des Nudges IA dans les organisations pour en dégager des pistes de recherche pour leur régulation éthique.

< MOTS-CLÉS >

éthique, nudges, persuasion, hypernudging, IA

< ABSTRACT >

With the rise of behavioral economics, Nudges have become widespread in various aspects of society and in the digital world. Recently, the combined progress in AI technologies and scientific knowledge, particularly in psychology, social sciences and neuroscience, has given rise to a new phenomenon called "Hypernudging", characterized by a change of scale in its ability to influence individuals' decision-making. The objective of this research, based on a socio-anthropological approach, is to better understand the ethical issues raised by the coupling of Nudges and AI in human-AI interactions. To this end, we conducted a literature review on the subject and carried out an exploratory survey with a sample of students and professionals working in various fields, such as banking, finance, marketing or media, as well as with two Nudge-specialist communication agencies. The aim was to examine the epistemological and sociotechnical implications of the Hypernudging and to confront them with the development of AI Nudges in organizations, in order to identify research avenues for their ethical regulation.

< KEYWORDS >

ethics, nudges, persuasion, hypernudging, AI

1. Introduction

Avec l'essor mondial de l'économie comportementale (Thaler et Sunstein, 2009 ; Whitehead *et al.,* 2014), les Nudges se sont largement répandus dans le monde numérique. Conçus autour d'architectures de choix, leur finalité est d'agir sur les biais cognitifs afin d'influencer les comportements et d'orienter les décisions des individus. On les retrouve sous différentes formes, par exemple dans les interfaces par l'introduction de détails dans le contexte utilisateur afin de guider ses choix (Weinmann *et al.,* 2016), dans les moteurs de recherche et les systèmes de recommandations (Jesse et Jannach, 2021), dans les parcours d'achat ou encore dans les systèmes de gamification des applications de sport, de bien-être (*Quantified self*) et de santé (Arruabarrena, 2021), parfois associés à des objets connectés tels que les

traceurs d'activité de « Nike+ » ou de « Fitbit » conçus pour orienter de manière « douce et ludique » la gestion des pratiques de santé. Plus récemment, les avancées combinées des technologiques de l'Intelligence Artificielle (IA) et des connaissances scientifiques, en particulier en psychologie, en sciences humaines et sociales et en neurosciences (Varazzani, 2017), ont conduit à l'émergence d'un phénomène nouveau, couplant les Nudges à l'IA[1] : il s'agit de « l'*Hypernudging* » (Wagner, 2021 ; Yueng, 2017), dont la capacité à agir sur les décisions individuelles opère désormais un changement d'échelle.

L'éthique des Nudges a déjà fait l'objet de nombreux débats dans la littérature scientifique depuis une dizaine d'années, notamment sur leur potentiel manipulatoire et coercitif, mais le sujet a été ravivé à la suite du scandale engendré par l'affaire Cambridge Analytica. L'utilisation massive de *Dark nudges* (Campione, 2020) lors des élections américaines a ainsi permis, grâce à la collecte illégales de données de centaines de milliers de comptes Facebook, de faire du profilage et du microciblage comportemental afin d'influencer le vote des électeurs. L'impact visé est double : d'une part, faire gagner plus de voix à un candidat (Ibid., 2020), et, d'autre part, faire perdre des voix aux autres candidats par l'utilisation de méthodes de persuasion suscitant le dénigrement ou le dégout, comme ce fut le cas pour la campagne anti-Clinton réalisée par Cambridge Analytica (Kaiser, 2020). Cette affaire a mis en lumière les dérives possibles des Nudges couplés aux algorithmes dans la manipulation de l'information. Mais la littérature scientifique et académique pointe la nature même des Nudges, et plus largement le paradigme comportementaliste, dans sa conception ontologique de l'individu, de sa liberté, de son l'autonomie et de son raisonnement (Schmidt, 2020). Selon cette logique, le risque majeur des Nudges-IA est qu'à terme les décisions des individus soient entièrement automatisées selon des plans d'action

1 Selon la définition du Parlement européen : « L'IA désigne la possibilité pour une machine de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité. ». En plus de cette définition, nous utilisons ici l'IA au sens élargi comme un ensemble de techniques et de méthodes automatisées mises en œuvre par des algorithmes de type Machine Learning et basés sur l'exploitation des données massives.

prescriptifs, que ce soit dans le monde du travail, de la consommation, de la santé ou encore de la politique.

L'objectif de cette recherche, menée selon une approche socio-anthropologique, est d'avoir une meilleure compréhension des questions éthiques soulevées par le couplage des Nudges et de l'IA dans les interactions humain-IA. Pour ce faire, nous avons réalisé une revue de littérature riche sur le sujet, afin d'examiner les principaux enjeux épistémologiques et sociotechniques que soulève le couplage Nudges-IA. Nous avons ensuite réalisé une enquête exploratoire auprès d'un public d'étudiants et de professionnels œuvrant dans différents domaines d'activités tels que la banque, la finance, le marketing, les médias et la communication afin d'étudier les pratiques de Nudging IA. Il s'agissait à la fois de voir comment ces pratiques étaient présentes et conçues dans ces organisations, et de d'interroger nos interlocuteurs sur les problématiques éthiques qu'elles soulevaient pour en dégager des pistes de recherche pour leur régulation éthique. Cette recherche ne prétend pas enquêter sur les effets de Nudges-IA sur les individus (utilisateurs finaux de services). Ce type de recherche nécessiterait une étude de grande ampleur mixant méthodes quantitatives et qualitatives réalisée sur un temps plus long.

1.1. Enjeux épistémologiques du couplage Nudges-IA

Dans la revue de littérature, la principale controverse concernant les Nudges et, plus indirectement, le recours aux sciences comportementales porte sur l'atteinte à l'autonomie des personnes dans la prise de décision. Si les tenants des Nudges, comme Sunstein (2015), affirment que les Nudges rendent finalement les individus encore plus libres en les aidant à faire de « bons choix », d'autres chercheurs argumentent qu'ils constituent une entrave à l'autonomie des individus (Keoing, 2019 ; Bovens, 2009), y compris dans leur droit à faire de « mauvais choix », concernant par exemple leur santé (Sætra, 2019 ; Hurd, 2016). Le problème avec les Nudges, c'est qu'ils « annulent et contournent les capacités de prise de décision rationnelle de l'agent autonome » (Mils, 2013, p. 30). A terme, c'est donc la capacité même de développement de l'autonomie propre à chaque individu qui pourrait être altérée (Hausman et Welch, 2010, p. 129). Selon ces mêmes auteurs, le risque est que

« l'exploitation des faiblesses de la prise de décision [finissent] par diminuer les capacités de prise de décision autonomes des individus » (Ibid, p. 129) et qu'au final leurs actions « reflètent plus les tactiques de l'architecture du choix plutôt que celles de la capacité à évaluer par soi-même des alternatives et produire ses propres délibérations » (Ibid, p. 130). Sur ces questions, des recherches récentes sur l'usage des Nudges pour lutter contre la désinformation, préconisé par certains chercheurs (Sunstein, 2021 ; Konstantinou *et al.,* 2019), montrent que la priorité devrait être davantage mise sur le développement de l'esprit critique et que « les Nudges gagneraient à intégrer des dispositifs stimulant la réflexivité et l'ouverture des utilisateurs » (Piguet, 2023).

Sur la question de la rationalité, rappelons que le Nudging s'appuie sur les approches de psychologie comportementale développées dans les années 70 par Daniel Kahneman, qui postule, selon le principe de rationalité limitée, que les biais cognitifs[2] conduisent les individus à faire des choix jugés irrationnels, c'est-à-dire à prendre de mauvaises décisions. Selon l'approche de Kahneman (2012), le cerveau fonctionnerait selon deux systèmes de décision : le système 1, qui relève de la pensée automatique, rapide, instinctive, émotionnelle, et le système 2, plus lent, plus réfléchi et plus logique. Les deux systèmes agiraient différemment selon leur mode d'action sur le comportement et les biais cognitifs associés. Les architectures de choix mobilisent le système 1, celui des comportements automatiques correspondant aux mécanismes cognitifs inconscients de bas niveau. Elles ne tiennent pas compte des mécanismes cognitifs de haut niveau, liés au raisonnement et à la compréhension. Le corolaire d'une conception automatique de la raison est que le Nudging s'inscrit dans une épistémologie à la fois *behavioriste*, qui considère le comportement humain sous l'angle du conditionnement lié à ses seuls automatismes, et *cognitiviste*, qui constitue une conception analogique des interactions humain-IA, dont la spécificité est d'assimiler le fonctionnement humain à une machine. Pourtant le débat sur l'analogie entre le cerveau et la technologie est déjà ancien, comme lorsque le philosophe Miguel Benasayag rappelle que

2 Voir le catalogue du CEBM https://catalogofbias.org/ et le codex des biais cognitifs de John_Manoogian sur Wikipédia

« nous ne nous sommes pas des machines » (Benasayag, dans Ravet, 2017) :

> *« Ce n'est pas le cerveau qui pense : c'est le corps pensant, situé affectivement dans un milieu, […]. En aucune façon la pensée ne peut être comparée à un simple flux logico-informatique qui circule dans un logiciel. L'ordinateur, même dans le cas de ce qu'on appelle l'apprentissage profond (« Deep Learning ») qui lui permet d'incorporer de lui-même de nouvelles données, fonctionne de manière autoréférentielle, par feedback, sans échange ouvert avec le monde, pas même avec la table sur laquelle il est posé. » (Benasayag, dans Ravet, 2017, p. 30). »*

Dans une réflexion plus large des Nudges dans une économie de l'attention, Yves Citton explique à ce sujet que :

> *« Le leurre principal, dans toute cette affaire, consiste sans doute à opposer de façon trop rigide Système 1 et Système 2 (alors qu'ils interagissent incessamment) et à ne voir partout que des courts-circuits, alors que chaque court-circuitage en forme de feedback entraîne ailleurs des rallongements de circuits porteurs de nouvelles possibilités de feed-forward. » (Citton, 2017, p. 33). »*

Cette remarque a des prolongements du point de vue de l'anthropologie de la communication humaine. Les Nudges, en réduisant la communication à une logique à la fois mécaniste (automatisation de la communication et des feedbacks) et opaque, en agissant sur des mécanismes inconscients, ne tiennent pas compte de la complexité des situations de communication. Or, selon nos différents travaux réalisés sur l'utilisation des Nudges dans l'action publique lors de la crise COVID (Arruabarrena, 2022), les intentions dissimulées des Nudges sont à la source d'un mode de communication paradoxale qui se caractérise par des formes de contradiction dans la communication humaine dues à des écarts de niveaux de communication entre le message (contenu) et le méta-message (relation au message/intention). Cela rend la communication dysfonctionnelle et génératrice de stress, d'anxiété, voire de blocage psychologique, et inhibe potentiellement les possibilités de construire du sens et d'apprendre des situations. Les effets psychologiques des Nudges sont également pointés en termes d'« effort mental » (Citton, 2017 ; Loewenstein et O'Donoghue, 2006) imposé de

facto par les architectures de choix dans la mesure où elles privilégient le statut quo en proposant des choix par défaut, mais à l'inverse cela peut aussi entrainer dans certain cas une forme d'inertie de la part des individus cherchant de fait à limiter leurs efforts dans leur prise de décision.

1.2. Enjeux sociotechniques du couplage Nudges-IA

Si les Nudges ne constituent pas un phénomène nouveau, ils opèrent un changement d'échelle inédit lorsqu'ils sont couplés à l'IA. En effet, les coups de pouces se retrouvent » suralimentés » (Eggers *et al.,* 2017) : ils sont dopés par leurs couplages avec des algorithmes de Machine Learning et des données massives. Si les Nudges s'adressaient jusqu'à présent au grand public de manière générique, par exemple avec des affiches publicitaires, de la signalétique ou des images sur un paquet de cigarettes, « ils sont aujourd'hui remplacés par un travail des algorithmes qui s'exercent en profondeur de manière bien plus dynamique, discrète, contextuelle et individuelle » (Wagner, 2021). Ce qu'il convient d'appeler « l'*Hypernudging* » (Yeung, 2017) ou encore le « *Big Nudge* » (Helbing *et al.,* 2019) constitue des phénomènes extrêmement plus puissants en raison « de leur mise en réseau, de leur mise à jour continuelle, et de leur omniprésence » (Yeung, 2017, p. 23). Le pouvoir des Nudges s'en retrouve démultiplié, ce qui va de pair avec des risques de dérives, bien plus importants qu'avec les Nudges traditionnels (Eggers *et al.,* 2017).

Force est de constater que les approches comportementales et les techniques de l'IA se combinent de façon symbiotique dans une logique commune, cybernétique et cognitiviste, fondée sur l'automatisation et des boucles de rétroaction qui lui confèrent une efficacité (Becks et Weis, 2022). Pour autant, d'autres chercheurs nous mettent en garde quant à ces l'évolutions. Selon Guszcza (2015) et sa « théorie du dernier km » à terme, les sciences comportementales pourraient contribuer à faire avancer l'IA dans ce qu'elle ne peut pas faire actuellement (améliorer la précision et la qualification des données individuelles) et l'IA pourrait en retour aider les sciences comportementales à s'améliorer (hyperpersonnalisation des Nudges).

D'un point de vue sociotechnique, le couplage Nudges-IA soulève un certain nombre d'enjeux éthiques qui ont été exposés dans la littérature scientifique et académique :

- **Le consentement.** Les Nudges, en modifiant des comportements et la prise de décision de manière opaque, ne se prévalent pas d'un consentement éclairé (Hansen et Jesperson, 2013) permettant de garantir le droit des personnes et de préserver ainsi leur autonomie dans leur prise de décision (Barton et Grüne-Yanoff, 2015). Ce phénomène est amplifié par l'automatisation et la complexité des flux de données et des interfaces.
- **Les données personnelles.** Avec le recours aux données massives, disponibles et variées[3], la collecte de données ne respecte pas forcément la confidentialité, sachant que « l'anonymat ne garantit pas contre les possibilités de caractérisation des comportements des individus, ni contre l'analyse prédictive de ces comportements » (Rouvroy, 2016).
- **Le profilage et le microciblage.** L'association de l'IA et des Nudges permet un profilage détaillé des comportements individuels grâce à la collecte de données qui fournit une de compréhension fine dont ils pensent réellement, y compris lorsqu'il s'agit de découvrir des opinons, des goûts et des préférences que les individus ne soupçonnent parfois pas eux-mêmes. Cette connaissance offre la possibilité de cibler de façon beaucoup plus précise les individus pour orienter leur choix, et ce en temps réel.
- **L'hyperpersonnalisation des Nudges.** Les effets des Nudges peuvent être suivis au fur et à mesure et continuellement ajustés par des boucles de rétroaction automatisées grâce aux algorithmes de Machine Learning qui apprennent en continu afin d'« influencer […] les comportements en temps réel » (Zuboff, 2022, p. 25) pour chaque individu particulier. Cette

3 Les données collectées sont de toute nature : données des clients, des citoyens traités par le secteur public et les agences gouvernementales, des médias sociaux, de reconnaissance faciale issue des déploiements dans le domaine public, ou encore issues du suivi de localisation et des objets connectés.

personnalisation du Nudge devient automatisable à grande échelle, pour tous les individus.

- **La prescription comportementale automatisée.** Le profilage fin des comportements, réinvestie auprès des utilisateurs sous la forme d'analyses prédictives ou prescriptives à grande échelle, comporte le risque que les choix soient massivement et entièrement guidés selon des plans d'action prescriptifs, que ce soit dans le monde du travail, de la consommation, de la santé, ou encore de la politique (Schmidt, 2020).

Face à ces nombreuses questions éthiques, certains chercheurs plaident pour davantage de transparence, de réglementation et de gouvernance des Nudges (Raj, 2021), que ce soit dans les organisations privées ou publiques (Goodwin, 2012, p. 90). Mais la question de la transparence des Nudges, si elle a été largement débattue, ne fait pas consensus, car elle est confrontée à certains paradoxes du Big Data et des Nudges. En effet, comment pratiquer la transparence alors que la collecte d'informations dans les techniques de profilage reste souvent opaque (Richards et King (2013, p. 42) ? Et comment rendre transparents les Nudges alors que leur visibilité réduit, voire annule les effets escomptés (Mills, 2013, p. 31) ? Ainsi, des tentatives de régulation ont commencé à émerger depuis 2022. En Europe, de nombreux textes sont en vigueur (RGPD, DSA et DMA) ou en cours d'entrée en vigueur, comme l'IA Act[4], approuvé en mai 2023 et visant à interdire des techniques subliminales et manipulatoires dans les systèmes IA. Aux États-Unis, la Commission fédérale du commerce a soumis en février 2022 au Congrès un projet de loi appelé « Social Media NUDGE Act »[5] en vue de réaliser des études sur les plateformes de médias sociaux, notamment sur les questions de mise en avant de contenus dangereux en lien avec les algorithmes. Néanmoins, ce travail de régulation part du principe qu'il existerait de « bons » ou de « mauvais » Nudges, au risque de tendre à légitimer davantage leur usage

4 IA Act – 1ere version du Texte adopté le 11/05/2023 https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf

5 Nudging Users to Drive Good Experiences on Social media https://www.congress.gov/bill/117th-congress/senate-bill/3608/text?s=1&r=74

plutôt qu'à répondre aux enjeux épistémologiques fondamentaux qu'ils soulèvent.

2. Méthodologie de recherche

Cette recherche, menée selon une approche socio-anthropologique, s'inscrit dans la continuité de nos travaux antérieurs menées depuis 2014 dans le cadre de thèse et de publications scientifiques (Arruabarrena et Nesvijevskaia, 2014) sur les pratiques du *Quantified self* et le design des dispositifs de quantification dans le domaine de la santé, du bien-être et des activités physiques. Une quarantaine d'entretiens ont été effectués dans le cadre de ces travaux. Cette recherche trouve également des prolongements dans nos activités d'enseignement depuis 2017 auprès d'élèves de Master en Data Sciences et Data Analysts au CNAM Paris dont les métiers ont rapidement évolué.

Pour répondre à la question de recherche, à savoir avoir une meilleure compréhension des questions éthiques que posent le couplage des Nudges avec l'IA dans les organisations et la société, nous avons réalisé une étude exploratoire entre septembre 2022 et avril 2023. Nous avons effectué 17 entretiens menés auprès de 5 femmes et de 12 hommes âgés de 21 à 42 ans, en ciblant des professionnels dans les secteurs de la banque, de la finance et des médias, ainsi qu'auprès de 2 agences de communication spécialistes des Nudges. Parmi les personnes interrogées, nous avons rencontré des Data Scientists, des Data Analysts, des ingénieurs informaticiens, des designers UX, des professionnels du Marketing ainsi que des communicants, résidents en France et pour quelques-uns d'entre eux au Canada. Précisons que les étudiants interrogés ont tous une expérience professionnelle, soit parce qu'ils sont en alternance, soit parce qu'ils effectuent des stages régulièrement.

Notre méthodologie s'est appuyée sur une approche par observation et par entretiens semi-directifs. La grille d'entretiens a été établie selon la revue de littérature en dégageant des thèmes de relance principaux quant aux pratiques de Nudging dans les organisations et à leur positionnement éthique : les configurations dans la conception des Nudges IA ; les méthodes utilisées ; les techniques et outils utilisées ; les questions éthiques que posent les Nudges couplés à l'IA. Les entretiens

ont fait l'objet d'une analyse thématique permettant de classer les réponses des personnes interrogées en fonction des questionnements.

Figure 1 : Tableau de profil des interviewés

Secteur (entreprises privées)	Profil de poste	Nombre	Tranche d'âge
Banque, Médias, Marketing	Data Scientists dont 2 professionnels confirmés et 2 étudiants en alternance depuis plus d'1 an	4	20-30
Banque, finance et Médias	Data Analysts, dont 2 professionnels confirmés et 2 étudiants (1 en stage au Canada et 1 en alternance en France)	4	20-30
Banque, finance et Médias	Ingénieurs informatique, soit 3 professionnels confirmés	3	20-30
Médias et Agences de communication	Designers UX dont 1 étudiant ayant réalisé plusieurs stages et 1 professionnel confirmé	2	20-30
Médias	Chef de projet marketing et Directeur du marketing rattaché à la cellule Data Sciences de l'entreprise, soit 2 professionnels confirmés	2	30-45
Agences de communication	Consultant en communication et Directeur d'Agence	2	30-45

L'échantillon de cette phase exploratoire étant réduit, nous n'avons pas fait de comparaison entre la France et le Canada. Au niveau méthodologique, deux points de vigilance sont à prendre en compte pour notre enquête : 1/il nous a été parfois difficile d'accéder aux outils, aux données et aux algorithmes pour des raisons de confidentialité invoquées par nos interlocuteurs ; 2/l'enquête étant de nature qualitative, et les personnes interviewées étant choisies sur la base du volontariat, certains

biais de sélection et de représentativité sont à considérer. Ce travail exploratoire constitue une première étape de cette recherche que nous envisageons de prolonger par une étude de plus grande ampleur mixant méthodes quantitatives et qualitatives afin de renforcer la validité de nos résultats.

3. Résultats

3.1. La nécessaire pluridisciplinarité dans la conception des Nudges IA

Les résultats de cette enquête exploratoire confirment ceux de la littérature scientifique quant à la présence des pratiques de Nudges-IA dans les organisations. Les différents entretiens ont permis de mettre en évidence des pratiques bien établies de design autour des Nudges et des algorithmes, malgré des niveaux de maturité très différents selon le type d'organisation, leur domaine d'activité, leur taille et les moyens mis en œuvre pour s'approprier les méthodes de Nudging. Plusieurs configurations de conception et de développement des Nudges IA se dégagent. Elles ont toutes en commun de travailler en s'appuyant sur des équipes pluridisciplinaires que les responsables interrogés jugent absolument nécessaires pour une « conception complexe de qualité ». En effet, ce type de conception ne repose pas seulement sur « des éléments de communication et de design, mais sur une imbrication subtile de psychologie et d'informatique » (Directeur du marketing). En ce qui concerne le type d'organisation, on distingue tout d'abord celles dotées de services de Data Science constitués de Data Scientists, de Data Analysts, de designers UX, de managers de l'innovation et de professionnels du markéting. Ces experts travaillent étroitement à la conception des Nudges IA. Il s'agit d'entreprises du secteur bancaire et des médias dont la dimension commerciale est déterminante dans leurs activités. Dans cette configuration, l'équipe de conception est souvent pilotée par le service marketing qui a une très bonne maîtrise des enjeux stratégiques et commerciaux ainsi que du potentiel de l'Hypernudging. La seconde configuration concerne des organisations dont la dominante métier est informatique. Il s'agit d'entreprises qui développent des algorithmes en interne ou à l'intention d'autres entreprises dans des

domaines divers comme la finance et le commerce. Les ingénieurs, les Data Scientists et les Data Analysts collaborent alors sur la création d'algorithmes tout en s'appuyant sur un service de Recherche et Développement dont les spécialistes en sciences comportementales sont essentiels pour concevoir des Nudges algorithmiques très précis. La dernière configuration concerne les petites et moyennes entreprises qui externalisent la conception des Nudges IA auprès d'agences de communication spécialisées dans le domaine, assez nombreuses sur le marché. Dans cette configuration, les consultants d'agence travaillent conjointement avec le marketing et l'informatique, afin d'élaborer une stratégie de *Nudging* adaptée au déploiement de service de données, que ce soit pour des applications internes ou externes à l'organisation.

3.2. Au-delà des Nudges, une variété de méthodes non-triviales pour une conception peu documentée

Au niveau des méthodologies de conception des Nudges, les organisations qui sont pilotées par le marketing et les agences de communication spécialisées s'inspirent des méthodes de l'UX design qui s'appuient sur le design persuasif (Fogg, 2009), et utilisent des approches telles que le design thinking ou encore des Framework collaboratifs comme le Nudge deck. Il s'agit d'un outil d'aide à la conception de Nudges technologiques développé par Carabon (2020) basé sur un système de cartes d'idéation fournissant des connaissances préalables sur le type d'interventions possibles, telles que les triggers. Les organisations à caractère technique s'appuient davantage sur des Frameworks issus de l'informatique persuasive (Zarouali et al., 2022). Ainsi, il apparait que dans la conception des Nudges-IA, on ne se limite pas à construire des architectures de choix. Au contraire, les designers ont accès à une variété de méthodes d'intervention sur les biais cognitifs qui dépassent largement celles des Nudges, même si celles-ci sont toutes désignées sous l'appellation de Nudging. Une palette importante de méthodes persuasives permettant des incitations comportementales est ainsi utilisée dans la conception des algorithmes qui traitent des données massives : les triggers (déclencheurs), le tuning (ajustement), le herding (l'aiguillonnage), le conditionnement, etc. Toutes ces méthodes sont complétées au niveau des interfaces par des méthodes de Nudges traditionnels, qui consistent à utiliser des indices visuels tels que

des symboles (icones, couleurs ou emplacement) pour capter l'attention et déclencher un comportement spécifique. Il s'agit également de présenter les avantages d'une action dans un langage clair et concis, de mettre l'accent sur la preuve sociale par la comparaison, d'insister sur la popularité d'un choix spécifique, d'utiliser des éléments de gamification pour rendre une tâche plus engageante et plus agréable, et d'encourager la répétition des comportements. En conséquence, on observe une forte convergence de toutes ces méthodes vers ce qu'on peut appeler le champ du design comportemental.

Dans notre analyse, il ressort également que les Nudges[6] en tant que méthode non triviale, sont élaborées au cas par cas selon le projet. Or, même si la démarche globale de conception via les Frameworks peut être consignée dans des documents de travail et que les modèles ainsi implémentés peuvent être capitalisés et réutilisés pour des développements ultérieurs d'autres services, dans de nombreux cas, le travail préliminaire d'élaboration détaillé d'un Nudge IA est très peu documenté. La non-trivialité des méthodes rend difficiles, voire impossibles la lisibilité et la traçabilité des méthodes constitutives du Nudge utilisé pour agir sur un biais une fois que celui-ci est implémenté. Ce manque de documentation est d'autant plus problématique qu'avec l'hyperpersonnalisation rendue possible par un suivi automatisé, les Nudges se complexifient dans la durée par l'apprentissage automatique qui opère des ajustements itératifs en fonction des données recueillies, comme c'est le cas pour le microciblage. Selon un des professionnels du marketing, c'est cette combinaison qui garantit une efficacité et une pertinence additionnelle quant à la connaissance client, dans la mesure où les attentes et les comportements des utilisateurs évoluent constamment. Les Nudges-IA font ensuite l'objet de réglages par monitoring, à commencer par des tests utilisateurs et ensuite ils sont automatisés par l'analyse comportementale effectuée en continu, afin de guider de façon personnalisée l'utilisateur final dans son expérience utilisateur, par exemple pour choisir un hôtel en fonction d'une promotion, pour visionner un film en fonction de son historique de films, ou encore pour sélectionner un service bancaire en fonction de ses

6 An Overview of the Various Types of Nudges - https://www.membershipinnovation.com/insights-and-ideas/an-overview-of-the-various-types-of-nudges

projets d'achat immobilier. Mais, comme le souligne un des Ingénieurs rencontré :

> *« Le problème, c'est qu'une fois qu'on met des Nudges dans les algorithmes, ils s'améliorent automatiquement par couches successives grâce au machine Learning, au point de ne plus pouvoir identifiés les différents réglages effectués. Il nous manque des briques pour tout évaluer, mais finalement c'est super efficace et ça marche très bien, pour nous c'est le plus important » (Ingénieur informaticien – Banque-Finance).*

Le risque ici est bien que les Nudges couplés à l'IA deviennent des boîtes noires qui fonctionnent par elles-mêmes et qui se complexifient avec le temps au fur et à mesure de leur entraînement. En ce sens, les systèmes de conception gagneraient sur le plan éthique à mettre en œuvre des méthodes d'explicabilité telles qu'elles se déploient aujourd'hui pour la transparence des algorithmes afin de rendre ce nouveau type d'ingénierie des Nudges IA compréhensible à l'humain.

3.3. Des risques éthiques souvent mal appréhendés face aux tensions entre innovation et régulation

Sur les enjeux éthiques liés aux Nudges-IA, nos interlocuteurs s'accordent pour expliquer que leur utilisation soulève un certain nombre de préoccupations éthiques, notamment celles concernant la collecte et la protection des données ainsi que l'utilisation des biais cognitifs pour influencer voire manipuler le comportement des utilisateurs. En ce sens, il est souvent souligné par les personnes rencontrées qu'il est important que les designers prennent en compte ces préoccupations, et qu'ils les utilisent de manière éthique et responsable, dans le respect de l'autonomie de l'utilisateur. Pourtant, il ressort bien des entretiens que très peu de designers sont formés aux sciences comportementales : leurs compétences se situent davantage au niveau opérationnel dans l'utilisation de méthodes appliquées de conception et, pour la plupart, leurs connaissances ne s'appuient sur aucun guide de bonnes pratiques sur les limites des Nudges, mis à part pour certains d'entre eux qui ont lu quelques articles académiques sur le sujet. Néanmoins la responsabilité attribuée aux designers et aux Data Scientists sur les questions éthiques

que soulèvent les Nudges est souvent mise en avant au détriment d'une responsabilisation collective de l'organisation.

Paradoxalement, quand on rentre dans un questionnement plus concret sur l'éthique quant aux risques associés à l'utilisation conjointe des Nudges et de l'IA avec les risques de dérives de l'Hperpersonnalisation, la minimisation des risques liés à l'usage des Nudges est largement répandue, quand bien même on reconnaît que leurs limites sont très difficiles à cerner. Seuls les *Darks Nudges* ou les *Sludges* sont considérés comme véritablement dangereux, notamment sur le plan politique et celui de la désinformation. Ce discours paradoxal se retrouve également dans les discussions concernant la collecte de données visant à faire du profilage et du microciblage automatisés. Dans le même temps, nos interlocuteurs mettent en avant leur attachement au respect du RGPD et leur besoin de collecter des données de différentes natures, par exemple :

> *« Pour les données personnelles nominatives, on respecte le RGDP, dans le cas du profilage et du microciblage, on est obligé de collecter différentes natures de données sur les goûts et les préférences des personnes, les données clients, les données des réseaux sociaux, les données Analytics. Pour les données indirectes, on part du principe que si un client nous a donné son consentement pour ses données clients, on peut exploiter d'autres données en lien avec leurs préférences dès l'instant où ça reste anonyme. A partir de là, nos algorithmes sont capables de mieux cibler et de proposer des offres qui correspondent mieux à ce que les gens veulent ». (Data Scientist – Banque).*

Globalement si les questions éthiques sont présentes dans les organisations, il existe une véritable tension dans les discours entre l'innovation et la régulation. Les organisations tentent à la fois de se saisir de l'innovation technologique que les Nudges IA représentent en tant que méthodes opérationnelles garantissant la performance, l'efficacité et la rentabilité, puisqu'elles sont peu onéreuses à mettre en place, et de satisfaire à la réglementation en répondant aux exigences minimales d'ordre juridique que ces méthodes impliquent, comme le souligne l'un des experts du marketing interrogé :

> *« Il est difficile de tenir les deux. La difficulté de l'éthique avec l'IA, c'est de comprendre les bonnes limites dans un domaine qui évolue très vite et qui est très concurrentiel. Au quotidien, on n'a pas forcément le temps ni les moyens de se poser la question de l'éthique à chaque modification d'un Nudge ou d'un algorithme, car cela demande un investissement énorme et parfois on est dépassé, on n'a pas toujours la connaissance à temps. Nous, ce qu'on essaie de faire c'est de couvrir au maximum les questions réglementaires et pour le reste on ajuste dès qu'il y a de nouvelles règles. » (Chef de projet marketing médias).*

Sur les enjeux éthiques du couplage des Nudges et de l'IA plus spécifiques, tels que le changement d'échelle et l'automatisation de la prise de décision à l'aide d'algorithmes, les témoignages des Agences de communication spécialisées ont confirmé qu'il ne s'agit pas toujours d'un sujet sensible pour les organisations. En effet, les entreprises n'ont pas toujours le temps et les moyens de s'approprier les mécanismes sous-jacents de l'IA et de comprendre les enjeux éthiques qu'elle soulève, comme c'est le cas aujourd'hui avec la déferlante de l'IA générative telle que Chat-GPT. Toutefois, ces agences ont bien compris le changement d'échelle qu'impliquait le couplage des Nudges et de l'IA :

> *« Avant, on faisait du design au cas par cas, maintenant on est en train d'industrialiser et d'augmenter l'efficacité des Nudges grâce aux algorithmes et aux datas. Nous, on essaie d'apporter notre expertise éthique, mais cela demande du temps, et les entreprises ne nous le donnent pas nécessairement. » (Consultant en communication - Agence Nudges)*

4. Discussion-Conclusion

Dans le cadre de cette recherche, nous avons tenté de faire ressortir les questions éthiques liées au couplage des Nudges et de l'IA en particulier dans leur conception dans les organisations. Si cette étude exploratoire nous a permis de voir la présence de pratiques de Nudges-IA bien installées dans les organisations, il reste néanmoins que les risques éthiques liés notamment au changement d'échelle restent très inégalement appréhendés. Pourtant le couplage Nudges-IA n'est pas neutre, il comporte de nombreux risques pour les individus et la société.

En effet, il impose un mode de raisonnement automatique dans les processus cognitifs, au détriment de ceux impliqués dans la compréhension. Cette généralisation pourrait entraver l'autonomie des personnes et, à terme, mener à des systèmes où la prescription comportementale automatisée par des algorithmes deviendrait la norme. Ce couplage affecte également l'attention et les modes de communication des individus qu'il peut rendre dysfonctionnel. En tant que technique d'influence, comme le montrent certains chercheurs en politique, le pouvoir de Nudges est important (Schmidt, 2017) : ils peuvent exercer un impact significatif sur les populations les plus vulnérables (selon leur niveau de connaissances, d'éducation, leur statut social ou encore leur handicap), ce qui augmente les inégalités sociales (Puaschunder, 2020). Ainsi, même si ce paternalisme libertarien prétend avoir une visée bienveillante sans régulation, il comporte des risques qu'il est impossible d'ignorer (Arruabarrena, 2021, 2022 ; Peeters et Shuilenburg, 2017). Les Nudges basés sur l'intelligence artificielle constituent ainsi une manifestation de biopouvoir (Foucault, 1079) algorithmique qui, en tant que stratégie d'influence, intègre et utilise les invariants biologiques et cognitifs de l'être humain afin de contrôler ses décisions et ses comportements.

Les technologies transforment la société, ce qui leur confère un rôle déterminant dans la construction de nos sociétés contemporaines. Or, le couplage des Nudges et de l'IA, tel qu'il est mis en œuvre actuellement selon une logique cognitiviste, favorise davantage la technique au détriment de l'humain et de la société. Selon nous, ce couplage qui atteint une mutation anthropologique inédite dans les interactions humain-IA, constitue une forme « d'aliénation machinique » basé sur un mauvais couplage entre l'homme et la machine (Simondon, 1958). Il y a donc une véritable réflexion à mener sur l'usage des Nudges dans les interactions humain-IA, c'est-à-dire sur les combinaisons entre technologies et procédés psychologiques dans la prise de décision automatisée. Plus largement, il s'agit de questionner la pertinence de l'exclusivité du design comportemental dans le couplage avec l'IA afin de réfléchir au bon dosage d'usages embarquant l'IA restituée sous la forme de Nudges, et de les conjuguer avec d'autres méthodes relevant de paradigmes tels que le constructivisme, le pragmatisme, etc. où la compréhension se trouve au cœur de la prise de décision. Ce questionnement doit alors être élargi bien

au-delà des usages où la compréhension est indispensable pour la prise de décision pour des raisons opérationnelles ou réglementaires (Nesvijevskaia *et al.,* 2021).

Dans cette perspective, et au regard du développement actuel des Nudges-IA dans les organisations dont les enjeux éthiques sont parfois mal perçus, il est crucial que ce type de couplage fasse l'objet de recherches fondamentales plus avancées visant à permettre une meilleure compréhension des mécanismes sous-jacents de ces interventions. Des programmes de recherche pourraient ainsi s'adosser aux avancées législatives comme l'IA Act qui interdit désormais l'utilisation de techniques subliminales et manipulatoires dans les systèmes IA. Ils contribueraient au développement d'outils d'évaluation des Nudges (et plus largement du design comportemental) et à l'identification des risques associés pour les individus. De telles pistes de recherches peuvent déjà être envisagées à l'instar de ce qui se fait pour l'explicabilité des algorithmes dans la décision par apprentissage automatique en termes de transparence, d'auditabilité, et de documentation (Nesvijevskaia, 2021) et de responsabilité. Il s'agit également de développer des pédagogies de formation et de sensibilisation à destination du public, en particulier des designers, des Data Scientists et des organisations professionnelles qui semblent parfois dépassées par les enjeux éthiques sous-jacents aux couplages de plus en plus complexes entre les technologies et la science.

Bibliographie

Arruabarrena Béa (2022). Gouverner par la communication comportementale en temps de crise sanitaire : L'argument comportementaliste comme nouveau paradigme de l'action publique en santé. *Approches Théoriques en Information-Communication (ATIC)* 5.2 : 79-90.

Arruabarrena Béa (2021). Objets connectés : penser les enjeux des technologies connectées sous l'angle de la médiation info-communicationnelle. *Tic&société*, vol. 15, N° 1-2.

Arruabarrena Béa et Nesvijevskaia Anna (2014). Quantified Self & Big Data : quelles implications dans les relations usagers et assureurs en santé ? Dans : *Penser Les Techniques et Les Technologies : Apports Des Sciences de l'Information et de La Communication et Perspectives de Recherches*. Présenté au XIXème Congrès de la Sfsic, Toulon.

Barton Adrien et Grüne-Yanoff Till (2015). From libertarian paternalism to nudging—and beyond. *Review of Philosophy and psychology*. 341-359.

Becks Eileen et Weis Torben (2022). Nudging to Improve Human-AI Symbiosis. *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, Pisa, Italy, pp. 132-133.

Bovens Luc (2009). The ethics of nudge. *Preference change: Approaches from philosophy, economics and psychology*. Dordrecht: Springer Netherlands. 207-219

Campione Chiara (2020). *The dark nudge era: Cambridge analytica, digital manipulation in politics, and the fragmentation of society*. Tesi di Laurea in Nudging: behavioral insights for regulation and public policy, Luiss Guido Carli, relatore Giacomo Sillari, pp. 55. [Bachelor's Degree Thesis]

Caraban Ana et al. (2020). The nudge deck: A design support tool for technology-mediated nudging. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*.

Citton Yves (2017). Le court-circuitage néolibéral des volontés & des attentions. *Multitudes*. 21-34.

Eggers William D. et al. (2017). *How Government Data Can Supercharge the Nudge* (21 juillet 2017). En ligne : https://www.governing.com/archive/col-government-data-behavioral-science-nudge-impact.html

Foucault Michel (1979). *Naissance de la biopolitique*. Cours au Collège de France 1978-1979, Gallimard, 368 p. Edition 2004.

Fogg Brian J. (2009). A behavior model for persuasive design. *Proceedings of the 4th international Conference on Persuasive Technology*.

Goodwin Tom (2012). Why we should reject nudge. *Politics* 32 (2). 85–92.

Guszcza Jim (2015). *The last-mile problem: How data science and behavioral science can work together*. Deloitte Review Issue 16. (27 January 2015). En ligne: https://www2.deloitte.com/insights/us/en/deloitte-review/issue-16/behavioral-economics-predictive- analytics.html.

Hausman Daniel M. et Welch Brynn (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy* 18.1 (2010): 123-136.

Helbing Dirk et al. (2019). Will democracy survive big data and artificial intelligence?. *Towards digital enlightenment: Essays on the dark and light sides of the digital revolution*. 73-98.

Hurd Heidi (2016). Fudging Nudging: Why 'Libertarian Paternalism' is the Contradiction It Claims It's Not. *Georgetown Journal of Law and Public Policy*, vol. 14, p. 703-734.

Jesse Mathias et Jannach Dietmar (2021). Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, *3*, 100052.

Kaiser Brittany (2020). *L'affaire Cambridge Analytica*. HarperCollins.

Kahneman Daniel (2011). *Thinking, fast and slow.* London: Penguin Books.

Koenig Gaspard (2019). *La fin de l'individu : voyage d'un philosophe au pays de l'intelligence artificielle*. Éditions de l'Observatoire.

Konstantinou Loukas et al., (2019). Combating misinformation through nudging. *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17*. Springer International Publishing.

Loewenstein George et O'Donoghue Ted (2006). We can do this the easy way or the hard way: Negative emotions, self-regulation, and the law.*The University of Chicago Law Review* 73.1.183-206.

Mills Chris (2013). Why nudges matter: a reply to Goodwin. *Politics* 33 (1).28–36.

Nesvijevskaia Anna, 2021. DATABOOK: a standardised framework for dynamic documentation of algorithm design during Data Science projects. *IASSIST Quarterly* 45.

Nesvijevskaia Anna, Ouillade Sophie, Guilmin Pauline, Zucker Jean-Daniel, 2021. The accuracy versus interpretability trade-off in fraud detection model. Cambridge University Press, *Data & Policy* 3.

Peeters Rik et Schuilenburg Marc (2017). The birth of mindpolitics: understanding nudging in public health policy. *Social Theory & Health* 15. 138-159.

Piguet Jean-Gabriel (2023). Nudges, désinformation et autonomie citoyenne. Une critique de Sunstein. *Éthique publique* [En ligne], vol. 24, n° 2.

Puaschunder Julia (2022). Artificial Intelligence and Nudging. *Advances in Behavioral Economics and Finance Leadership: Strategic Leadership, Wise Followership and Conscientious Usership in the Digital Century*. Cham: Springer International Publishing. 133-196.

Raj Vijay (2021). *The Ethics of Nudge : Towards a governance structure for the ethical use of nudge theory by Governments*. EN ligne : https://osf.io/q79ku/download.

Ravet Jean-Claude (2017). Nous ne sommes pas des machines : entrevue avec Miguel Benasayag. *Relations* 792 (2017): 30-33.

Richards Neil M. et King Jonathan H. (2013). Three paradoxes of big data. *Stan. L. Rev. Online* 66 : 41.

Rouvroy Antoinette (2016). Des données et des hommes : droits et libertés fondamentaux dans un monde de données massives. *Rapport Conseil de l'Europe*. https://rm.coe.int/16806b1659

Sætra Henrik Skaug (2019). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society* 59 : 101130.

Simondon Gilbert (1958). *Du mode d'existence des objets techniques*. Paris, Aubier, p. 23-36

Schmidt Andreas T. et Engelen Bart T. (2020). The ethics of nudging: An overview. *Philosophy Compass*. 2020; 15:e12658.

Schmidt Andreas T. (2017). The power to nudge. *American Political Science Review* 111.2: 404-417.

Sunstein Cass R. (2015). The ethics of nudging. *Yale Journal on Regulation*, vol. 32, nº 2, p. 413-450.

Sunstein Cass R. (2021). *Liars. Falsehood and Free Speech in an Age of Deception,* New York, Oxford University Press.

Thaler Richard H. et Sunstein Cass. R. (2009). Nudge: Improving decisions about health, wealth, and *happiness*. Penguin.

Varazzani Chiara (2017). The Brains Behind Behavioral Science. *Behavioral Scientist.* (22 May 2017). https://behavioralscientist.org/brains-behind-behavioral-science/

Wagner Dirk. (2021). On the emergence and design of AI nudging: the gentle big brother?. *ROBONOMICS: The Journal of the Automated Economy*, *2*, 18. Retrieved from https://journal.robonomics.science/index.php/rj/article/view/18

Weinmann Markus et al. (2016). Digital nudging.*Business & Information Systems Engineering,* 58 433-436.

Whitehead Mark et al. (2014). *Nudging all over the world: Assessing the impacts of the behavioural sciences on public policy.* ESRC Negotiating Neuroliberalism Project Report. http://changingbehaviours.wordpress.com

Yeung Karen (2017). "Hypernudge": big Data as a mode of regulation by design, *Inf. Commun. Soc.* 20 (1): 118–136.

Zarouali Brahim, et al. (2022). The algorithmic persuasion framework in online communication: conceptualization and a future research agenda. *Internet Research* 32.4 : 1076-1096

Zuboff Shoshana (2022). *L'âge du capitalisme de surveillance*. Paris, France : Zulma.