Comité national pilote d'éthique du numérique (CNPEN)

Les enjeux éthiques des agents conversationnels

Appel à contributions

~~**ouvert jusqu'au 30 septembre 2020 à minuit**~~

Compte tenu de l'intérêt suscité par cet appel à contributions, la date limite de réception des réponses est reportée au

31 octobre 2020 à minuit

Envoi des réponses à l'adresse cnpen-consultation-chatbots@ccne.fr

Le Comité national pilote d'éthique du numérique (CNPEN) a été créé en décembre 2019 à la demande du Premier ministre. Constitué de 27 membres, ce comité réunit des spécialistes du numérique, des philosophes, des médecins, des juristes et des membres de la société civile. L'une des trois saisines soumises par le Premier ministre au CNPEN concerne les enjeux éthiques des agents conversationnels, appelés communément *chatbots*, qui communiquent avec l'utilisateur humain par la voix ou par écrit. Ce travail du CNPEN vient en prolongation des travaux initiés par la CERNA, Commission d'éthique de la recherche en sciences et technologies du numérique de l'alliance Allistene.

Dans ce document, nous sollicitons l'avis des lecteurs en posant des questions sur les enjeux éthiques liés aux chatbots. Chacun est invité à répondre soit à quelques questions de son choix soit à l'ensemble des questions posées.

Répondez-vous à ce questionnaire :

- *À titre personnel (préciser vos nom et prénom si vous le souhaitez)*
- *Au titre de vos activités professionnelles ou au nom d'une organisation :*
  - *Chercheur ou Institut de recherche (préciser le nom de votre institution)*
  - *Société ou groupe de sociétés (préciser laquelle)*
  - *Association de consommateurs ou assimilé (préciser laquelle)*
  - *Autorité publique (préciser laquelle)*
  - *Consultant professionnel*
  - *Think thank (préciser lequel)*
  - *Autre :*

Objectifs de ce document :

Le Comité national pilote d'éthique du numérique (CNPEN), créé en décembre 2019 sous l'égide du CCNE pour les sciences de la vie et de la santé, a été saisi par le Premier ministre pour élaborer en particulier un avis sur les enjeux éthiques des agents conversationnels (*chatbots*). Il a aussi dans ses objectifs « d'engager une discussion collective pour développer une approche partagée des innovations présentes et futures. Cette dimension est fondamentale pour s'assurer que la technique et l'innovation continuent à servir le bien commun. ». C'est pourquoi le Comité engage une consultation des parties prenantes et des citoyens avec pour objectifs de les sensibiliser aux enjeux éthiques et d'enrichir sa réflexion.

Utilisation et protection de vos données personelles :

Les données personnelles demandées (*adresse mail, nom, prénom, profession, institution de rattachement*) ou celles que vous pourriez fournir spontanément dans votre réponse au questionnaire ne seront traitées que si elles sont utiles à l'analyse et à la réflexion du comité. Toutes les données personnelles récoltées seront conservées sur les serveurs du CCNE ou de ses prestataires. Elles seront traitées de manière confidentielle uniquement par le personnel du CNPEN ou les membres du groupe de travail du CNPEN sur les agents conversationnels ; elles ne seront pas traitées de manière automatisées. Elles seront conservées au maximum dix-huit mois après la clôture de la consultation et jusqu'à douze mois après la publication de l'avis du comité.

Les résultats de cette analyse nourriront l'avis du comité sur les agents conversationnels, qui sera rendu public. Les contributions n'y seront pas citées nommément sans l'accord explicite de leurs auteurs.

Dans les conditions définies par la Loi Informatique et Libertés du 6 janvier 1978 et par le Règlement Européen sur la Protection des Données Personnelles entré en vigueur le 25 mai 2018, chaque contributeur bénéficie d'un droit d'accès aux données le concernant, de rectification, d'interrogation, de limitation, de portabilité et d'effacement. Chaque contributeur peut également, pour des motifs légitimes, s'opposer au traitement de ses données personnelles. Le contributeur peut exercer l'ensemble des droits mentionnés ci-dessus en s'adressant au CNPEN à l'adresse : cnpen-consultation-chatbots@ccne.fr.

INTRODUCTION

Qu'est-ce qu'un agent conversationnel ?

Un agent conversationnel, appelé communément chatbot, est un programme informatique qui interagit avec son utilisateur dans la langue naturelle de celui-ci. Ces termes regroupent tant les agents vocaux que les chatbots écrits.

Le plus souvent, un agent conversationnel ne constitue pas une entité indépendante mais il est intégré au sein d'un système ou d'une plate-forme numérique, comme un smartphone ou une enceinte vocale[1]. Sur le plan de l'apparence visuelle, les chatbots peuvent aussi être intégrés à un agent conversationnel animé, représenté en deux ou trois dimensions sur un écran, voire faire partie d'un robot social, y compris humanoïde. Le dialogue avec l'utilisateur ne représente alors qu'une des fonctions du système global.

L'histoire des agents conversationnels prend ses origines dans le jeu de l'imitation d'Alan Turing[2]. La compréhension du langage intéresse Turing dans la mesure où elle se manifeste à travers des réponses qui paraissent intelligibles et sensées à un examinateur (« test de Turing »). Dès 1991, un concours annuel est organisé afin de soutenir le développement de chatbots capables de passer le test de Turing.

Le premier agent conversationnel de l'histoire de l'informatique est le programme ELIZA de Joseph Weizenbaum[3], qui est aussi l'un des premiers leurres conversationnels. ELIZA simule un dialogue écrit avec un psychothérapeute rogérien en reformulant tout simplement la plupart des répliques de l'utilisateur « patient » sous forme de questions. Aujourd'hui, l'expression « effet ELIZA » désigne la tendance à assimiler de manière inconsciente le dialogue avec un ordinateur à celui avec un être humain.

D'un point de vue technique, comment ça marche ?

La conception et le fonctionnement d'un agent conversationnel se divisent en plusieurs modules de traitement automatique du langage naturel (TALN) : schématiquement, un chatbot peut inclure des modules de reconnaissance de la parole (pour les agents conversationnels vocaux), de traitement sémantique (hors et en contexte), de gestion de l'historique du dialogue, de gestion des stratégies de dialogue, d'accès aux ontologies, de gestion des accès aux connaissances externes (base de données ou internet), de génération de langage et de synthèse de la parole (pour les agents conversationnels vocaux).

Un agent conversationnel suit des règles décidées et transposées en code par des concepteurs humains ou obtenues par apprentissage. Les chatbots apprenants, par exemple XiaoIce de Microsoft Chine[4], sont aujourd'hui encore assez rares parmi les applications commercialisées,

---

[1] "Google Assistant", "Google Home", "Apple Siri", "Amazon Alexa" et "Amazon Echo", "Yandex Alisa", "Mail.ru Marusia", "Baidu DuerOS", "Xiaomi XiaoAI", "Tencent Xiaowei", "Samsung Bixbi", "Orange Djingo", etc.

[2] A. Turing, "Computing Machinery and Intelligence", *Mind* 59(236) 433–460, 1950.

[3] J. Weizenbaum, "ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine", *Communications of the Association for Computing Machinery* 9, 36-45, 1966.

[4] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum, "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot", *Computational Linguistics* 46(1), 53-93, 2020.

mais leur proportion n'aura de cesse de croître avec l'avancement de la maîtrise de cette technologie.

Ces dernières années, développer soi-même un chatbot rudimentaire ou dédié à une seule tâche est devenu relativement facile grâce à la disponibilité de nombreux outils de conception, comme "LiveEngage", "Chatbot builder", "Passage.ai", "Plato Research Dialogue System", etc.

Quelques défis de recherche concernant la conception des agents conversationnels

- Apprendre de manière adaptative en faisant évoluer la base de connaissances en cours d'utilisation.
- Être capable de converser librement sur des sujets génériques.
- Saisir le « sens commun », le caractère ironique ou le sens au « second degré » d'un énoncé
- Mettre en place une stratégie de dialogue.
- Détecter les émotions et les intentions de l'utilisateur.

Quelques défis de recherche concernant la compréhension des capacités des agents conversationnels par les utilisateurs

- Quelles données les chatbots enregistrent-t-ils ? Sont-elles anonymisées ?
- Comment peut-on mener des audits du comportement des chatbots (mesure automatique ou/et évaluation humaine) ?
- Les répliques sélectionnées par les chatbots sont-elles explicables ? Les chatbots peuvent-ils les rendre eux-mêmes plus compréhensibles ?
- Quels paramètres du profil de son interlocuteur les chatbots calculent-ils ? Les humains en sont-ils conscients ?
- L'idée que l'utilisateur se fait de la stratégie du chatbot correspond-elle à la stratégie réelle mise en place dans le chatbot ?

Questions éthiques

Le langage est un élément constitutif de l'identité de l'être humain et le fondement de sa vie en société. Les agents conversationnels sont ainsi naturellement comparés à un être humain, que leur interlocuteur soit informé, ou non, de leur caractère artificiel. Cet aspect naturel du dialogue est susceptible d'influer sur l'être humain : c'est le problème fondamental de l'éthique des chatbots. Leur déploiement étant un phénomène récent, on ne dispose pas de données expérimentales suffisantes pour évaluer leurs effets sur l'être humain à long terme.

Depuis peu, les performances de la reconnaissance de la parole permettent l'utilisation des interfaces vocales. Outre le dialogue langagier, la voix porte des informations de diverses natures, par exemple sur l'âge, le sexe, la corpulence, la langue maternelle, l'accent, les lieux de vie, le milieu socio-culturel, l'éducation, l'état de santé, la compréhension ou les émotions de la personne qui parle. De nombreuses questions éthiques sont liées à ces aspects de la vie humaine.

À l'instar des systèmes techniques en général et des systèmes autonomes en particulier (par exemple, la reconnaissance automatique d'images ou la conduite autonome des véhicules), les agents conversationnels doivent répondre à un grand nombre d'exigences en termes de sécurité, transparence, traçabilité, utilité, protection de la vie privée, etc. Les systèmes de chaque type

mettent ces propriétés en œuvre en fonction du contexte spécifique de leur utilisation. Dans tous les cas, il s'agit de contraintes de premier plan pour le concepteur comme pour l'utilisateur.

Certains agents conversationnels provoquent des tensions éthiques nouvelles, par exemple liées à l'impossibilité d'expliquer en langue naturelle la chaîne des décisions aboutissant à telle ou telle recommandation médicale. Des préconisations sont formulées à cet égard dans l'avis de la CERNA sur les questions éthiques de la recherche en apprentissage machine[5].

[5] http://cerna-ethics-allistene.org/Publications%2bCERNA/apprentissage/index.html

CONSULTATION

I. Les facteurs éthiques dans l'utilisation des chatbots

1) Confusion de statut. Plusieurs facteurs contribuent à faire confondre un agent conversationnel avec un être humain. Un effacement des distinctions de statut peut advenir comme une brève illusion ou, au contraire, il peut persister tout au long d'un dialogue. Il peut également être volontaire ou spontané, avoir des conséquences psychologiques ou juridiques, donner lieu à des manipulations plus ou moins graves. Cette confusion de statut a pour cause un phénomène plus général.

Quelle que soit la nature de son interlocuteur, l'être humain projette sur lui spontanément des traits humains : pensée, volonté, désir, conscience, représentation interne du monde. Ce comportement est qualifié d'« anthropomorphisme ». L'interlocuteur apparaît alors comme un individu autonome doté de pensée propre, qu'il exprime à travers sa parole.

A ce jour, seule une loi de l'État de Californie[6] impose explicitement de mentionner l'existence d'une interaction avec un chatbot lorsque cette interaction entend inciter à l'achat ou vendre des produits ou services dans le cadre d'une transaction commerciale ou influencer le vote dans un cadre électoral. Il n'existe pas d'équivalent à cette disposition dans le droit français ou européen même si une réflexion est désormais engagée sur ce point[7].

1.1 Faut-il informer l'utilisateur de la nature de son interlocuteur (être humain ou machine) ? Si oui, quelles informations sur le chatbot faut-il communiquer à l'utilisateur (finalités, corpus d'entraînement, nom du concepteur, etc.) ?

1.2. Pensez-vous qu'en Europe, il faudrait adopter un cadre législatif comparable à celui de l'État de Californie ?

1.2 Remarque libre :

2) Attribution de nom propre. Souvent, l'être humain donne à un agent conversationnel un nom, comme par exemple les enfants le font avec leurs poupées.

Parfois, l'attribution du nom est voulue par le concepteur : s'adresser à la machine par un nom peut aider à mieux réaliser sa fonction, par exemple dans les secteurs d'assistance aux personnes ou de divertissement. Dans ces cas, l'utilisation du nom renforce la réaction émotionnelle de l'utilisateur.

6 https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=BPC&division=7.&title=&part=3.&chapter=6.&article

7 High-Level Expert Group on Artificial Intelligence | Shaping Europe's digital future

Actuellement, ce recours au nom et à la réaction émotionnelle sert encore souvent à masquer le manque de performances sémantiques et contextuelles des agents conversationnels. Attribuer un nom à la machine relève de la dynamique de projection, c'est-à-dire d'anthropomorphisation de cette machine. Or, lorsque l'agent conversationnel lui-même emploie son « nom » dans un dialogue, se pose alors la question de l'autoréférence : à qui ou quoi exactement renvoie ce nom ?

2.1 L'utilisateur devrait-il pouvoir choisir le nom et le genre du nom (masculin, féminin, neutre) porté par un chatbot ou ce choix relève-t-il du concepteur ?

2.2 Un chatbot pourrait-il ou devrait-il se voir attribuer un nom humain (par exemple « Sophia »), un nom non-humain (par exemple « R2D2 ») ou bien aucun nom ?

2.3 Remarque libre :

3) Malmener les chatbots. La projection des qualités humaines sur les agents conversationnels est un phénomène courant et important. En particulier, les utilisateurs pourraient maltraiter un agent conversationnel.

Tandis que votre chatbot vous rappelle les gestes barrière pendant une épidémie, vous pourriez réagir en l'insultant ou en lui ordonnant de se taire. En outre, cela pourrait avoir une incidence sur les enfants qui entendent cet échange.

Les assistants vocaux généralistes (Siri, etc.) se font parfois insulter par les utilisateurs. Dans ce cas, ils répondent selon des stratégies prédéterminées par leurs concepteurs.

3.1 Insulter un chatbot dans une conversation est-ce un acte moralement répréhensible ? Pensez-vous qu'il est admissible de se servir du chatbot comme un souffle-douleur ?

3.2 Un chatbot insulté par son interlocuteur devrait-il pouvoir répondre à l'utilisateur en l'insultant au retour ?

3.3 Si un chatbot répondant à un nom féminin voire ayant une voix féminine est malmené, y voyez-vous un geste de maltraitance envers les femmes ? La même question se pose pour les noms masculins.

3.4 Remarque libre :

4) Confiance dans les chatbots. Une certaine confiance de l'utilisateur envers les finalités du chatbot est nécessaire pour la réalisation des tâches fonctionnelles du chatbot.

La confiance n'est pas seulement un phénomène psychologique émergent mais relève d'un effort technique : les concepteurs des agents conversationnels cherchent à l'établir et à la maintenir, mais pourraient également se poser la question d'éviter qu'elle soit accordée au chatbot de manière irréfléchie.

L'évaluation du niveau de confiance des utilisateurs envers les comportements et performances du chatbot est un important sujet de recherche.

4.1 Si la réponse « je ne sais pas » d'un chatbot vient en conflit avec la préservation de la confiance de l'utilisateur, par exemple dans le cas d'un service après-vente, faut-il privilégier la confiance en modifiant la réponse ?

4.2 Afin de gagner la confiance, le chatbot peut-il se présenter comme un.e « assistant.e / conseiller.ère / ami.e » de l'utilisateur ?

4.3 Remarque libre :

5) Les conflits des chatbots. Si la plupart des systèmes de dialogue sont conçus pour une tâche spécifique, de nombreux autres sont des agents conversationnels généralistes. Leur interaction avec l'être humain peut participer à un conflit. On se pose alors la question du rôle de l'agent conversationnel dans ce conflit et du jugement qui va tomber sur lui. *Par exemple, un chatbot pourrait donner de fâcheux conseils à son utilisateur, lui mentir, ou encore se comporter en délateur en appelant la police s'il détecte à tort ou à raison une menace.*

Les recherches actuelles portent sur le développement et l'utilisation des systèmes capables de s'adapter aux utilisateurs, à leurs desiderata, intentions et croyances en leur répondant comme le ferait un proche. Ces réponses adaptées voire « intelligentes » en réaction aux questions ou comportements des humains ne peuvent qu'engendrer chez les utilisateurs des croyances sur des « compétences » ou des supposés « états d'esprit » de la machine. L'humain s'adapte ainsi aux agents conversationnels avec lesquels il dialogue, soit en s'en méfiant, soit au contraire en leur donnant un certain « crédit de vérité ». En s'appuyant sur son « crédit de vérité », un chatbot pourrait proférer un mensonge.

La tension émerge lorsque le chatbot, par exemple, répond à une question de l'utilisateur relative à sa santé. Un médecin peut le cas échéant cacher toute la vérité à son patient dans le souci du bien-être de celui-ci.

5.1 Le mensonge proféré par un chatbot est-il plus ou moins acceptable que le mensonge humain ? La réponse dépend-elle du contexte (assistant vocal, éducation, psychothérapie, recrutement, etc.) ?

5.2 Si les chatbots peuvent mentir aux utilisateurs, qui et comment devrait décider des buts admissibles et des limites de tels comportements ?

5.3 Remarque libre :

6) La manipulation (*nudge*) des chatbots. Prix Nobel d'économie, l'Américain Richard Thaler a mis en lumière le concept de nudge, qui consiste à inciter les individus à changer de comportement sans les contraindre, par la seule utilisation de leurs biais cognitifs. Dans le cas des chatbots, les nudges sont définis comme des suggestions ou manipulations, manifestes ou cachées, conçues pour influencer le comportement ou les émotions d'un utilisateur.

Les agents conversationnels pourraient ainsi devenir un moyen d'influence des individus à des fins mercantiles ou politiques. Mais le nudge est aussi souvent mis en œuvre pour surveiller notre santé ou pour améliorer notre bien-être (faire plus d'exercice physique, consommer moins d'alcool, arrêter de fumer, etc.).

6.1 Tous les nudges sont-ils permis ? Comment peut-on distinguer les bons des mauvais nudges ?

6.2 Le concept de consentement libre et éclairé dans le cadre d'un agent conversationnel capables de « nudge » a-t-il encore un sens ?

6.3 Remarque libre :

7) Les chatbots et le libre choix. Lors d'un dialogue, les chatbots évaluent plusieurs réponses possibles pour en donner une seule. Dans le cas des systèmes de recommandation, ce choix unique pourrait limiter la liberté de l'utilisateur de choisir de manière autonome, en dérobant à sa vue toute la palette d'options disponibles. Cela génère en outre un risque d'enfermement (*filter bubble*), problème renforcé par le faible niveau de paramétrage proposé par les systèmes commercialisés actuellement.

Par exemple, à la demande de commander une pizza, le chatbot propose de commander chez un restaurateur particulier. Cela peut être le fournisseur plus proche géographiquement, le mieux noté sur un site donné ou encore celui qui possède un accord commercial avec le concepteur du chatbot. Or il propose un choix unique tandis qu'il existe au voisinage quinze pizzerias qui proposent le service demandé. Ce choix unique peut poser un problème éthique lié à la liberté et à la discrimination.

7.1 Dans l'exemple cité, souhaiteriez-vous que le chatbot explicite tous les choix ou plusieurs choix ?

7.2 Pensez-vous qu'une information transparente de l'utilisateur sur les critères de choix des recommandations par le chatbot soit une solution satisfaisante aux problèmes éthiques du libre choix et de la discrimination ?

7.3 Remarque libre :

8) Les émotions des chatbots. Les émotions sont souvent mélangées dans la vie de tous les jours. En conséquence, la détection et l'identification des émotions des utilisateurs dépendent d'un grand nombre de facteurs contextuels, culturels et idiosyncrasiques. L'informatique émotionnelle comprend trois grands domaines : détecter les émotions des humains, raisonner sur ces informations pour modifier la stratégie du dialogue et générer une expressivité émotionnelle par le langage ou le comportement non-verbal du chatbot.

Par exemple, ayant reconnu que l'utilisateur est stressé, un agent conversationnel peut simuler l'empathie et exprimer la compréhension de l'état de l'utilisateur.

8.1 Est-il souhaitable de construire des chatbots qui détectent les émotions des êtres humains ? Précisez la réponse selon le contexte d'utilisation.

8.2 Est-il souhaitable de construire des chatbots qui simulent des émotions des êtres humains ? Précisez la réponse selon le contexte d'utilisation.

8.3 Remarque libre :

9) Les chatbots et les personnes vulnérables. Un chatbot peut occuper toute l'attention d'une personne vulnérable en remplaçant, comme dans le cas des enfants autistes, le difficile contact avec les personnes humaines. Ce phénomène provoque souvent des jugements polarisés : d'un côté, le bien-être de la personne peut être amélioré ; de l'autre, il l'est au dépens de sa socialisation humaine « standard ».

Par exemple, un enfant autiste pourrait préférer l'interaction très nourrie et prolongée avec un chatbot à celle avec un parent ou un pédagogue. Un jeune enfant pourrait apprendre et imiter les comportements émotionnels de la machine au lieu de ceux des humains. Une personne âgée pourrait vouloir faire le deuil de son chatbot ou l'enterrer si elle lui est très attachée et qu'il ne fonctionne plus.

9.1 Quelles finalités de l'interaction entre un chatbot et une personne vulnérable (surveillance, éducation, accompagnement, divertissement) sont acceptables ? La réponse dépend-elle de l'âge de la personne (enfant, personne âgée) ou de son statut (patient, personne en convalescence) ?

9.2 Les utilisateurs, notamment les personnes vulnérables, sont susceptibles de s'attacher profondément à des chatbots, ce qui peut entraîner une modification durable de leur façon de vivre ou de leurs interactions sociales. Ce phénomène est-il inquiétant ? Pourquoi ?

9.3 Remarque libre :

10) Les chatbots et la mémoire des morts. Si le droit à la vie privée s'éteint à la mort de la personne, l'utilisation post-mortem de ses données, par exemple de sa voix, par un chatbot pour faire « revivre » cette personne peut néanmoins poser problème quant à l'atteinte possible au principe de respect de la dignité de la personne humaine.

Un journaliste américain est parvenu à créer un chatbot, le « dadbot », à partir des souvenirs qu'il avait de son père[8]. Il échange avec ce chatbot « comme si » il s'agissait d'un échange avec son père.

10.1 Pensez-vous que les chatbots sont un moyen envisageable pour faire « vivre » la mémoire ou la manière de s'exprimer propres à une personne décédée ? De tels usages porteraient-ils atteinte au principe de respect de la dignité de la personne humaine ?

10.2 Quelle évolution du concept de mort envisagez-vous en tenant compte des possibilités offertes par les chatbots ?

10.3 Remarque libre :

11) Surveillance par les chatbots. Si certains chatbots font partie des systèmes exclusivement consacrés à l'interaction humain-machine, d'autres fonctionnent dans des environnements partagés. Les chatbots capables d'enregistrer la voix pourraient ainsi surveiller les interactions autour d'eux, que celles-ci soient humaines ou avec d'autres chatbots. Cette capacité implique des enjeux éthiques et juridiques liées à la protection de la vie privée, à l'exploitation des données personnelles sans consentement, au risque de violation du secret personnel ou professionnel ainsi qu'à l'introduction de failles de sécurité. La divulgation par les chatbots des contenus enregistrés à l'insu des personnes peut s'apparenter à la délation.

[8] James Vlahos. *Talk to me, Amazon, Google, Apple, and the Race for Voice-Controlled AI*. Random House, 2019.

Par exemple, en cas d'écart à la diète que le médecin a imposée à un patient, le chatbot l'en informe, voire se met en contact l'organisme de soins de santé.

Autre exemple, un chatbot peut « tenir compagnie » des personnes vulnérables ou âgées en surveillant leur comportement.

11.1 Dans les exemples cités, pensez-vous que le comportement du chatbot est justifié ? Comment, dans ce cas, l'utilisateur peut-il exprimer son consentement ? Qu'en est-il si les chatbots sont placés dans des espaces partagés ?

11.2 Donnez d'autres exemples de situation dans laquelle la surveillance par un chatbot vous paraît justifiée.

11.3 Si un chatbot est insulté par son utilisateur, cette information doit-elle être communiquée par le chatbot à une tierce partie, par exemple son concepteur ?

11.4 Remarque libre :

12) Les chatbots et le travail. Les chatbots présenteront des opportunités et des risques pour les entreprises selon les contextes de leur utilisation (évaluation, recrutement, divertissement, etc.). L'introduction d'agents conversationnels dans les équipes peut induire des effets organisationnels selon les secteurs industriels, notamment du point de vue de la charge informationnelle et émotionnelle, de la temporalité du travail, du sentiment de cohésion ou d'isolement des travailleurs, des effets des chatbots sur le moral des employés ainsi que les problèmes d'égalité et de reconnaissance au mérite au sein des entreprises.

Par exemple, dans le secteur médical, l'aide à l'action humaine (médecins psychiatres, médecins généralistes, infirmiers, agents des centres d'appel d'urgence, etc.) par des chatbots pourrait provoquer des effets sur la profession dans sa totalité ainsi que sur le bien-être des patients et des personnels soignants et sur la relation entre eux.

12.1 Existe-t-il des métiers ou des pratiques humaines dans lesquels le recours aux chatbots devrait être encouragé ou prohibé ?

12.2 Comment et à quelle échelle temporelle envisagez-vous l'évolution des métiers à la suite de l'introduction des chatbots ? Précisez votre réponse selon un ou plusieurs cas d'usage.

12.3 Par quels moyens (législatif, code de bonne conduite, etc.) le recours aux chatbots devrait-il être encadré ?

12.4 Remarque libre :

13) Effets à long terme sur le langage. À moyen et long termes, l'utilisation des chatbots peut avoir une incidence durable sur le langage humain et peut-être également sur les habitudes de vie.

Par exemple, si les chatbots répondent par des phrases courtes, linguistiquement pauvres, sans politesse aucune, les humains risquent d'imiter ces tics langagiers lorsqu'ils s'adressent à d'autres humains.

13.1 Comment envisagez-vous l'évolution du langage sous l'influence des chatbots ? Cette influence peut-elle être jugée comme bonne ou mauvaise ?

13.2 Quelle échelle temporelle peut-on envisager pour cette évolution ?

13.3 Remarque libre :

II. Les facteurs éthiques dans la conception des chatbots

14) Problème de spécification. Les lois et les règles de conduite dans la société sont formulées dans une langue naturelle. Leur traduction dans un langage informatique exige une « spécification » : définition de tous les termes dans un cadre formel. Souvent, la spécification complète est impossible : par exemple, le terme « humain » peut inclure des humains qui seraient facilement identifiables par un système informatique apprenant, mais aussi des humains que le système ne parviendra pas à identifier comme tels car absents des données d'apprentissage. Quels que soient la base d'apprentissage et l'algorithme déployé, les erreurs d'identification sont inévitables : par nature, la langue humaine admet la multiplicité des significations.

Pour les chatbots, le problème de spécification se traduit, par exemple, par la difficulté de distinguer systématiquement et sans erreur, l'usage ironique ou satirique d'un concept ou d'une expression de son usage indicatif standard.

14.1 Quelles erreurs commises par les chatbots seraient acceptables et lesquelles ne le seraient pas ? Précisez la réponse selon le contexte (santé, éducation, divertissement, service après-vente, etc.).

14.2 Si un chatbot n'est pas capable de trouver une réponse, doit-il le dire explicitement ?

14.3 Quelles conséquences sur le comportement des utilisateurs la réponse « je ne sais pas », fréquemment donnée par les assistants vocaux actuels, entraîne-t-elle ? Si vous avez vécu cette expérience, décrivez-la.

14.4 Remarque libre :

15) Les métriques et les fonctions d'évaluation. Dans un agent conversationnel, les fins recherchées par le concepteur donnent lieu à la définition d'une métrique ou d'une fonction d'évaluation, qui quantifie la mesure de « bonne réponse » ou « réplique adéquate » pour le système. Cette métrique est encodée au préalable. La métrique d'un chatbot peut aussi tenir compte des facteurs émergents, qui apparaissent pendant la conversation, par ailleurs susceptibles de causer des ruptures dans la compréhension humaine du comportement du système. Souvent, cette qualité du dialogue est mesurée par le degré d'engagement de l'utilisateur, c'est-à-dire sa volonté à poursuivre le dialogue avec le chatbot. La métrique d'engagement utilise la longueur des échanges comme des marqueurs paralinguistiques (rire, sourire, hésitation, hochement de tête, etc.) de satisfaction ou d'intérêt ; or, dans l'état actuel des recherches, elle tient rarement compte du contenu sémantique des échanges. Cela peut défavoriser ceux qui ne comprennent pas le procédé d'évaluation de l'agent conversationnel et en outre donner lieu à des comportements manipulateurs de la part des utilisateurs.

En avril 2016, le chatbot Tay *de* Microsoft*, qui avait la capacité d'apprendre en continu à partir de ses interactions avec les internautes, avait appris à tenir des propos racistes.* Tay *a été rapidement retiré par* Microsoft.

Malgré cette expérience, DeepCom*, un autre chatbot développé par* Microsoft China *en 2019 afin de commenter des nouvelles sur les réseaux sociaux, a été reconnu par ses concepteurs eux-mêmes comme étant susceptible de générer des contenus biaisés, (par exemple, discriminants) voire de la propagande, à la suite de fortes réactions dans la communauté de recherche*[9]*. La première version de la publication postulait : « Compte tenu de la prévalence des articles de presse en ligne avec commentaires, il est très intéressant de mettre en place un système de commentaire automatique des nouvelles avec des approches construites à partir des données ». Dans la version révisée, les auteurs affirment : « Il existe un risque que des personnes et des organisations utilisent ces techniques à grande échelle pour simuler des commentaires provenant de personnes à des fins de manipulation ou de persuasion politique ».*

15.1 Faudrait-il que l'utilisateur soit informé du fait que la stratégie de dialogue d'un chatbot puisse être adaptée au cours de la conversation ?

15.2 Comme expliqué plus haut, l'utilisateur peut manipuler la métrique d'un chatbot à ses propres fins. S'il le fait, le concepteur partage-t-il l'éventuelle responsabilité pour les résultats de cette manipulation ou devrait-il en être dédouané ?

15.3 Avez-vous vécu des exemples personnels liés, selon votre interprétation, aux métriques particulières des chatbots ?

15.4 Remarque libre :

16) Les finalités de l'agent conversationnel : Les finalités d'un chatbot, c'est-à-dire les buts qui lui sont assignés, sont définies par ses concepteurs, et le chatbot cherche à les satisfaire dès sa mise en marche. Si cela ne pose pas de problèmes excessifs pour les chatbots dédiés à une ou plusieurs tâches connues au préalable, la spécification des finalités peut s'avérer complexe pour un chatbot généraliste car elles ne sont pas toutes énumérables au moment de la conception.

Ces finalités peuvent être très diverses : des systèmes après-vente aident à réparer des produits défectueux, des conseillers médicaux cherchent à améliorer l'état du patient, des services d'aide au recrutement, etc.

D'autres systèmes possèdent des finalités plus vagues : certains chatbots sont conçus afin de converser librement avec l'utilisateur sur tous les sujets. Que la perception des finalités ou le jugement que l'on porte sur elles puissent évoluer, cela ne supprime guère cette distinction

[9] Microsoft Used Machine Learning to Make a Bot That Comments on News Articles For Some Reason

fondamentale entre un agent conversationnel et un humain qui peut agir sans finalité prédéterminée et peut ne pas rendre sa finalité transparente aux autres.

16.1 Doit-on révéler la finalité d'un chatbot à l'utilisateur ? Si oui, à quel moment et sous quelle forme ? Si non, pourquoi ?

16.2 Devrait-on accepter qu'un chatbot capable d'apprentissage en interaction (par exemple, un agent conversationnel généraliste) puisse être dirigé vers une finalité particulière à travers une influence intentionnelle ou involontaire de la part des utilisateurs (par exemple, inciter la personne à faire un don ou à acheter un produit particulier) ? Précisez la réponse selon le contexte (santé, éducation, divertissement).

16.3 Remarque libre :

17) Les biais d'apprentissage. Un système apprend à partir de données sélectionnées par un « entraîneur » (agent humain responsable de leur sélection). L'existence de biais dans les données d'apprentissage est une source majeure des conflits éthiques, notamment à travers la discrimination ethniques, culturelles ou encore de genre.

Par exemple, des données de parole enregistrées peut contenir uniquement des voix d'adultes alors que le système est censé interagir aussi avec les enfants, ou un corpus de textes peut utiliser statistiquement plus fréquemment des pronoms de genre féminin que ceux de genre masculin.
Le système reproduira alors ces biais issus d'un corpus d'apprentissage, sauf s'il est équipé d'outils spécialement conçus dans le but de les corriger, ce qui présuppose déjà la connaissance des biais possibles. Or, certains biais pourraient ne pas être connus à l'avance.

17.1 Considérez-vous qu'un agent conversationnel devrait être sans biais ? Est-ce possible ? Précisez la réponse selon le contexte (santé, recrutement, service après-vente, éducation, sécurité, assistant vocal domestique).

17.2 Pensez-vous que les chatbots devraient imiter les biais humains ou les corriger ?

17.3 Remarque libre :

18) Instabilité de l'apprentissage. Des erreurs sont inévitables lorsqu'un système apprenant classifie une donnée qui ne ressemble pas, ou qui ressemble faussement, à celles contenues dans le corpus utilisé pendant son apprentissage. Dans le cas des agents conversationnels, cela recouvre les homophones, homographes, homonymes ou autres exemples d'ambiguïté linguistique.

Un cas simple est celui des erreurs d'orthographe : le comportement du chatbot dans ce cas diffère totalement de celui de l'être humain. Par exemple, l'utilisateur humain reconnaît un mot même s'il contient plusieurs erreurs, tandis qu'à cause de l'instabilité, un algorithme cesse de reconnaître correctement un mot contenant une ou deux fautes d'orthographe.

18.1 L'apprentissage des chatbots étant instable, il induit des erreurs parfois évidentes. Êtes-vous prêt à tolérer ces erreurs davantage que les erreurs humaines ? Précisez la réponse selon le contexte.

18.2 Les erreurs des chatbots provoquent-elles des sentiments ou des réactions différentes par rapport aux erreurs humaines ? Lesquelles ?

18.3 Remarque libre :

19) Explicabilité et transparence. La transparence d'un système signifie que son fonctionnement n'est pas opaque ou incompréhensible pour l'homme. Elle s'appuie notamment sur la traçabilité des répliques sélectionnées par un agent conversationnel. L'explicabilité signifie qu'un utilisateur peut appréhender le comportement du chatbot. Les problèmes de transparence et d'explicabilité sont provoqués par différents facteurs, notamment par le fait que, contrairement à l'être humain, un système informatique ne comprend pas le sens des phrases qu'il génère ou qu'il perçoit.

Ainsi, un chatbot, qui n'a pas de représentation du monde, *est susceptible de formuler des phrases qui ne correspondent à aucune réalité (« lait noir »), de répondre sans tenir compte du contexte (« Comment vas-tu ? » - « Il fait beau ») ou d'employer un lexique désagréable ou prohibé*.

Les effets immédiats sur l'utilisateur provoqués par un tel dialogue peuvent être importants (réaction émotionnelle forte, rupture dans la compréhension, abandon du dialogue ou débranchement du système). La question de responsabilité se pose alors à l'égard des concepteurs et des entraîneurs des agents conversationnels. La dimension esthétique (certaines paroles peuvent être étranges mais belles) suffit-elle à dédouaner le chatbot du besoin d'imiter toujours la parole humaine ?

19.1 À quelle réaction peut-on s'attendre de la part d'un utilisateur en situation de rupture de compréhension dans un dialogue avec le chatbot ? Précisez la réponse selon les finalités de celui-ci et le contexte (par exemple, santé, assistant vocal généraliste, divertissement, recrutement).

19.2 Lorsque l'utilisateur donne spontanément un sens à des répliques peu compréhensibles du chatbot, ce phénomène relève-t-il d'une attitude ludique ou pose-t-il un problème éthique ?

19.3 Remarque libre :

20) Impossibilité d'évaluation rigoureuse. Un agent conversationnel fournit une réponse en appliquant des stratégies de dialogue qui dépendent de l'interprétation. Les modèles les plus avancés utilisent de grands corpus de données pour apprendre.

L'évaluation de ce système de dialogue, par essence dynamique, est difficile au moins sur deux plans : *a) la prédiction des entrées générées par l'utilisateur n'est souvent pas possible; b) les aléas de l'apprentissage contribuent à la difficulté de reproduire le comportement du système*.

Or, l'incertitude théorique et pratique va de pair avec les techniques d'apprentissage qui procurent aux systèmes leur grande efficacité.

20.1 Est-ce acceptable qu'un chatbot profère des phrases « incongrues », qu'aucun être humain n'a jamais utilisées, ce qui serait susceptible d'influencer son interlocuteur ?

20.2 Un chatbot devrait-il se limiter à un ensemble prédéterminé de phrases ou, à l'inverse, en générer librement ? Précisez la réponse selon le contexte (divertissement, service après-vente, éducation, assistant vocal généraliste).

20.3 Remarque libre :

Merci beaucoup pour votre contribution !

L'envoi se fait à l'adresse cnpen-consultation-chatbots@ccne.fr