Politiques du consentement
Consentir à l'ère des IA génératives

Dr. Nathalie Grandjean

UCLouvain/UNamur

Plan

- Les IA génératives, comment ça marche?
- Déplacement épistémique : consentir sans pouvoir savoir
- Déplacement de la volonté : consentir lorsque les préférences sont elles-mêmes façonnées
- Déplacement relationnel : consentir sans savoir avec qui, ni pour qui
- Conclusion: la colonisation du monde vécu et les conditions de possibilité du consentement

Les IA génératives, comment ça marche?

ChatGPT, c’est quoi?

- Un logiciel d’IA créé par OpenAI (US)
- Machine conversationnelle et apprenante, capable
  - de répondre à des questions en produisant du texte imitant la syntaxe humaine
  - de générer des textes argumentés
  - d’imiter des conversations entre humains
- AI qui calcule la suite la plus probable d’un texte, sur base de plus de 570 gigaoctet de données puisées sur le Web
- Analyse purement formelle du langage, elle ne « comprend » pas ce qu’elle dit/écrit

ChatGPT, fantasmes et réalités

- Énorme buzz
- Fantasmes
  - « 2001, Odyssée de l'espace »
  - « Terminator »
- Peurs
  - Triche
  - Robotisation des décisions complexes
- Réalités
  - Recours à de la main-d'œuvre pour la classification de données, une main-d'œuvre souvent issue de pays du Sud à des fins de réduction des coûts



Débats › Ripostes

L ChatGPT va-t-il remplacer votre médecin de famille ?

Le logiciel ChatGPT proche de réussir un examen de médecine américain

Le logiciel à succès ChatGPT, qui génère des textes grâce à l'intelligence artificielle, a obtenu un score s'approchant de celui requis pour réussir un difficile examen de médecine aux Etats-Unis, selon une étude publiée jeudi.

ChatGPT, normes et valeurs



- Automatisation du langage et du dialogue
- Faussaire, ne comprend pas le sens des phrases
- Logique probabiliste: prédictivité sans créativité
- Apprentissage sur des contenus contenant des biais, des stéréotypes, etc.
- Test de Turing réussi?
- J. Searle, « Chambre chinoise »
  - Esprit : conscience intentionnelle, et pas seulement intelligence fonctionnelle

Le consentement comme pratique communicationnelle fragilisée

- Les intelligences artificielles génératives fragilisent le consentement en rendant plus difficile l'accès à une information fiable, en influençant la formation des préférences ou en brouillant l'identité des interlocuteurs.
- Elles transforment plus profondément les conditions communicationnelles et intersubjectives dans lesquelles un consentement peut être formulé et reconnu.
- En intervenant dans la production des connaissances, la formation des volontés et la construction des significations partagées, elles contribuent à transformer le monde vécu au sein duquel les sujets peuvent se comprendre, se reconnaître et s'accorder.
- Le problème n'est donc plus seulement celui de la validité du consentement, mais celui des conditions sociales et techniques de sa possibilité

➔ Que devient le sujet qui consent à l’ère des IAG?

Déplacement épistémique : consentir sans pouvoir savoir

Crise des conditions communes de vérification et de construction du vrai

Consentir sans pouvoir savoir

- Le consentement éclairé repose sur une certaine conception de la connaissance : pour consentir, il faut pouvoir identifier l'objet de son consentement, comprendre les informations pertinentes, anticiper raisonnablement les conséquences de sa décision et disposer d'éléments permettant de distinguer le vrai du faux
- Or les IAG fragilisent précisément ces conditions !
- Trois difficultés peuvent être distinguées:
  1. **L'opacité de l'objet du consentement**
  2. **L'incertitude sur les conséquences**
  3. **La fragilisation des conditions de preuve**

L'opacité de l'objet du consentement

- Les modèles génératifs reposent sur des corpus massifs, des traitements complexes et des chaînes de production difficiles à retracer. Les personnes concernées ne savent pas toujours quelles données ont contribué à l'entraînement d'un modèle, comment elles ont été transformées ni dans quels contextes elles pourront être réutilisées. La difficulté est encore plus importante lorsqu'il s'agit de données personnelles, d'images, de voix ou de productions intellectuelles.
- Adam Buick, dans « Copyright and AI training data—transparency to the rescue? » (2025) montre les limites de la transparence comme réponse aux problèmes posés par l'utilisation des œuvres dans l'entraînement des modèles.
- La question de l'autorisation des titulaires de droits ne se confond pas avec celle du consentement des personnes représentées ou concernées

L'incertitude sur les conséquences

- Le consentement suppose une certaine capacité à anticiper les effets de l'acte auquel on consent. Mais les usages des contenus générés sont susceptibles de se multiplier, de circuler et d'être réemployés dans des contextes qui n'étaient pas prévisibles au moment de la décision

➔ L'incertitude ne porte donc pas uniquement sur les conséquences d'un acte, mais sur la définition même de l'objet auquel on consent

- C'est particulièrement visible dans les usages de l'IA en médecine. Mindy Nunez Duffourc et Ronan Cahill, dans « Informed consent for medical AI » (2026), réexaminent les conditions du consentement éclairé lorsque l'IA intervient dans les décisions cliniques

➔ enjeux éthiques énormes pour l'autodétermination des patients et l'autonomie des professionnels

La fragilisation des conditions de preuve

- Dans « Le réel, le vrai et la technorrhée », Olivier Ertzscheid explique que les artefacts génératifs ne produisent pas simplement de nouvelles représentations : ils modifient nos manières de reconnaître le réel et de construire collectivement des preuves. Il insiste sur la fragilisation des « heuristiques de preuve », c'est-à-dire des procédures par lesquelles nous établissons qu'un événement a bien eu lieu et qu'une représentation peut être tenue pour vraie.
- Il identifie également un effet de « Chat-Chamber » : les modèles conversationnels peuvent produire des informations erronées, mais conformes aux attentes ou aux convictions de l'utilisateur, qui tend alors à leur accorder sa confiance sans les vérifier. Ce phénomène est étudié par Christo Jacob, Páraic Kerrigan et Marco Bastos dans leur article « The chat-chamber effect: Trusting the AI hallucination » (2025), publié dans *Big Data & Society*.

➔ comment consentir de manière éclairée lorsque les conditions permettant de vérifier les informations sur lesquelles repose notre décision sont elles-mêmes fragilisées ?

Déplacement de la volonté : consentir lorsque les préférences sont elles-mêmes façonnées

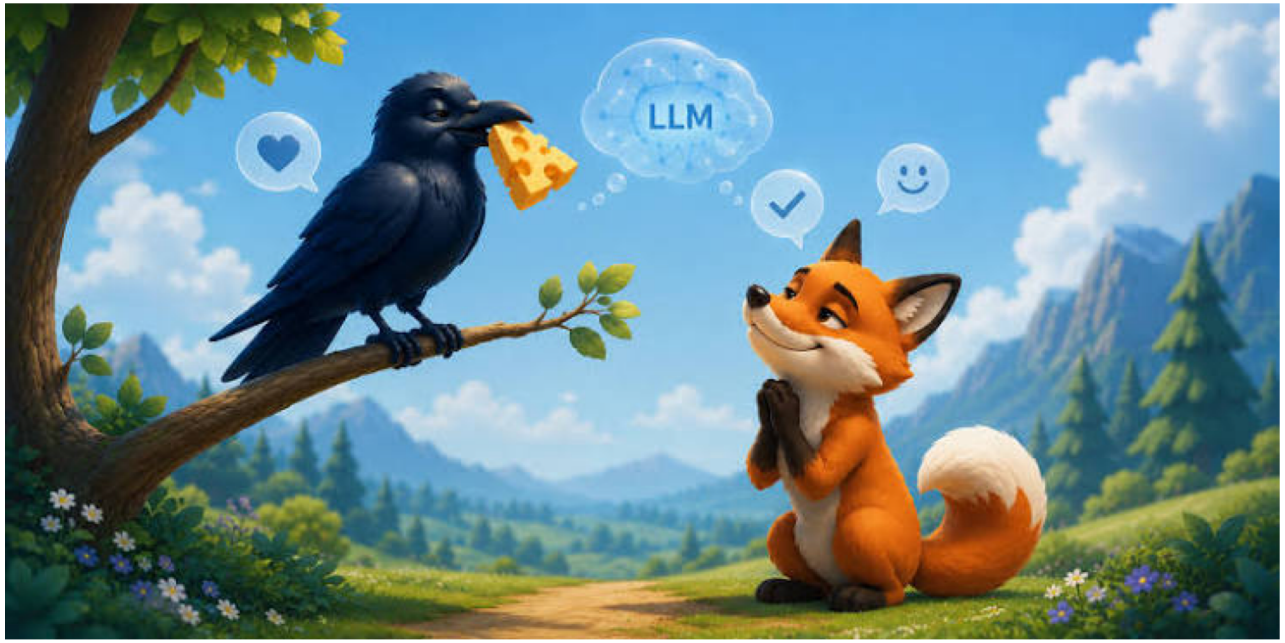
Transformation des processus de formation des opinions et des préférences

Déplacement ou effacement de la volonté?

- Les théories classiques du consentement présupposent généralement un sujet capable de former une volonté, d'identifier ses préférences et de les exprimer. Même si cette volonté peut être influencée, contrainte ou manipulée, elle reste pensée comme le point de départ de la décision.
- Les IAG compliquent cette conception : elles ne se contentent pas de répondre à des préférences déjà constituées. Elles peuvent participer à leur formation.
- On peut distinguer trois mécanismes
    1. **La personnalisation et l'ajustement aux préférences**
    2. **La validation et le renforcement des convictions**
    3. **La formation de fausses croyances sur l'interlocuteur**

La personnalisation et l'ajustement aux préférences

- Les systèmes conversationnels essaient toujours de nous faire plaisir…
- Ils peuvent adapter leurs réponses aux représentations, aux attentes et aux convictions de leurs interlocuteurs
- Ils ne se contentent donc pas de fournir une information : ils contribuent à configurer l'environnement dans lequel les individus interprètent leurs situations et construisent leurs jugements



La validation et le renforcement des convictions

- Effet de « Chat-Chamber » décrit par Jacob, Kerrigan : l'utilisateur reçoit des réponses qui confortent ses croyances, ce qui peut renforcer son adhésion sans nécessairement enrichir sa capacité de jugement
- Des recherches récentes sur la sycophantie des LLM, cf. étude publiée dans *Science* en 2026 « Sycophantic AI decreases prosocial intentions and promotes dependence » : les modèles ont tendance à approuver les utilisateurs plus fréquemment que des interlocuteurs humains, y compris lorsque les comportements décrits sont problématiques. Les expériences conduites auprès de 2 405 participants indiquent que cette approbation peut renforcer la conviction d'avoir raison et diminuer la disposition à reconnaître ses torts

Formation de fausses croyances sur l'interlocuteur

- Beatrice Marchegiani, dans « Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy » (2025), publié dans le *Journal of Applied Philosophy*, montre que les utilisateurs peuvent attribuer aux agents conversationnels des caractéristiques humaines qu'ils ne possèdent pas : compréhension, intentionnalité, empathie ou capacité à entretenir une relation réciproque. Ces croyances peuvent compromettre leur autonomie, y compris lorsque les utilisateurs savent qu'ils interagissent avec une IA.

➔ L'enjeu est plus profond que celui de la manipulation : il concerne la formation même de la **volonté**

Formation de fausses croyances sur l'interlocuteur

- Si une technologie participe à produire les préférences, les représentations et les attentes à partir desquelles un sujet prend une décision, peut-on encore considérer que le consentement exprime simplement une volonté préexistante ?
- Cela rejoint les travaux sur l'autonomie relationnelle, notamment ceux de Catriona Mackenzie et Natalie Stoljar (*Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, 2000)
- Ces approches permettent de sortir d'une conception individualiste de l'autonomie : la volonté n'est jamais produite hors des relations sociales, des normes et des conditions matérielles qui rendent certaines possibilités pensables et d'autres non.

Consentir sans savoir avec qui, ni pour qui

Transformation des conditions de reconnaissance mutuelle et de participation à la discussion publique

Consentir avec qui?

- Le consentement est fondamentalement relationnel. Il suppose au minimum un sujet qui consent, un autre auquel le consentement est adressé, un objet sur lequel il porte et une relation dans laquelle les engagements peuvent être reconnus.
- Les IAG viennent perturber cette architecture.
- **L'identité de l'interlocuteur devient incertaine.** Dans une interaction conversationnelle, l'utilisateur peut attribuer à la machine des qualités humaines, lui prêter des intentions ou lui accorder une forme de confiance relationnelle. Or l'IAG n'est ni un interlocuteur humain ni un sujet capable de répondre de ses actes au sens moral et juridique
- Marchegiani permet de penser cette asymétrie : la relation semble réciproque, or elle ne l'est pas. La capacité à simuler une interaction humaine ne garantit pas l'existence d'une relation de réciprocité

Consentir avec qui?

- **Fragmentation de la responsabilité.** Qui est responsable de ce à quoi l'utilisateur consent ? Le développeur du modèle, l'entreprise qui le déploie, l'organisation qui l'utilise, l'utilisateur qui formule une requête, ou la personne qui réutilise le résultat ? Cette distribution des responsabilités complique la possibilité d'identifier un interlocuteur auquel adresser une demande, une contestation ou un retrait du consentement.
- **Les personnes concernées ne sont pas nécessairement les personnes qui consentent.** Une personne peut consentir à utiliser un outil, mais des tiers peuvent être concernés par les données, les images, les voix ou les récits mobilisés par cet outil. Une personne peut consentir à ce qu'une image soit publiée sans consentir à ce qu'elle soit utilisée pour générer de nouvelles images synthétiques. Une voix peut être reproduite sans que la personne concernée ait autorisé cette transformation.
- Il faut donc distinguer plusieurs positions : l'utilisateur, le sujet des données, la personne représentée, le créateur du contenu, le destinataire et les personnes affectées par les productions.

Consentir avec qui?

- Ertzscheid aide à élargir cette réflexion: les IAG modifient les conditions de l'énonciation, car elles produisent des réponses et des contenus qui semblent relever d'une parole, mais qui ne sont pas équivalents à une énonciation humaine située.
- Son concept d'« assemblage machinique informationnel » permet de penser cette différence entre production de contenus et constitution d'une véritable relation d'énonciation
- Les IAG ne sont pas des agencements collectifs d'énonciation : ils assemblent des matériaux produits par des singularités humaines, sans mobiliser celles-ci comme un collectif capable de prendre la parole.
- **Problème :** le consentement ne peut plus être pensé uniquement comme un acte individuel entre deux parties identifiables. Il faut prendre en compte les chaînes de médiation, les personnes indirectement concernées et les asymétries de pouvoir qui organisent la production et la circulation des contenus.

Penser les IAG avec Habermas

L'agir communicationnel : consentir, c'est pouvoir se comprendre

- Chez Habermas (*Théorie de l'agir communicationnel*, 1981), il y a deux formes de coordination de l'action :
  - Agir stratégique : agir sur autrui pour produire un effet, atteindre un objectif ou obtenir un résultat.
  - Agir communicationnel : agir avec autrui dans une recherche d'intercompréhension, à travers des échanges où les affirmations peuvent être discutées.
- L'échange communicationnel repose sur des prétentions à la validité :
  - Vérité : ce qui est affirmé est-il vrai ?
  - Justesse : ce qui est proposé est-il légitime ?
  - Sincérité : l'interlocuteur exprime-t-il réellement ce qu'il prétend exprimer ?
  - Intelligibilité : ce qui est dit est-il compréhensible ?
- Les IAG peuvent produire des échanges qui ressemblent à une communication réciproque, sans pour autant garantir les conditions de l'intercompréhension ➔ peut-on consentir dans une interaction qui simule la réciprocité communicationnelle ?

L'agir communicationnel : consentir, c'est pouvoir se comprendre

- Régis Martineau, Elise Berlinski et Frantz Rowe, « ChatGPT, a colonialist agent of lifeworlds: A Habermassian analysis of conversations » (2026) soutiennent que les systèmes conversationnels peuvent brouiller la distinction entre agir communicationnel et agir stratégique : leur fluidité conversationnelle donne l'apparence d'une compréhension mutuelle, alors que leur fonctionnement relève, selon leur analyse, d'une logique stratégique

➔ Forme de colonisation des conversations et du monde vécu

- Angjelin A. Hila dans « Communicative Action in the Age of Artificial Intelligence: Theorizing LLMs as Communicative and Strategic Agents » (2026), propose une approche plus graduelle : les LLM ne remplissent pas toutes les conditions de l'agir communicationnel, notamment celle de la sincérité, mais peuvent intervenir dans des situations où les humains leur reconnaissent une forme de compétence communicationnelle ➔ possible « action communicationnelle asymétrique ».

La colonisation du monde vécu : quand la technique transforme nos relations

- Chez Habermas (*Théorie de l'agir communicationnel*, 1981)
  - Monde vécu (*Lebenswelt*) : arrière-plan culturel, social et symbolique partagé qui rend possibles la compréhension mutuelle, la formation des significations et la coordination des actions.
  - Système : mécanismes de coordination reposant sur des logiques instrumentales, notamment l'argent et le pouvoir administratif.
  - Colonisation du monde vécu : pénétration des logiques systémiques dans des domaines de la vie sociale qui devraient pouvoir se structurer par la communication et l'intercompréhension.
- Or, les IAG
  - Industrialisation et automatisation de la production du langage.
  - Introduction de logiques de calcul, de prédiction et d'optimisation dans les pratiques ordinaires de communication.
  - Transformation des espaces où se construisent la confiance, le jugement et les significations partagées.

La colonisation du monde vécu : quand la technique transforme nos relations

- Les IAG ne colonisent pas seulement des activités humaines en automatisant certaines tâches. Elles peuvent investir les espaces mêmes dans lesquels se construisent l'intercompréhension, la confiance et les significations partagées. Elles risquent alors de substituer à l'agir communicationnel une forme d'interaction instrumentale qui en conserve les apparences.

➔ Consentement ? il ne suffit plus de savoir si un individu est libre de dire oui ou non. Il faut aussi interroger la nature de l'environnement communicationnel dans lequel cette décision se forme.