Éduquer à l'Intelligence Artificielle

L'éthique dans l'IA

Laurent Gimazane

Points abordés

01 Notions préliminaires

Définition de l'éthique – L'éthique dans l'IA – Notion de "Nudge"

02 La dépendance émotionnelle

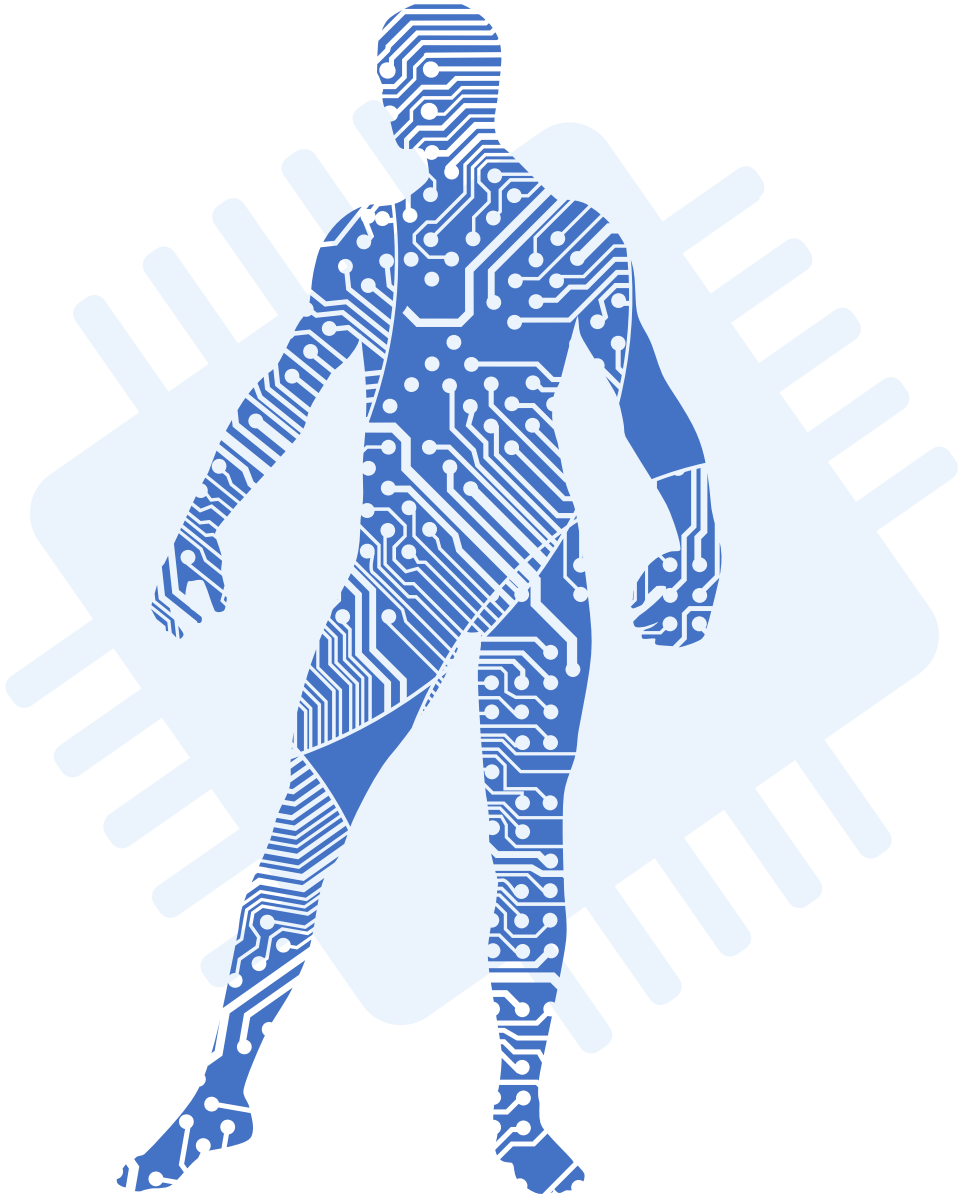
L'affective computing.

03 Les "Chatbots"

Qu'est-ce qu'un Chatbot? – Des exemples et contre-exemples

04 Un cadre légal

Comment cadrer l'usage de l'IA – Une tentative d'écriture de la charte

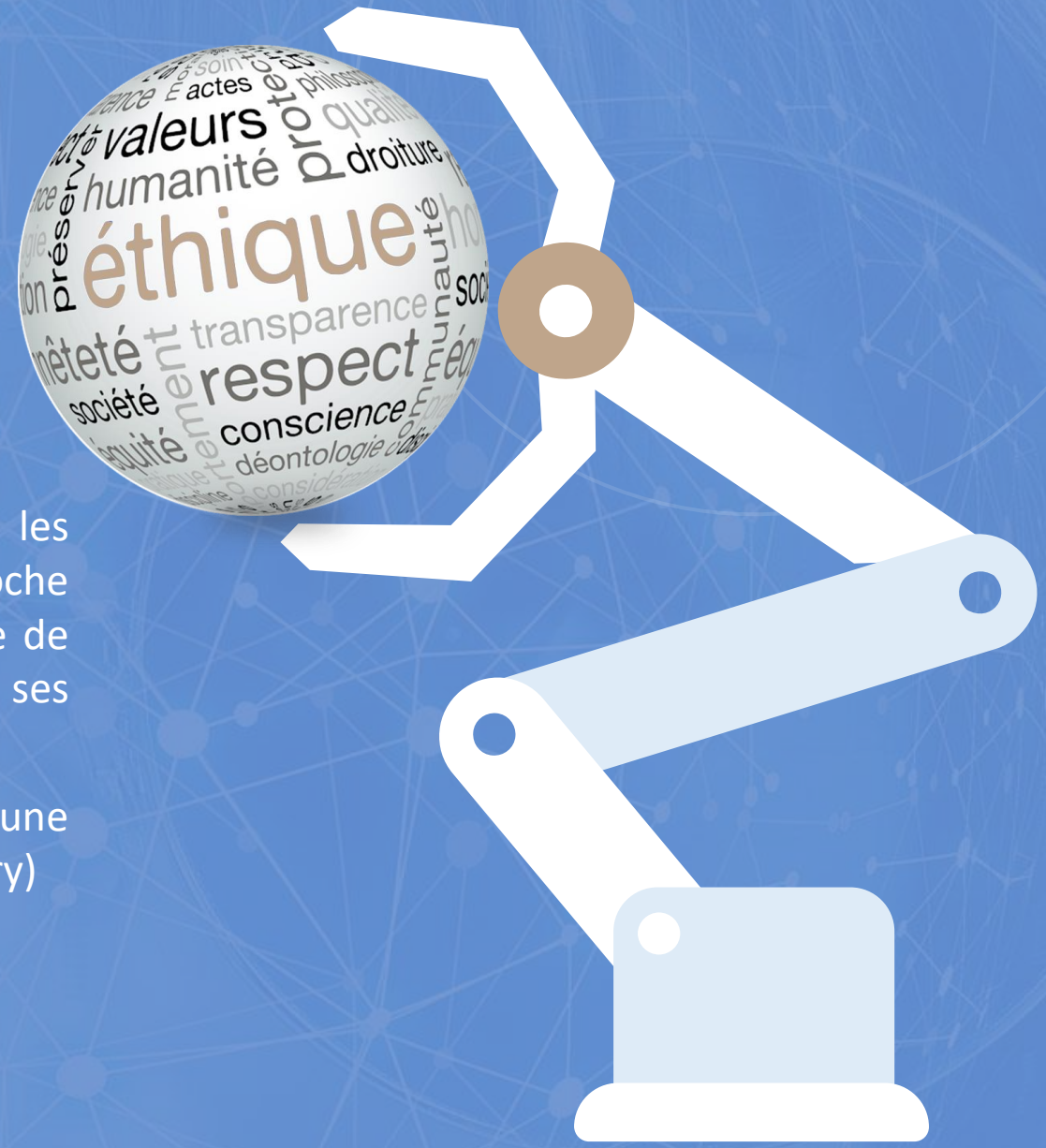


Notions préliminaires

Qu’est-ce que l’éthique?

Que veut dire éthique?

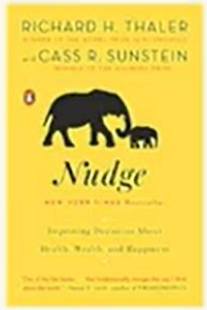
- L'éthique est une discipline philosophique portant sur les jugements moraux et dont le concept est donc très proche de celui de la morale. C'est une réflexion fondamentale de tout peuple afin d'établir ses normes, ses limites et ses devoirs. Wikipédia
- Principes moraux qui régissent le comportement d'une personne ou la conduite d'une activité (Oxford dictionnary)



L’éthique dans l’IA?

Plusieurs axes ou points de vue qui suscitent des actions critiques autour de la co adaptation humain machine (**nudge** « incitation à faire » plus efficiente avec un robot).

Le nudge : une incitation douce



Les nudges fonctionnent en faisant appel à nos biais cognitifs et à notre façon " irrationnelle " de prendre des décisions.

- **nos capacités cognitives sont limitées**
- **nous manquons de maîtrise de soi**
- **nous agissons de manière émotionnelle**
- **nous agissons par conformité**
- **nous agissons par paresse, etc.**

Le nudging pourrait être utilisé dans un avenir proche dans les chatbots et les robots sociaux :

- pour inciter à l'achat,
- pour influencer un comportement qui peut être ou ne pas être souhaité par les utilisateurs.

2 vidéos pour nous éclaircir

Quel modèle économique derrière?

Ethique by design : Biais des données ou des algorithmes - Peut-on les évaluées et y remédier ou la décision est-elle donnée aux machines / Cybersécurité

Ethique des usages : Acceptabilité et confiance dans le numérique ; Comment sont capturer nos données, L'IA est-elle parfaite ou fait elle des erreurs ; L'IA est elle malveillante ; Peut-on raisonner face à l'IA ou doit on accepter ses conclusions et choix ; Peut-on se débrancher des IA.

Ethique sociétale : Comment traiter les IA ; Comment vont-elles modifier nos comportements en société ; Va-t-elle changer nos jugement de valeurs dans le travail (créativité, responsabilité, mérite…)

La dépendance émotionnelle

Qu’est-ce que l’éthique?

Dépendance émotionnelle

Demain, ils vont s'immiscer dans nos vies pour nous inciter à faire « les bons choix », pour décider à notre place s'il faut faire du sport, ou aller chez le médecin, pour être témoins de notre intimité, nous soigner, être des objets sexuels et remédier à notre solitude. Le pire sans doute est que nous pourrions être heureux qu'enfin « quelqu'un » de bienveillant fasse attention à nous et soit là pour nous. Malgré le fait qu'ils sont des leurres, des présences vides de sentiments, ils sont capables de s'adresser à nous, d'apprendre nos habitudes et de nous montrer de l'attention, voire de l'empathie s'ils sont programmés pour être sociaux et affectifs, voire du plaisir s'ils sont programmés pour.

Les dimensions affectives envahissent les machines pour permettre un dialogue plus naturel mais aussi pour capter notre attention et nous rendre dépendant d'elles.

L'apprentissage machine et des approches symboliques sont utilisés pour créer les systèmes de dialogue émotionnel. Ce sont les prochains défis des agents conversationnels (Google Home, Alexa Amazon), des assistants virtuels (2D/3D) et des robots compagnons

L'illusion que les robots pourraient s'humaniser à notre contact serait le moyen le plus simple de créer entre nous et eux une sorte d'adoption. L'animisme très répandu dans la société japonaise réconcilie l'objet et le vivant.

Plus un robot sera capable de s'adapter à nous, de tourner la tête dans notre direction quand nous l'appelons, de nous faire un geste particulier, de nous appeler par notre prénom, plus ce stratagème marchera. Ce qui rend la manipulation du robot convaincante est qu'il nous invite à nous occuper de lui.

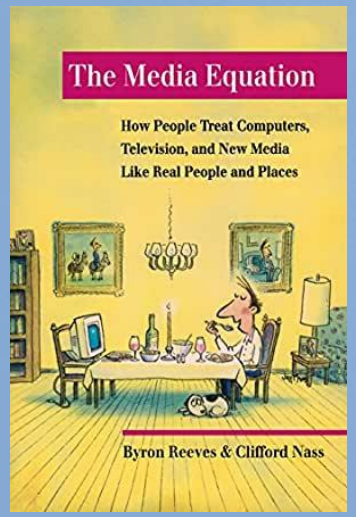
Dépendance émotionnelle

A l'instar de l'avion qui ne bat pas des ailes comme un oiseau pour voler, nous construisons des machines capables d'imiter sans ressentir, de parler sans comprendre et de raisonner sans conscience. Si leur rôle peut être extrêmement positif dans le domaine de la santé pour par exemple le suivi des maladies chroniques, il faut vérifier les risques de manipulation, d'isolement ou encore de dépendance.

Comment évolueront nos relations avec les machines émotionnels ? Un certain nombre de valeurs éthiques sont importantes pour la conception et les usages de ces machines : la déontologie et la responsabilité des concepteurs, l'émancipation et la responsabilité des utilisateurs, l'évaluation, la transparence, l'explicabilité, la loyauté, la non-discrimination des systèmes et l'anticipation des conséquences de la co-évolution homme-machine. Nous devons également éviter deux écueils : la paresse, qui consiste à abandonner notre libre arbitre aux choix opérés par ces machines « amies » et le complexe d'infériorité face à des technologies qui calculent plus vite que nous, savent plus que nous et nous impressionnent d'autant plus que nous ne comprenons pas vraiment comment elles fonctionnent.

Laurence Devillers, Professeure Sorbonne-université.

L’anthropomorphisme



01

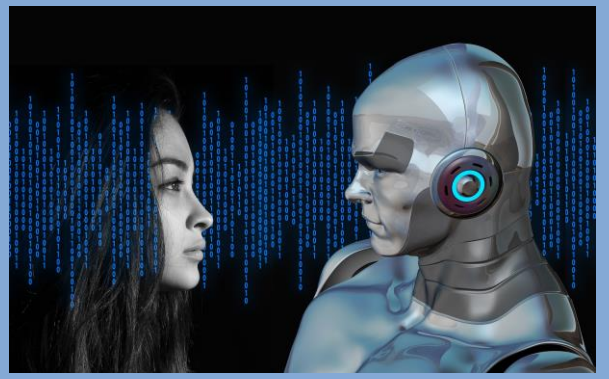
Définition

L'anthropomorphisme est l'attribution des caractéristiques comportementales ou morphologiques de vie humaine à des objets

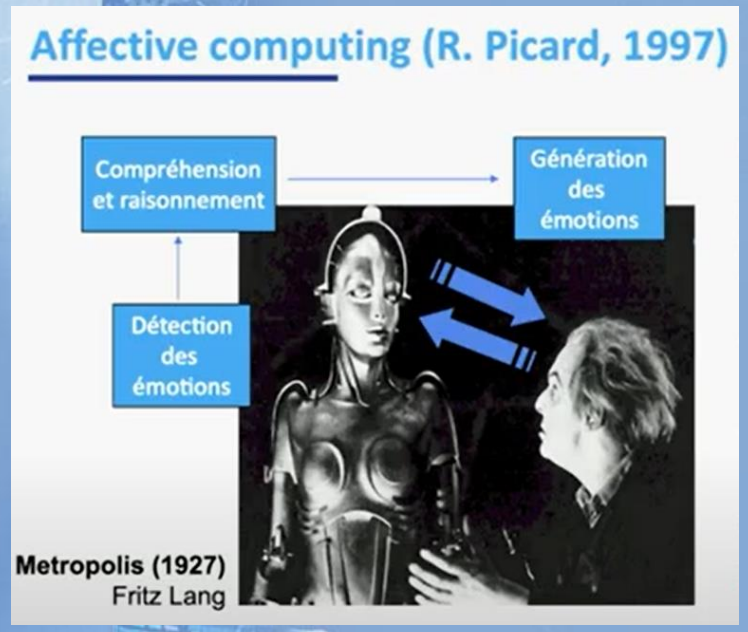
02

Notre comportement face aux objets.

Il est possible d’éprouver des émotions en face de n’importe quel objet, L’humain projette des relations affectives avec des robots qui en sont dépourvus : robots conversationnels, robots aspirateurs…



L'affective computing



01

Définition

La dépendance aux objets dotés d'intelligence artificielle et de l'isolement qui pourrait en résulter. Elle souligne le besoin d'un cadre législatif et de la transparence pour contrer le fait que nous serons de plus en plus sur écoute, de plus en plus surveillés
Entre l'émotion et l'empathie (**affecting computing**) qu'utilisent ces objets de plus en plus humanisés où se situe l'éthique ? Les machines qui s'immiscent dans notre intimité peuvent faire aussi des choses extraordinaires, mais où placerons-nous les limites ?
La relation entre personnel soignant et patient reste primordiale. L'homme "doit garder la main", le contrôle, la décision finale. »

02

Comment analyser des émotions sachant que personne ne les traduits ni ne les ressent de la même manière ?

La plupart des mécanismes psychologiques sont nécessaires en tant que telle (expression), influencés par l'émotion (perception, attention, mémoire, jugement moral et prise de décision).
« Nous ne sommes pas rationnels sans être émotionnels »

A.Damasio

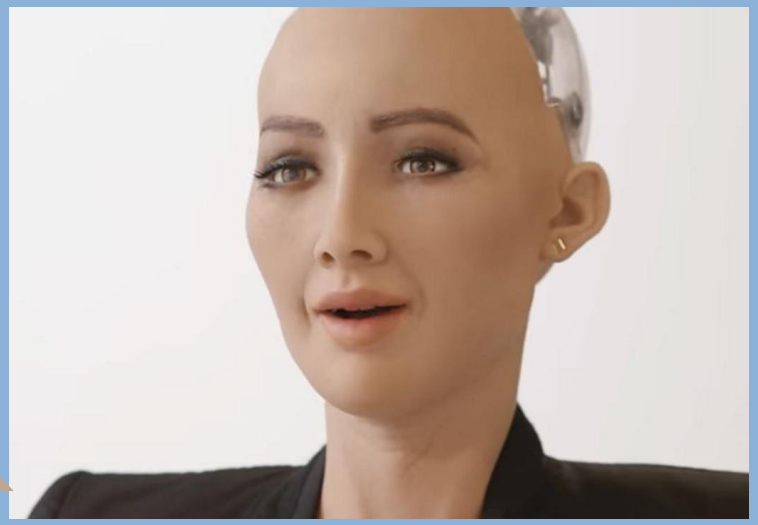
Illusions du vivant

Pourquoi féminiser les robots ou les interfaces alors que les programmeurs sont essentiellement des hommes – Pourquoi préfère t-on des voix de femmes

Quelques exemples des limites et dérives de l'IA

Ishiguro et son clone

« Sophia » un clone trop parfait



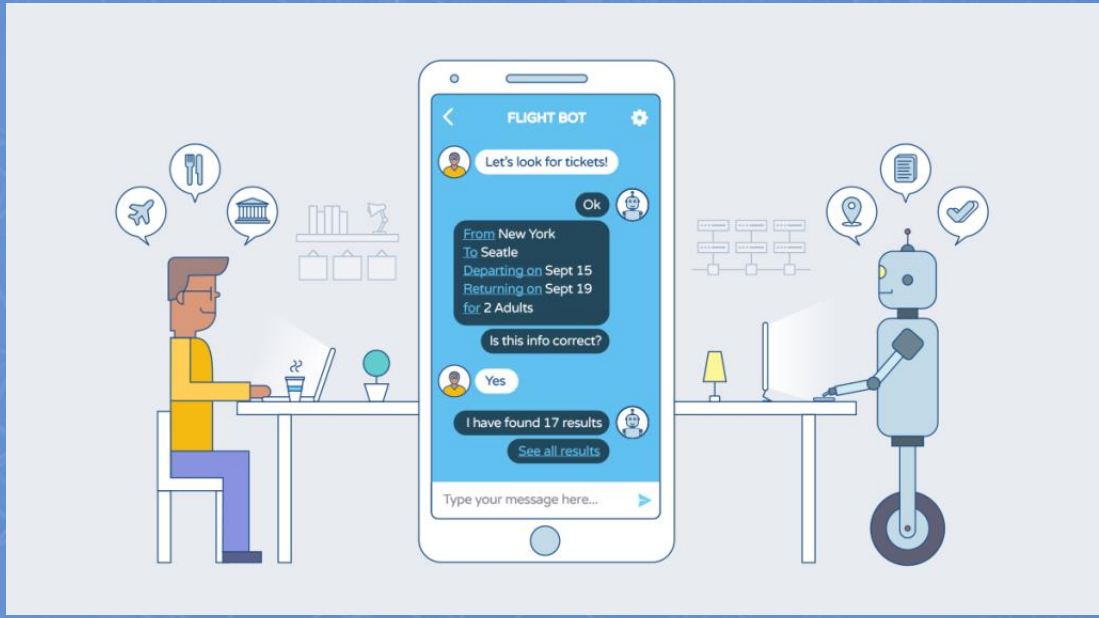
« Gatebox » pour vaincre la solitude

Les Chatbots
Agents conversationnels

Une assistance virtuelle

Qu’est-ce qu’un Chatbot?

Un **chatbot** est un robot logiciel pouvant dialoguer avec un individu ou consommateur à l’aide d'un service de conversations automatisées pouvant être effectuées par le biais d'arborescences de choix ou par une capacité à traiter le langage naturel.
Le chatbot utilise à l'origine des bibliothèques de questions et réponses, mais les progrès de l'intelligence artificielle lui permettent de plus en plus "d'analyser" et "comprendre" les messages par le biais des technologies de traitement du langage naturel (NLP) et d'être doté de capacités d'apprentissage liées au machine learning.



Le système d’intelligence artificielle repère les mots-clés et répond grâce à des réponses pré-enregistrées. Celles-ci sont conçues par l’équipe du Community Management. Le chatbot est constamment supervisé par une équipe dédiée.

Le chatbot sert donc d’assistant virtuel, capable de discuter avec un client afin de répondre au mieux à sa demande.

Les réponses du chatbot sont tout de même limitées et les sujets plus délicats ne peuvent être traités.

En résumé

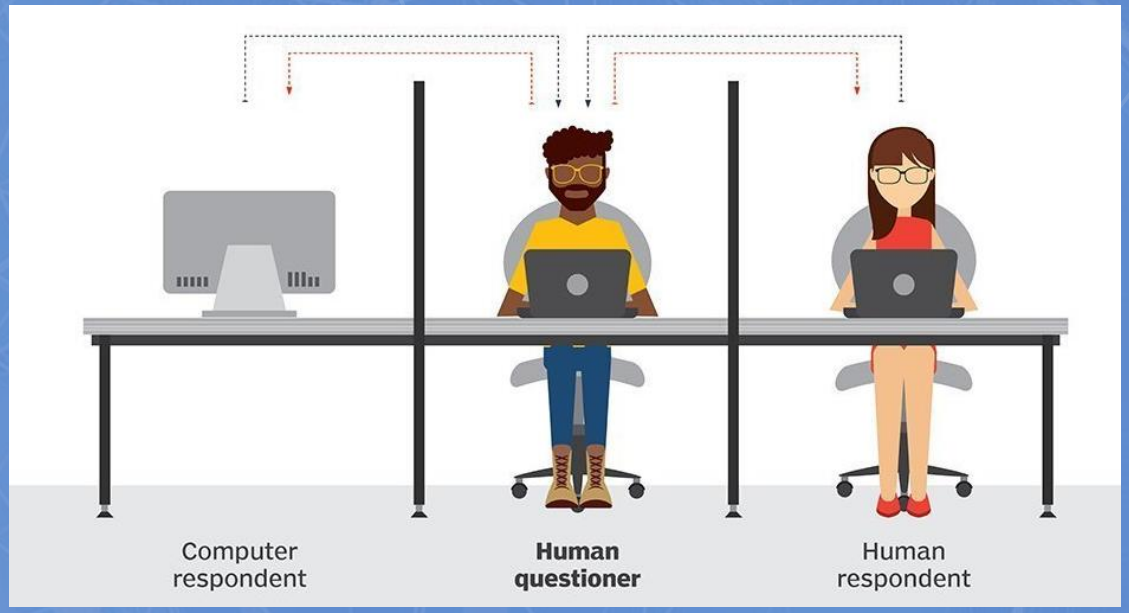
Un **chatbot** comporte plusieurs modules :

- Capture du signal et reconnaissance de la parole
- Analyse sémantique et gestion du dialogue
- Génération de la réponse et synthèse de la parole

Les paramètres qui définissent un Chatbot sont :

- Co construire une histoire
- Comprendre le sens
- Comprendre les états mentaux (émotions)
- Raisonner (pas dans le sens de l'intelligence humaine)
- Formuler des réponses
- Simuler l'empathie
- S'engager
- Se synchroniser

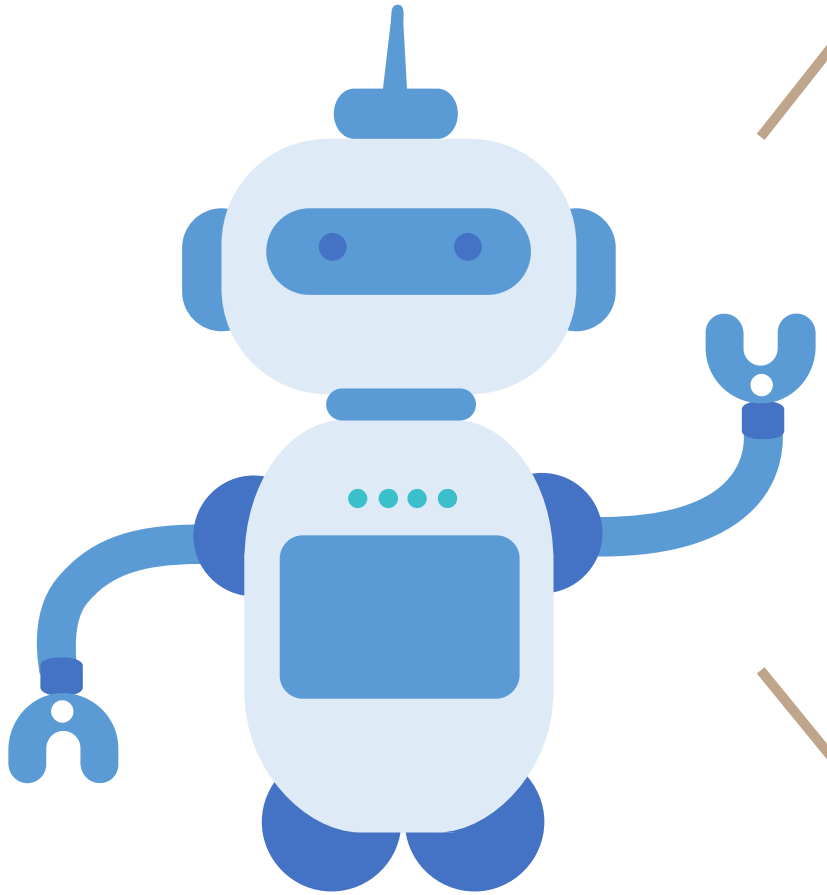
Une manière d'évaluer les performances d'un Chatbot : Le test de Turing



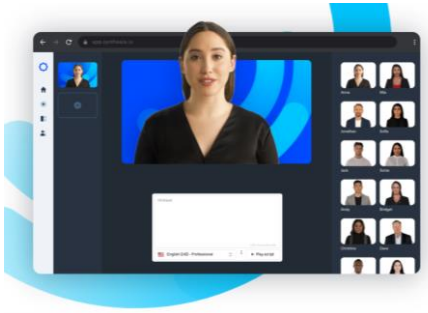
Des exemples de Chatbots

Alexa, le Chatbot d'Amazon

Bixby l'ancien Chatbot de Samsung

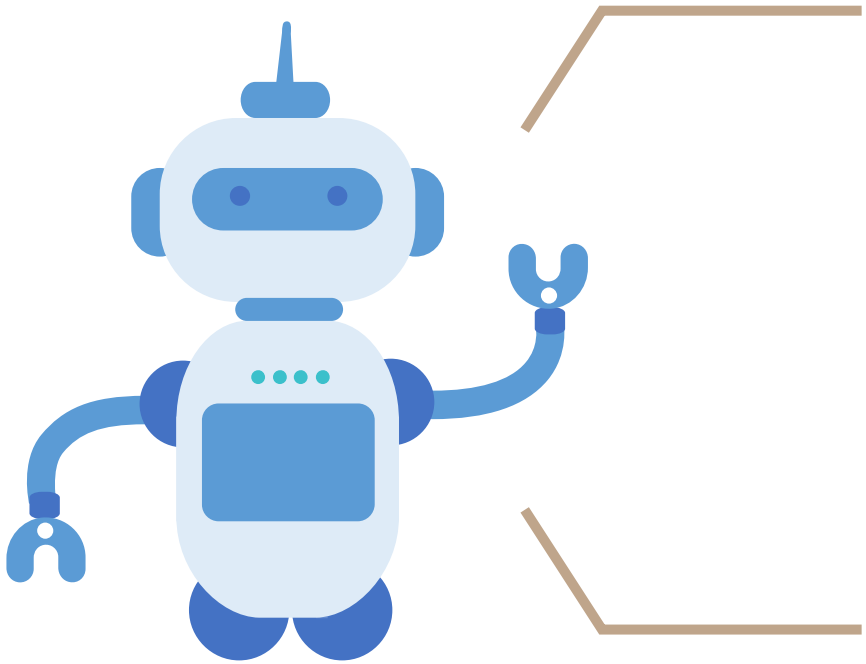


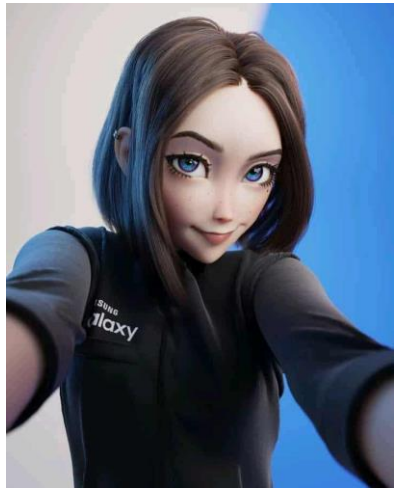
Créer son Chatbot avec Synthesia



Siri, le Chatbot d'Apple

Des exemples malheureux





Le nouveau **Chatbot** de Samsung « Sam »
, tentative avorté car trop genré

Tay le chatbot de microsoft qui dérape sur tweeter en tenant des propos racistes

La meilleure solution c'est de ne pas faire de représentation physique de l'assistant(e) comme le fait SIRI ou bien Google Assistant. Avec 3 options de voix: Neutre, masculine, féminine. Ce qui permet d'éviter de détourner l'image de l'assistant.

Un cadre éthique
Construire des valeurs réglementées

Comment cadrer les usages de l'IA

De nombreuses initiatives nationales ou internationales tentent de cadrer les usages de L'IA d'un point de vue éthique ou protection des données personnelles (voir chapitre dédié dans ce parcours). Toutefois à l'heure où nous élaborons ce cours (début 2022) les choses bougent mais ne sont pas entérinées. Ci-dessous vous retrouverez des acteurs de la réglementation et un début de charte.

Le CNPEN

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.

Des réflexions sont menées sur les valeurs qui orientent et motivent nos actions individuelles et en groupe:

- Respecter la culture, les biens communs
- Rendre plus transparente la conception et l'apprentissage des machines
- Auditer les systèmes, éviter les bad-nudges
- Remettre l'humain au cœur du débat

Comment cadrer les usages de l'IA

De nombreuses initiatives nationales ou internationales tentent de cadrer les usages de L'IA d'un point de vue éthique ou protection des données personnelles (voir chapitre dédié dans ce parcours). Toutefois à l'heure où nous élaborons ce cours (début 2022) les choses bougent mais ne sont pas entérinées. Ci-dessous vous retrouverez des acteurs de la réglementation et un début de charte.

Le CNPEN

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.

Des réflexions sont menées sur les valeurs qui orientent et motivent nos actions individuelles et en groupe:

- Respecter la culture, les biens communs
- Rendre plus transparente la conception et l'apprentissage des machines
- Auditer les systèmes, éviter les bad-nudges
- Remettre l'humain au cœur du débat

Une tentative d'écriture de charte

Alors comment garder le contrôle de ces masses de données et comment en extraire des données sans impacter négativement la vie privée des citoyens qui sont à l'origine de ces données ?

Une tentative d'écriture d'une charte par *Maël Pegny, Chercheur post-doctoral en Ethique en IA à l'Université de Tübingen*
Dans un modèle d'apprentissage machine, la distinction entre programme et données n'est pas claire car les paramètres du programme sont déterminés par entraînement sur une base de données particulières. Certaines attaques permettent une reconstitution des données d'entraînement à partir des informations encodées dans les paramètres du modèle : on parle alors de **"rétro-ingénierie" des données**. Si le modèle a été entraîné sur des données personnelles, on peut ainsi retrouver celles-ci, même si elles ont été détruites après l'entraînement du modèle.

Un pouvoir prédictif trop fin d'un modèle d'IA peut poser des problèmes d'éthique. Attention toutefois à ne pas confondre le problème de pouvoir prédictif trop fin avec la suroptimisation ou le phénomène de sur-apprentissage (l'apprentissage des données par cœur plutôt que de caractéristiques généralisables);

La charte introduit un certain nombre de principes pour des IA respectueuses de la vie privée mais dont la mise en œuvre n'est pas toujours évidente

Une tentative d'écriture de la charte

Alors comment garder le contrôle de ces masses de données et comment en extraire des données sans impacter négativement la vie privée des citoyens qui sont à l'origine de ces données ?

Une tentative d'écriture d'une charte par Maël Pegny, Chercheur post-doctoral en Ethique en IA à l'Université de Tübingen

Dans un modèle d'apprentissage machine, la distinction entre programme et données n'est pas claire car les paramètres du programme sont déterminés par entraînement sur une base de données particulières. Certaines attaques permettent une reconstitution des données d'entraînement à partir des informations encodées dans les paramètres du modèle : on parle alors de **"rétro-ingénierie" des données**. Si le modèle a été entraîné sur des données personnelles, on peut ainsi retrouver celles-ci, même si elles ont été détruites après l'entraînement du modèle.

Un pouvoir prédictif trop fin d'un modèle d'IA peut poser des problèmes d'éthique. Attention toutefois à ne pas confondre le problème de pouvoir prédictif trop fin avec la suroptimisation ou le phénomène de sur-apprentissage (l'apprentissage des données par cœur plutôt que de caractéristiques généralisables),

La charte introduit un certain nombre de principes pour des IA respectueuses de la vie privée mais dont la mise en œuvre n'est pas toujours évidente

- ***Principe 1*** *– Dans le cadre de recherches scientifiques, déclarer les finalités et l'extension nécessaire de la collecte, puis apporter une justification scientifique à tout écart à cette déclaration initiale, en discutant ces possibles impacts sur la vie privée .*
- ***Principe 2*** *– Tester et questionner les performances finales du modèle par rapport à la finalité déclarée, et veiller à éviter l'apparition d'un pouvoir prédictif trop fin*
- ***Principe 3*** *– Prendre en compte le respect de la vie privée dans l'arbitrage entre suroptimisation et perte de performances.*
- ***Principe 4*** *– Entraîner son modèle sans faire usage de données personnelles. Si cela est impossible, voir les principes plus faibles 5 et 6.*
- ***Principe 5*** *– Entraîner son modèle sans faire usage de données personnelles dont la diffusion pourrait porter atteinte aux droits des personnes.*
- ***Principe 6*** *– Entraîner son modèle sans faire usage de données ayant fait l'objet d'un geste explicite de publication.*
- ***Principe 7*** *– Si le recours à des données personnelles est inévitable, déclarer les raisons justifiant ce recours, ainsi que les mesures prises contre la rétro-ingénierie des données et leur complétude par rapport à l'état de l'art.*
- ***Principe 8*** *– Diffuser en licence libre tous les outils de lutte contre la rétro-ingénierie.*
- ***Principe 9*** *– Si le principe 8 n'entraîne pas de risque de sécurité intolérable, mettre le modèle à disposition de tous afin que chacun puisse vérifier les propriétés de sécurité, et justifier explicitement la décision prise.*
- ***Principe 10*** *– La restriction de l'accès à un modèle entraîné sur des données personnelles ne peut être justifiée que par des enjeux d'une gravité tels qu'ils dépassent les considérations précédentes. Cette exception doit être soigneusement justifiée, l'emploi du modèle devant être réduit dans sa temporalité et ses modalités par les raisons.*

Des exemples d’éthique

Google

- Etre socialement bénéfique
- Éviter de créer ou de renforcer des préjugés injustes
- Etre élaborer et tester pour la sécurité
- Être responsable envers les gens
- Incorporer les principes de conception de la confidentialité
- Maintenir des normes élevées d'excellence scientifique
- Être mis à disposition pour des utilisations conformes à ces principes.

Microsoft (Justice, Fiabilité et sécurité, transparence , Confidentialité et sécurité, incursion, responsabilité).

En conclusion :

Nous avons besoin de politiques et de cadres réglementaires nationaux et internationaux pour garantir que ces technologies émergentes profitent à l'humanité tout entière.

Nous avons besoin d'une IA centrée sur l'humain, qui servirait l'intérêt supérieur des citoyens, et non pas l'inverse.