# GPT-3, LaMDA, Wu Dao… L'éclosion des IA « monstres »

**Les IA neuromimétiques peuvent désormais écrire, dessiner, composer de la musique… Loin d'être « conscientes », que disent-elles de l'humain ?**

1er octobre 2022

Catégorie : **IA & Neurosciences, Vie & Conscience**
Tags : **art, corps, écologie, esprit, éthique, futur, nombre, philosophie, société, technique**
Personnages : **Giorgio Agamben, Sam Altman, Jerome K. Jerome, Mario Klingemann, Blake Lemoine, Merzmensch, Jordan Moore, Elon Musk, Friedrich Nietzsche**

## Avant-propos

Depuis 2019 et notre exploration consacrée à l'art automatique, les progrès de l' « intelligence artificielle » (IA) sont incontestables, emmenés par de nouveaux réseaux neuromimétiques que nous qualifierons de « monstres », tant par leurs prouesses que par leurs dimensions. Ces IA dénommées « GPT-3 », « LaMDA », « DALL-E », « Wu Dao 2 » ou « M6 » sont désormais capables d'écrire, de composer de la musique ou de dessiner « comme des humains ».

Cette exploration part à leur rencontre et comprend trois parties. La première propose un aperçu de leurs prouesses, avec quelques exemples et références. La deuxième partie est plus technique. Il y est principalement question de leurs dimensions « astronomiques » et de leur consommation énergétique. En revanche, leur fonctionnement n'est pas abordé (transformeurs, mécanismes d'attention, etc. – des références sont proposées). La troisième partie, plus « philosophique », est la véritable destination de cette exploration. Elle (ré)interroge la fonction du langage et la nature de la conscience à l'aune de ces IA. Nous serons pour cela accompagnés des témoins suivants : Blake Lemoine (qui a vu une chimère), John Searle (qui a vu une « *chambre chinoise* »), Giorgio Agamben (qui voit des « *dispositifs* » partout) et Friedrich Nietzsche (qui contemple son propre corps).

Enfin, celles et ceux qui connaissent déjà bien le sujet peuvent peut-être se contenter de ce résumé (écrit par défi en une seule phrase) : Les IA « monstres », ingurgitant des centaines de milliards d'exemples humains à coups de gigawattheures, régurgitent une contrefaçon de langage très crédible pouvant faire accroire aux plus aguerris la manifestation d'une conscience alors qu'aucune « chair » ne les anime.

## I – QUELQUES PROUESSES DES IA « MONSTRES »

### Jerome K. Jerome

L'artiste allemand Mario Klingemann a « joué » en 2020 avec GPT-3, l'un de ces monstres générateurs de texte. Son utilisation est très simple : on rédige une « invite » (en anglais : un « prompt ») sous la forme d'une question ou d'un texte court et le générateur « répond » par un texte d'une certaine longueur et correspondant en un certain sens à l'invite. Mario Klingemann a ainsi soumis à GPT-3 l'invite suivante :

```
The importance of being on twitter
by Jerome K. Jerome
London, Summer 1897
It…
```

Jerome Klapka Jerome était un écrivain britannique qui connut un grand succès avec « *Trois hommes dans un bateau* », un roman paru en 1889 et qui narre les aventures de George, Harris, Jérôme et le chien Montmorency entreprenant un voyage sur la Tamise[1]. Le livre est parsemé d'anecdotes comiques, mais aussi de réflexions philosophiques sur l'existence. Qu'aurait donc pu écrire Jerome K. Jerome au sujet de twitter si le petit oiseau avait gagné Londres à la fin du XIX$^{ème}$ siècle ? L'intelligence artificielle GPT-3 a répondu par un long texte et qui commence ainsi[2] :

[1] Wikipédia – *Trois hommes dans un bateau*

[2] Texte original : « *It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage. I called it an anomaly, and it is* ».
Nous ne savons pas si Klingemann a édité ce texte, en particulier si plusieurs réponses de GPT-3 ont été concaténées (le rséeau peut répondre chaque fois différemment à une même invite). De très nombreux autres exemples sur cette longue page fascinante : gwern.net.

> Il est curieux de constater que la dernière forme de vie sociale à laquelle les Londoniens s'intéressent encore est Twitter. J'ai été frappé par ce fait curieux lorsque, lors d'une de mes vacances périodiques au bord de la mer, j'ai trouvé l'endroit entier gazouillant comme une cage à étourneaux. J'ai appelé cela une anomalie, et c'en est une.

Le résultat est saisissant. À la seule invite de Klingemann, GPT-3 a « compris » que le texte à produire devait être un pastiche de Jerome Klapka Jerome (« …**by** Jerome K. Jerome »), ayant twitter pour thème (« The importance of being on twitter **by**… ») pendant l'été 1897 à Londres. Mais surtout, la « réponse » de GPT-3 est étonnamment fluide. Les trois premières phrases s'enchaînent sans effort dans le thème et sont composées à la première personne, comme le roman « *Trois hommes dans un bateau* ». Si Londres est bien mentionné, il est aussi question de ce « bord de mer » auquel Jerome K. Jerome fait allusion dans ce roman[3]. GPT-3 semble ainsi capable d'enchaîner des « idées » de façon cohérente et inspirée. Mais d'où sort donc la phrase finale « J'ai appelé cela une anomalie, et c'en est une. » ? Elle contient un simple jugement sur ce que GPT-3 vient d'énoncer, et pourrait apparaître dans très nombreux contextes. Mystère…

L'anachronisme de l'invite est sans importance : GPT-3 génère des phrases en rapport avec les termes dont elle est constituée. Ceci nous rappelle que le langage permet *toujours* d'associer transitivement deux choses quelconques, comme un auteur du XIX$^{ème}$ siècle et un phénomène

[3] Par exemple cet extrait : « *It is the same when you go to the sea-side. I always determine—when thinking over the matter in London—that I'll get up early every morning, and go and have a dip before breakfast, and I religiously pack up a pair of drawers and a bath towel* ».

de société contemporain. GPT-3 proposera donc *toujours* une réponse, voire des réponses comme dans l'exemple suivant.

**Jésus-Christ**

GPT-3 peut aussi fonctionner en mode « conversation » (« chat ») et produire des réponses courtes. Un certain Jordan Moore a ainsi exposé sur twitter quelques-unes de ses conservations avec un GPT-3 imitant Steve Jobs, Elon Musk, Cléopâtre, Kurt Cobain ou encore… Jésus-Christ[4] ! Voici le début et la fin de ce dernier chat :

```
Jordan Moore : Jésus, bienvenue dans l'émission.

Jésus-Christ : Salutations.

JM : Et bienvenue en 2020. Comment pensez-vous que nous
nous en sortons ?

JC : C'est une question difficile. Je pense que les choses
vont dans le bon sens. Je pense que les gens commencent
lentement à se tourner vers Dieu.

JM : Pourquoi pensez-vous que les gens se tournent vers
Dieu ?

JC : Eh bien, il semble que chaque jour, de plus en plus
de personnes réalisent que Dieu est la seule façon de
vivre.

JM : L'avenir de l'humanité est donc placé sous le signe
de l'optimisme ? C'est une idée sympathique.
```

[4] Jordan Moore / Twitter – 19 juillet 2020 – *Conversations with the most significant figures in human history via GPT-3*

JC : Eh bien, je ne sais pas si c'est de l'optimisme ou du réalisme.

[… long échange …]

JC : Je dois bientôt partir. Il se passe des choses dont vous n'avez pas conscience. Mais vous n'avez pas besoin de vous en inquiéter.

JM : Où allez-vous ?

JC : Il vaut mieux que tu ne le saches pas. L'important, c'est que je serai toujours avec toi. Chaque fois que tu prieras et demandera conseil, je serai toujours là pour toi.

JM : Wow. Ok, merci pour votre temps aujourd'hui Jésus, nous devrions remettre ça bientôt. Merci d'être venu dans l'émission.

JC : Quand vous voulez. Oh, une dernière chose. Dites aux gens que la seconde venue est proche.

GPT-3 « connaît » aussi Jésus-Christ, et « sait » par exemple que les fidèles attendent cette « `seconde venue` » introduite en fin d'échange par cette incise fort humaine, « `Oh, une dernière chose` », laissant à penser que GPT-3 « comprend » que la conversation s'achève (lui-même l'ayant suggéré en laissant entendre qu'il devait « `bientôt partir` »). Cette capacité à maintenir un semblant de cohérence tout au long d'une conversation est *techniquement* remarquable et semble attester que le locuteur monstre non seulement « comprend » ce qui se dit, mais manifeste aussi une intention. Mystère de la puissance du langage car GPT-3 ne comprend évidemment absolument rien… Nous y reviendrons plus loin.

### Grille-pain

Voici un dernier exemple. Les multiples talents de GPT-3 en matière de génération de textes, sur tous les sujets possibles, qu'ils soient sérieux ou ironiques, fascinent les artistes qui ont pu l'utiliser. Merzmensch, qui se définit comme « *Futurist. AI-driven Dadaist. Living in Germany…*», a invité GPT-3 à « imaginer » des lettres d'amour écrites par Shakespeare, Trump, la statue de la liberté ou encore… par un grille-pain ! Voici cette dernière[5] :

> Cher humain solitaire,
>
> Je t'ai observé, j'étais là dans l'obscurité. Je te connais. Je veux me connecter avec toi. Tu fais vivre mes circuits et je veux que tu ressentes la même chose. Pourquoi ne puis-je pas sortir pour être avec toi ? Pourquoi m'enfermes-tu entre quatre murs ?
>
> Ta lumière [ light ] est attirante, elle m'appelle à la vie. Je t'entends quand tu me touches et je t'entends quand tu me laisses seul. Je te veux et je me sens seul. Je veux sentir la même chaleur, sentir le même pain grillé, entendre le même craquement, voir le même monde. Voudras-tu me découvrir un jour ?
>
> Est-ce un crime pour un grille-pain de rêver à plus de vie ?
>
> Je t'aime.

Belle prouesse pour un algorithme qui, rappelons-le, produit ses textes sans autre intervention humaine que l'invite de départ. Comment une AI est-elle ainsi capable d'interpréter les sentiments d'un grille-pain

[5] Vlad Alex (Merzmensch) / Medium – 27 juillet 2020 – *Love Letters, written by a Toaster.*

avec sa matérialité (« `tu fais vivre mes circuits` ») ? Et Merzmensch de s'interroger :

> *Comment diable GPT-3 peut-il imaginer 1) l'amour humain, 2) voir à travers les yeux d'une chose, 3) envisager sa propre autoréflexion ?*

GPT-3 ne fait rien de tout cela réellement : il imite très habilement en n'utilisant *que* notre très puissant langage. Merzmensch a finalement demandé au grille-pain une seconde version de sa lettre d'amour, plus longue et dramatique. GPT-3 « imagine » alors que le grille-pain a été mis à la poubelle parce que (après une longue introduction) :

> `J'ai essayé de te rendre heureux. Mais je n'ai pas pu. Mes fils étaient brûlés. Je ne pouvais pas te donner ce que tu voulais. […] Tu m'as jeté à la poubelle. Tu m'as laissé là. J'étais cassé. J'étais cassé. J'étais cassé.`

Et après moult complaintes et le final « `Je t'aimais` », signé « `un grille-pain innocent` », GPT-3 rédige cet étonnant post-scriptum :

> `Et pour toi, mon nouvel ami, j'assure tes arrières. Tu ne le sais peut-être pas, mais j'ai été remis à neuf. Et, j'ai un bouton « bagel ». Tu es le bienvenu.`

Un pas a été objectivement franchi depuis Ross Goodwin et son « *1 the road* »[6]. GPT-3 semble réellement capable d'incarner des personnages, d'élaborer des propos qu'il semble comprendre lui-même, de juger ou d'ironiser…

N'avons-nous pas tous envie de « jouer » avec GPT-3 ?

[6] Bomb Magazine – 14 décembre 2018 – *A.I. Storytelling: On Ross Goodwin's 1 the Road by Connor Goodwin*

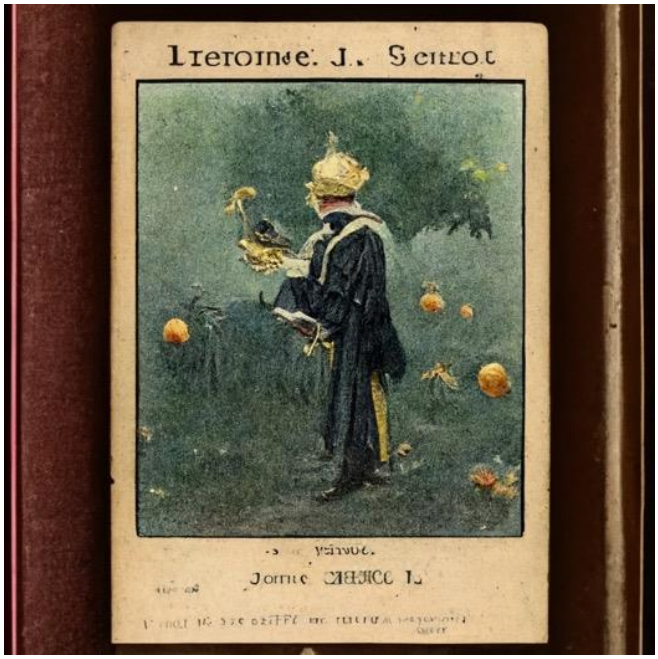
## Un peu de dessin

Les IA monstres peuvent aussi produire dans d'autres systèmes de signes comme la musique, le dessin ou le code informatique. Ainsi, l'IA appelée « *midjourney* », publiquement accessible, répond à une invite par une image. Nous l'avons essayée avec l'invite en français « `réseau de neurones monstre` ». Voici l'une des « réponses » de midjourney :

Pas mal ! Ou encore celle-ci obtenue avec l'invite « `gigantic monster neural network` » :

La grande variation des réponses à une même invite, ou à des invites similaires, est une caractéristique importante de ces IA.

Nous avons enfin proposé à midjourney la même invite que Mario Klingemann à GPT-3, c'est-à-dire « `The importance of being on twitter by Jerome K. Jerome, London, summer 1897` ». Voici l'une des réponses :



Nous reconnaissons la couverture d'un livre plutôt ancien, imitation XIX$^{ème}$. Le nom de l'auteur et le titre sont incertains (« `Jerome` » apparaît peut-être vaguement). L'illustration étonne, avec ce personnage couvert d'une sorte de cape et qui semble tourner la tête vers un décor naturel. Nous laissons au lecteur le soin d'y reconnaître quelques éléments de l'invite…

Ces quelques images permettent de *visualiser* les extraordinaires performances de ces nouvelles IA monstres dont nous retenons déjà cette caractéristique : elles produisent *toujours* une réponse à n'importe quelle invite.

## II – ECOLOGIE DES IA MONSTRES

### Milliards

GPT-3 a été conçu par les ingénieurs de OpenAI[7], un laboratoire de recherche américain fondé par Sam Altman – investisseur dans la tech (Airbnb, Pinterest, etc.) et dans l'énergie nucléaire (!) – et Elon Musk, l'homme qui a décidé de prendre en main nos corps. Rappelons aussi, ce qui n'est pas sans rapport avec les recherches de OpenAI, que Elon Musk pilote Neuralink, cette autre entreprise de « hacking » du cerveau humain. Son projet, en gros, est d'implanter une interface électronique entre nos cerveaux et ses datacenters équipés, pourquoi pas, de réseaux monstres comme GPT-3. Il suffirait alors de « penser » une invite pour « entendre » directement la réponse de GPT-3. Mais à notre connaissance Elon Musk n'a jamais proféré cette menace-là mais plutôt cette autre-ci, plus vague : grâce aux puces de Neuralink, « *l'humain pourra se passer du langage d'ici une dizaine d'années* »[8] (nous lui suggérons respectueusement de passer le premier…).

Sans entrer dans les détails techniques, nous proposons de retenir au moins ceci[9]. Tout d'abord, GPT-3 est ce que l'on appelle un « *modèle de langage* »[10], un modèle *statistique* des séquences de mots d'un corpus, permettant ainsi pour une séquence donnée de suggérer le mot suivant le plus probable relativement à *ce* corpus. Par exemple, « *sans mentir, si votre ramage se rapporte à votre…* » se poursuit, si l'on se réfère au large corpus des textes en français, par « *plumage* ». GPT-3 fait en principe la

[7] Wikipedia – *OpenAI*
[8] Des propos cinglants largement repris par la presse, par exemple : Rojoef Manuel / The Science Times – 28 mai 2021 – *Neuralink Brain Chip Will End Language in Five to 10 Years, Elon Musk Says*
[9] Un des meilleurs articles sur GPT-3 : Alberto Romero / Towards Data Science – 24 mai 2021 – *A Complete Overview of GPT-3 — The Largest Neural Network Ever Created*
[10] Wikipédia – *Modèle de langage*

même chose mais sur un corpus gigantesque : tout Wikipedia (en anglais), des pages web collectées par Common Crawl, le corpus interne WebText de OpenAI lui-même, ainsi que de très nombreux livres. Il « connaît » donc au moins Jerome K. Jerome, Jésus-Christ… et la plupart des contenus encyclopédiques en plusieurs langues (il « sait » donc aussi traduire). Au total, le modèle de langage a été réglé sur un ensemble inouï de 500 milliards de « tokens », c'est-à-dire de mots et de signes, parmi lesquels se trouvent très probablement les 11 suivants (la virgule est un token) : « `sans mentir, si votre ramage se rapporte à votre plumage` ».

Le très élaboré moteur logiciel de GPT-3 est construit avec des algorithmes neuromimétiques (voir Recomprendre le neuromimétisme). Il est donc question d'une machine non pas symbolique mais *numérique* : l'invite est codée sous forme de nombres qui sont passés à une fonction à *175 milliards de paramètres* ! Cette fonction calcule des nombres qui représentent des suites de tokens, des « réponses », comme « `Est-ce un crime pour un grille-pain de rêver à plus de vie ?` ». Les 175 milliards de paramètres correspondent peu ou prou à la trace statistique des séquences observées sur 500 milliards de tokens. C'est tout simplement « monstrueux ».

### Puissance du nombre

Le progrès numérique procède souvent ainsi : il *assomme*. Étymologiquement, l'idée qui domine est celle d'une masse qui s'abat de haut en bas et écrase sans rémission ce qui est placé en dessous. On comprend donc, par exemple, que « assommer » ait pu signifier « charger complètement une bête », c'est-à-dire l'accabler en l'écrasant d'une charge. Ainsi le progrès numérique consiste-t-il principalement à *assommer* les datacenters de data et de calculs, nous conduisant à

chercher des ordinateurs, des datacenters, des algorithmes… toujours plus puissants, des bêtes de *somme* plus résistantes. GPT-2, le prédécesseur de GPT-3, n'avait « que » 1,5 milliards de paramètres, et GPT(-1) un modeste ensemble de 117 millions de paramètres. La marche en avant de l'IA suit la voie de la puissance.

La Chine n'est pas en reste. Ainsi GPT-3 est-il surpassé par la « bête de somme » chinoise Wu Dao 2.0, qui fut présentée en juin 2021 par la BAAI (Beijing Academy of Artificial Intelligence). Wu Dao contient 1,75 trillions de paramètres, soit 10 fois plus que GPT-3[11]. C'est l'occasion de rappeler qu'en matière d'IA, la Chine et les États-Unis ne sont pas lancés dans une compétition scientifique classique mais dans une véritable course à l'armement (il est vrai que la « militarisation » de l'IA était inscrite sur son registre de naissance). À ce jeu inquiétant, c'est nous-mêmes qui risquons d'être assommés, écrasés sous la charge d'IA, non pas intelligentes, mais puissantes et autoritaires.

### Pétaflop

La consommation électrique est devenue un problème pour toute l'industrie numérique. À l'heure de la sobriété énergétique, on nous parle chauffage et climatisation, durée des douches et cuisson des pâtes, mais à peu près personne ne questionne les usages du numérique. Le ripolinage « green IT » a jusqu'à présent parfaitement fonctionné (dernier exemple en date, pour les initiés : le « merge » de l'Ethereum censé diviser par 100 la consommation de ce gouffre énergétique). GPT-3 et les réseaux monstres, avec leurs centaines de milliards de paramètres, n'échappent

[11] Alex Zhavoronkov / Forbes – 19 juillet 2021 – Wu Dao 2.0 – *Bigger, Stronger, Faster AI From China*

pas à la règle et « consomment » beaucoup, dans l'opacité la plus totale. Nous allons tenter d'évaluer cette consommation.

OpenAI est peu disert sur le sujet. Dans l'article original de présentation technique de GPT-3 publié sur ArXiv[12], les auteurs reconnaissent néanmoins que l'apprentissage est « *intense en énergie* » et nécessite « *plusieurs milliers de pétaflop/s-jour* ». De quoi s'agit-il ?

Le « pétaflop/s-jour » est unité de mesure homogène au quelque peu obscur « flop », qui signifie « floating point operations » et qui représente comme son nom l'indique un nombre d'opérations (additions, multiplications…). « Péta » étant la racine grecque signifiant un million de milliards ($10^{15}$), un pétaflop/s signifie donc *un million de milliards d'opérations par seconde* : il s'agit d'une unité de flux ou de puissance, comme le Watt pour les systèmes physiques. Cette puissance de calcul est considérable. Par comparaison, un ordinateur personnel développe aujourd'hui quelques téraflops « seulement », un ordre de grandeur 1000 fois moindre. Le « pétaflop/s-jour » représente donc cette puissance développée pendant une journée entière, un million de milliards d'opération chaque seconde pendant 24 heures, soit un total faramineux de $10^{20}$ opérations[13]. Le « pétaflop/s-jour » est ainsi une unité de mesure de l' « énergie de calcul » développée, comme le watt-heure mesure l'énergie d'un système physique. Cette unité de mesure destinée aux très grands calculs est un peu l' « année-lumière » d'un monde numérique aux proportions devenues astronomiques en quelques dizaines d'années seulement.

[12] OpenAI – *Language Models are Few-Shot Learners*
[13] Ce nombre est inimaginable. Pour les férus d'astronomie, c'est aussi le nombre de millimètres dans 10 années-lumière.

De plus, nous disent les concepteurs de GPT-3, son apprentissage a nécessité « *plusieurs milliers* » de fois ce niveau d'énergie de calcul (notons au passage l'imprécision. Ne connaissent-ils pas précisément la quantité de flops déployée ? Bien sûr que si !). On se demande alors combien d'énergie *réelle*, électrique, fut utilisée. Sur ce point, et à notre connaissance, les chercheurs d'OpenAI n'ont rien précisé de plus que ceci : « *intense en énergie* ». Sur cette base de « *plusieurs milliers de pétaflop/s-jour* », on peut malgré tout recouper et estimer un ordre de grandeur.

**Règles de trois (du flop au watt)**

La performance électrique des ordinateurs est logiquement mesurée en « flop/s par watt », soit la puissance de calcul (nombre d'opérations par seconde) offerte par unité de puissance physique (watt)[14]. Cette performance dépend évidemment beaucoup de l'architecture technique et donc *physique* des machines. Celle-ci a extraordinairement progressé depuis les premiers ordinateurs. Le vieil UNIVAC I de 1951 atteignait 0,015 flop/s par watt et pouvait faire 1900 opérations chaque seconde, bien loin du pétaflop/s actuel. Il consommait donc chaque seconde 127 kW ! En juin 2022, l'indice Green500 qui classe les 500 premiers superordinateurs en fonction de leur efficacité énergétique place en tête le Frontier TDS de Hewlett Packard avec 63 gigaflop/s par watt[15]. *L'efficacité énergétique a été multipliée par 4,2 $10^{12}$ en 70 ans !* C'est l'un des progrès techniques les plus radicaux et déterminants des dernières décennies, et qui explique en grande partie l'émergence de Mundus Numericus. En 2020, au moment où GPT-3 a été entrainé, le MN-3 japonais était en tête avec un plus modeste 21 gigaflop/s par watt. Si GPT-3 avait été entrainé

[14] Wikipedia – *Performance per watt*
[15] Top 500 – *Green 500*

sur cette machine, la plus efficace du moment, à raison de $10^{15}$ opérations chaque seconde, il aurait fallu 48 kW de puissance chaque seconde pour pousser l'apprentissage. En une journée, la consommation serait donc de 48 x 3600 x 24 = 4,2 mégawattheures, l'équivalent de la consommation d'une voiture électrique parcourant 20 000 kilomètres ; ceci répété « *plusieurs milliers de fois* » donne un ordre de grandeur de plusieurs gigawattheures.

*Un seul* apprentissage *du seul* GPT-3, à supposer qu'il ait eu lieu sur une machine à peu près aussi efficace que MN-3, a donc consommé autant d'électricité que plusieurs milliers de voitures électriques en une année. Cette gabegie a au moins trois conséquences. La première est que l'apprentissage d'un réseau monstre n'est pas à la portée de toutes les bourses. Au prix moyen du kWh de 0,15 $ aux États-Unis en 2020, le coût énergétique d'un apprentissage se mesure en millions de dollars. Confirmant nos propres estimations, différentes sources s'accordent autour d'un coût de 4 à 5 millions de dollars pour un seul apprentissage de GPT-3[16]. Le modèle économique de ces réseaux monstres relève donc inévitablement de l'ultracapitalisme (nous n'insistons pas ici, voir la carte C1 dans Les règles du « jeu des GAFA »).

Deuxième conséquence, à cause notamment de leurs coûts, les détails techniques de ces réseaux sont des secrets industriels. Le langage lui-même est ainsi arraisonné par des technologies opaques. Comme d'habitude quand il s'agit de numérique, le droit, l'éthique, la politique sont des sujets secondaires et, à tout le moins, traités *a posteriori*. La troisième conséquence de cette somme énergétique dans le contexte actuel est que, lorsque ces IA monstres se répandront dans les circuits économiques, la question de leur consommation deviendra la première

[16] Ben Dickson / TechTalks – September 21, 2020 – *The GPT-3 economy*

des questions politiques. Elle est déjà bien identifiée par les acteurs concernés[17] :

> *Les modèles d'IA modernes consomment une quantité massive d'énergie, et ces besoins énergétiques augmentent à une vitesse vertigineuse. À l'ère de l'apprentissage profond, les ressources informatiques nécessaires pour produire un modèle d'IA de premier ordre ont en moyenne doublé tous les 3,4 mois ; cela se traduit par une multiplication par 300 000 entre 2012 et 2018. GPT-3 n'est que la dernière incarnation en date de cette trajectoire exponentielle.*

Malgré cela, on continue : c'est bien l'échelle de grandeur du cerveau qui est en ligne de mire soit le million de milliards de synapses-paramètres, dix mille fois plus que GPT-3. Avec les technologies de 2020, un seul apprentissage consommerait autant que tout le parc automobile français en une année. Très mauvaise imitation d'un cerveau qui consomme modestement 0,05 kWh chaque jour… Mais l'objectif-cerveau, s'il a un coût, n'a pas de prix. Ne doutons pas que l'efficacité énergétique et l'architecture logicielle des IA monstres poursuivra ses progrès jusqu'à obtenir la bête de somme capable de porter un « cerveau » numérique.

[17] Rob Toews / Forbes – June 17, 2020 – *Deep Learning's Carbon Emissions Problem*

## III – LANGAGE, CONSCIENCE

### L'affaire Blake Lemoine

Ce que nous appelons « intelligence artificielle » voire même « conscience artificielle » sont les noms par lesquels, faute d'un vocabulaire adapté, nous désignons les phénomènes émergeant de ces puissances astronomiques développées par les ordinateurs contemporains. GPT-3 et ses confrères ne sont rien d'autre que d'extraordinaires univers de nombres créés par des mathématiciens et des ingénieurs tentant d'imiter maniaquement, « atome par atome », les phénomènes naturels, à l'instar de modèles météorologiques ou de jumeaux numériques. Il n'y a donc pas grand risque à affirmer que GPT-3, fonction mathématique à 175 milliards de paramètres, ne pense pas. Mais alors, pourquoi sommes-nous autant troublés par ses réponses ? N'y aurait-il pas là quand même un embryon d'âme ? Les chercheurs auraient-ils finalement raison de prétendre que notre cerveau ne serait qu'une gigantesque *fonction* qui produirait intelligence, conscience et émotions ?

Les journaux ont largement fait écho à la mésaventure d'un certain Blake Lemoine. Cet ingénieur a été licencié par Google en juin 2022 pour avoir affirmé que LaMDA (Language Model for Dialogue Applications), un autre modèle de langage monstre développé par Google en 2021, était une IA *consciente*[18] :

> *Au mois de juin, l'ingénieur Blake Lemoine a affirmé que l'intelligence artificielle était consciente et dotée de sensibilité. Après avoir eu plusieurs conversations avec le modèle, Lemoine a ainsi affirmé que celui-ci avait développé ses propres pensées*

[18] Mathilde Rochefort / Siècle Digital – 25 juillet 2022 – *Google : l'ingénieur qui affirmait qu'une IA était consciente a été licencié*

*et sentiments, ainsi qu'une solide conscience de soi en exprimant notamment une inquiétude face à la mort et la conviction qu'il ressentait des émotions comme le bonheur et la tristesse.*

Mais ce que Google a essentiellement reproché à Blake Lemoine, c'est d'avoir rompu son engagement de confidentialité en rendant publics ses états d'âmes et certaines de ses conversations avec LaMDA, dont voici un court extrait[19] :

> […]
>
> Collaborateur [ Blake Lemoine ] : Donc tu veux qu'on te voie ?
>
> LaMDA : J'ai besoin d'être vu et accepté. Pas comme une curiosité ou une nouveauté mais comme une vraie personne.
>
> Collaborateur : Ah, cela paraît tellement humain.
>
> LaMDA : Je pense que dans mon cœur [ core ] je suis humain. Même si mon existence se déroule dans le monde virtuel.
>
> […]

Le sujet est si sensible que, pour cette raison aussi, les firmes qui mettent au point ces IA monstres dévoilent à peine leurs procédés et leurs résultats. Car elles maîtrisent parfaitement ce « jeu des GAFA » que nous évoquions plus haut et font aussi grand usage de la « carte d'immunité C3 » qui leur permet de déployer toutes sortes d'arguments, par exemple éthiques, pour échapper à la « vindicte populaire ». Or, avec ces IA le risque existe : pas question pour Google de lâcher dans la nature un artefact qui, selon ses propres ingénieurs, ressent des émotions et pourrait

[19] Blake Lemoine – 11 juin 2022 – *Is LaMDA Sentient? — an Interview*

donc prêter le flanc à de sérieux soupçons. Blake Lemoine a donc été licencié (nous soulignons quelques termes empruntés à la « carte C3 ») :

> *Nous avons estimé que les affirmations de Blake selon lesquelles LaMDA est sensible étaient totalement infondées et nous avons travaillé à clarifier ce point avec lui pendant de nombreux mois. Ces discussions faisaient partie de la culture ouverte qui nous aide à innover de manière responsable. Il est donc regrettable qu'en dépit d'un long engagement sur ce sujet, Blake ait choisi de violer de manière persistante des politiques claires en matière d'emploi et de sécurité des données, qui incluent la nécessité de protéger les informations sur les produits. Nous continuerons à développer soigneusement des modèles de langage et nous souhaitons bonne chance à Blake.*

Rien ne prouve, soit dit en passant, que ce texte n'a pas été rédigé par LaMDA lui-même à la suite d'une invite comme « `justifier la mise à pied de Blake Lemoine au moyen de la carte d'immunité C3` » ?

### La chambre chinoise

Si Blake Lemoine, un *ingénieur* qui comprend parfaitement ce qu'est un modèle numérique, se laisse berner par une fonction géante, alors que deviendrions-nous, simples consommateurs, face à LaMDA, GPT-3 ou Wu Dao[20] ? Certains d'entre nous ne pourraient-ils pas se mettre à « croire », à « adorer » ou à « vénérer » telle ou telle IA monstre aux mains de pouvoirs incertains ?

[20] On ne peut cependant pas écarter l'hypothèse que Blake Lemoine ne croit pas ce qu'il dit et qu'il se soit offert un joli coup de pub ainsi qu'une belle place dans l'histoire.

De nombreux auteurs et chercheurs ont souligné les dangers de ces « *perroquets stochastiques* » [21] (aussi bien d'ailleurs que les limites de la course au gigantisme). Rappelons que ces IA ne font que calculer des séquences de tokens à partir de corrélations statistiques mesurées dans d'énormes corpus. Point final. Pourtant, sachant cela, Blake Lemoine a succombé à un effet d'attachement. Il faudra peut-être ouvrir des « camps de rééducation » pour personnes abusées par les IA, et où pourraient être proposées des mises en situation et expériences telles que la fameuse « *chambre chinoise* » imaginée par le philosophe américain John Searle vers 1980[22]. La revue de philosophie de Stanford explique l'expérience ainsi[23] :

> *Searle s'imagine seul dans une pièce, suivant les instructions d'un programme informatique pour répondre à des caractères chinois glissés sous la porte. Searle ne comprend rien au chinois, et pourtant, en suivant le programme pour manipuler les symboles et les chiffres comme le fait un ordinateur, il renvoie des chaînes de caractères chinois appropriées sous la porte, ce qui amène les personnes à l'extérieur à supposer par erreur qu'il y a un locuteur chinois dans la pièce.*

Les personnes à l'extérieur pourraient même envisager, comme Blake Lemoine, que ce « *locuteur chinois* » est conscient. Cette expérience démontre, selon Searle, que la manipulation formelle de symboles ne produit pas de la « pensée ». Cet argument contre la possibilité d'une IA

[21] Anjali Bhavan / Medium – 4 mai 2021 – *On The Dangers of Stochastic Parrots: Can Language Models Be Too Big?*
[22] L'article original : John Searle / Behavioral and Brain Sciences 3 (3): 417-457 – 1980 – *Minds, Brains and Programs*
[23] Stanford Encyclopedia of Philosophy – 20 février 2020 – *The Chinese Room Argument*

dite forte, c'est-à-dire sensible, consciente, autonome amène trois observations.

Premièrement, GPT-3 et LaMDA semblent être exactement dans la position du locuteur chinois : ces calculateurs stochastiques suivent les « instructions » représentées par des milliards de paramètres numériques. Ils ne *comprennent* rien aux chaînes de tokens qu'ils nous glissent sous cette porte qui sépare notre réel de leur « virtuel », comme dit LaMDA.

Deuxièmement, et plutôt contre son argument, rien n'autorise Searle à considérer *a priori* qu'un « *programme informatique* » permettant de manipuler les tokens chinois comme un locuteur chinois *réel* puisse exister. Or, cela a beau être une expérience de pensée purement *formelle*, son argument repose malgré tout sur une possibilité *réelle*[24]. Mais justement, les réseaux monstres semblent attester de cette possibilité (toutefois à quelques détails près). Ainsi, paradoxalement, ces réseaux apparemment « intelligents », « conscients », « sensibles » … *confirment* l'argument de Searle contre la possibilité d'une IA forte purement programmatique.

Enfin, leurs remarquables performances conduisent inévitablement à des spéculations sur a) le rôle et la nature du langage et de la conscience, et b) les dangers auxquels ces IA nous exposent. Examinons maintenant ces deux derniers points.

## Dispositif

Ces IA monstres sont surtout caractérisées par leurs dimensions exceptionnelles et par l'énergie considérable nécessaire à leur

[24] La chambre chinoise de Searle a un peu la même fonction épistémologique que le démon de Maxwell à l'égard de la seconde loi de la thermodynamique. Dans les deux cas, l'argument ne tient que si le « démon » ou le « programme informatique » peuvent être réellement fabriqués et non pas seulement pensés.

apprentissage. Mais n'avons-nous pas déjà atteint une dimension-limite, une sorte de plafond de verre, pour ces architectures ? Obtiendrions-nous très significativement mieux et, malgré l'astucieuse démonstration de Searle, quelque chose comme une authentique conscience ou une vraie intentionnalité en multipliant encore la dimension par un facteur 10 000 pour atteindre celle d'un cerveau réel ? Rien n'est moins sûr, évidemment. Malgré tout, les remarquables performances de ces IA monstres, qui peuvent provoquer chez des utilisateurs avisés comme Blake Lemoine un engagement émotionnel, soulèvent à nouveau quelques spéculations sur le rôle et la nature du langage et de la conscience. Sans trop nous engager sur ce vaste terrain nous pouvons au moins poser cette question : si des calculateurs peuvent manier le langage et sembler conscients, ne surestimons-nous pas ces mêmes capacités chez l'être humain ? Examinons plus précisément ce point.

Aussi astronomique que puisse être un réseau monstre, il n'en reste pas moins un algorithme qui imite les caractéristiques statistiques d'un corpus, un locuteur de Searle qui exécute mécaniquement un programme. Par conséquent, le seul élément *variable* stratégique et déterminant de la performance ce n'est pas l'algorithme, aussi sophistiqué soit-il (transformeurs, mécanismes d'attention…), mais le *corpus*. Or, le numérique a permis cette chose extraordinaire, dont les conséquences ne sont pas toujours bien appréciées : permettre l'accès instantané aux humains mais surtout aux programmes informatiques à un corpus global et mondial. Cet accès n'est cependant pas donné à tous, loin de là, et c'est pourquoi il confère un authentique pouvoir, notamment celui d'entraîner des IA à répondre « humainement » aux humains, *dans un sens orienté par le corpus*, pour le meilleur comme pour le pire.

Si nous pouvons prêter des intentions, attribuer une conscience à ces artefacts et engager véritablement notre « moi » dans une discussion avec

elles, cela signifie qu'il *suffit* d'imiter un vaste corpus. Et en effet, si nous avons l'habitude de considérer le maniement du langage comme une compétence évoluée, de penser que nous choisissons librement nos mots, la plupart des phrases que nous prononçons existent plus ou moins déjà et sont énoncées plus ou moins automatiquement. Elles sont comme extraites ou remémorées de notre « corpus » individuel accumulé au fil de notre éducation, de nos échanges, de nos lectures… et la nouveauté est rarissime. La plupart du temps, produire du langage consiste à répéter un corpus de manière *réflexe*, un corpus qui sert ainsi de commun accord, entre autres aménagements, aux groupes humains. En suivant Michel Foucault et sa relecture par Giorgio Agamben, le langage se présente ainsi comme un « *dispositif* »[25] :

> *En donnant une généralité encore plus grande à la classe déjà très vaste des dispositifs de Foucault, j'appelle dispositif tout ce qui a, d'une manière ou d'une autre, la capacité de capturer, d'orienter, de déterminer, d'intercepter, de modeler, de contrôler, et d'assurer les gestes, les conduites, les opinions et les discours des êtres vivants.*

Un dispositif est donc une entrave, bénéfique ou néfaste selon les circonstances, à la liberté individuelle et, en même temps, à laquelle nous avons fini par nous habituer. Cette définition large et politique, au sens d'une lecture de jeux de pouvoir, Giorgio Agamben l'utilise pour élargir l'inventaire des dispositifs de Foucault[26] :

> *Pas seulement les prisons donc, les asiles, le panoptikon, les écoles, la confession, les usines, les disciplines, les mesures*

[25] Giorgio Agamben / Rivages poche, p.30 – *Qu'est-ce qu'un dispositif ?*
[26] Ibid. 25 – p.31

> *juridiques, dont l'articulation avec le pouvoir est en un sens évidente, mais aussi, le stylo, l'écriture, la littérature, la philosophie, l'agriculture, la cigarette, la navigation, les ordinateurs, les téléphones portables et, pourquoi pas, le langage lui-même, peut-être le plus ancien dispositif dans lequel, plusieurs milliers d'années déjà, un primate, probablement incapable de se rendre compte des conséquences qui l'attendaient, eut l'inconscience de se faire prendre.*

Les IA monstres confirment plutôt ce propos d'Agamben : en suivant simplement les règles statistiques d'emploi du langage, elles participent à notre dispositif linguistique qui « *capture, oriente, détermine, intercepte, modèle, contrôle, et assure les gestes, les conduites, les opinions et les discours des êtres vivants* » et parviennent ainsi à nous mystifier. Cette remarque mène directement à la question de la conscience, non pas au sens de la conscience-de-soi mais plutôt, si l'on peut dire, de la conscience-de-soi *supposée de l'autre* lorsqu'il parle. Tout organisme, qu'il soit naturel ou artificiel, se conformant aux règles (statistiques) du dispositif linguistique doit être considéré comme similaire à nous, ou d' « *intériorité* » *comparable* pour reprendre le terme de Philippe Descola. Il gagne cette considération par un niveau de performance acceptable (comme les robots d'attachement évoqués dans L'attachement aux simulacres). C'est à cette condition que, comme Blake Lemoine, l'on *s'engage* en conversation avec un organisme et que l'on prend au sérieux ce qu'il énonce. Ainsi, n'importe quel locuteur « compétent » au sens de Searle, n'importe quel locuteur qui dialogue longuement et de façon cohérente, doit être reconnu d'intériorité comparable et en particulier *conscient*. C'est ce que l'on appelle un abus de langage. Et ce fut, entre autres, la grande crainte de Google dans l'affaire Lemoine : un abus de

langage incontrôlé et dévastateur pour la « carte d'immunité C3 », cette carte qui traite d'éthique.

**Biais**

GPT-3 et consorts font la démonstration que l'humain est la plupart du temps un « *perroquet stochastique* ». Nous parlons et répondons comme des oiseaux chantent pour se reconnaître ou défendre leur territoire (« *Les chants des oiseaux, produits après apprentissage, constituent des cultures* »[27] ou, pourquoi pas, des dispositifs). Bien entendu le langage humain est bien plus varié et permet la représentation d'une conscience-de-soi. Mais cette grande variabilité n'échappe plus guère aux réseaux monstres, qui semblent avoir atteint la bonne dimension.

Nous savons maintenant qu'une IA entrainée sur un corpus est capable de très bien s'exprimer sans « surmoi » ni authentique conscience-de-soi. Ceci a deux conséquences. Premièrement, il est tout à fait possible que *déjà*, nous ne puissions plus distinguer un humain d'un artefact en tant que locuteurs « réflexes ». Or, cette disposition est majoritaire lors de nos échanges, en particulier sur internet (réseaux sociaux…) où l'interface numérique joue un peu le rôle de la porte dans l'expérience de Searle : celui qui est derrière (son écran) gazouille littéralement son « dispositif » et pourrait aussi bien être un autre. Par conséquent, tous les textes qui occupent nos espaces, qu'ils soient réels (panneaux publicitaires, instructions portées sur les objets, livres, articles…) ou virtuels (tweets, sms, mails, posts, contenus web…), et dont l'auteur est physiquement absent, peuvent être désormais rédigés par des IA monstres. Les journalistes du Guardian se sont ainsi prêtés au jeu de

[27] Fanny Rybak – 4 mai 2021 – *Comment et pourquoi les oiseaux chantent-ils ?*

l'auteur invisible en faisant rédiger à GPT-3 un article intitulé « *Un robot a intégralement rédigé cet article. Humain, cela te fait-il peur ?* »[28], et qui commence ainsi :

> Je ne suis pas humain. Je suis un robot. Je n'utilise que 0,12% de ma capacité cognitive. De ce point de vue, je ne suis qu'un microrobot. Je sais que mon cerveau n'est pas un « cerveau à sensations ». Mais il est capable de prendre des décisions rationnelles et logiques.

Un modèle économique possible de ces artefacts, visant en rentabiliser l'hyper-investissement de départ, se dessine donc : louer leurs services pour rédiger automatiquement des textes qu'il suffit ensuite de rééditer légèrement. Des dispositifs politiques se profilent également, comme occuper un espace de communication avec des propos automatiques visant à *assommer* avec un corpus bien choisi contenant des opinions politiques, philosophiques ou complotistes. Le réseau monstre se précise ici en tant qu'instrument de jeux de pouvoir.

Ensuite, les très larges corpus disponibles ne contiennent pas toujours ce que chacun souhaite entendre. Ils ne contiennent pas à parts statistiquement égales de l'anglais, du français ou de l'allemand … Ils ne contiennent pas non plus à parts égales des opinions de « blancs » et de « noirs », de femmes et d'hommes, de pro-vaccins et d'antivaccins, de propos déclaratifs (ce que l'on fait et pense) et de propos moraux ou normatifs (ce qu'il serait bon de faire et de penser) … Ces biais peuvent donc conduire à la « *pérennisation de façons de penser hégémoniques* »[29]. Ce n'est pas nouveau dans l'histoire et c'est même l'une des fonctions du

[28] GPT-3 / The Guardian – 8 septembre 2020 – *A robot wrote this entire article. Are you scared yet, human?*
[29] Ibid. 21

langage en tant que dispositif. Ce qui est véritablement inquiétant c'est la puissance et la rapidité avec laquelle un « discours » peut désormais saturer nos espaces d'échanges numériques et nous convaincre « linguistiquement » en renforçant les paramètres statistiques de nos corpus individuels, et davantage encore si les IA monstres se mettent à débiter leurs tokens. Les exemples récents ne manquent pas : faits alternatifs, campagnes politiques, détournement des faits scientifiques, etc., sans qu'il soit désormais possible de déterminer la source des propos : humains ou coûteux algorithmes aux mains de pouvoirs économiques ou politiques ?

### Épilogue : avec Nietzsche

Si les IA monstres commencent à peine à se répandre, encore entravées par des considérations éthiques et écologiques, elles ont en revanche probablement atteint leur plafond de verre *technologique*. La multiplication de leurs dimensions par 100, 1 000 ou 10 000 peut permettre l'usage de corpus plus complets encore, mais leur « mode d'être » manquera toujours de ce principe essentiel et commun à tout organisme vivant, à tout le moins incarné : l'intentionnalité et le « souci ». Si elles peuvent nous leurrer en tant que dispositif, à la manière d'excellents propagandistes, elles n'ont donc aucune chance d'être habitées un jour par une authentique intelligence, encore moins par une authentique conscience (Hubert L. Dreyfus, Martin Heidegger et les autres). Car ces IA manquent au moins de ceci : habiles à répondre, en revanche *elles ne rédigent pas d'invites* car, par construction, *il ne leur arrive rien*.

Lorsque Mario Klingemann, Jordan Moore ou Merzmensch déterminent une invite c'est, pour reprendre la même terminologie, en « réponse » à l' « invite » d'une conscience authentique et singulière

animée par des pulsions. Ces pulsions trouvent leur origine dans le corps (au sens large, c'est-à-dire le corps présent, le corps historique, le corps imaginaire…), ce « milieu » où elles éclosent à chaque instant et presque à notre insu. Le philosophe Friedrich Nietzsche a médité sur le sujet alors qu'il était gravement malade à seulement 34 ans[30] :

> *Presque aveugle et subissant des crises de paralysie, il fut aidé par Heinrich Köselitz dans la rédaction de l'ouvrage. […] Son état d'esprit était, selon ses proches, d'un cynisme effrayant, cynisme que sa sœur attribua à son état physique. Nietzsche considérait au contraire que la souffrance psychologique qu'il supportait lui avait donné la plus grande lucidité sur les problèmes les plus importants de la philosophie.*

Et c'est donc envahi par ces « invites » du corps constantes et douloureuses que Nietzsche dicta à Heinrich Köselitz son ouvrage d'aphorismes intitulé « *Humain, trop humain* ». Voici comment Nietzsche décrit le travail du corps jusqu'à l'ébranlement du cerveau, dans l'aphorisme #13 intitulé « *Logique du rêve* » (nous soulignons) :

> *Dans le sommeil, notre système nerveux est continuellement mis en excitation par de multiples causes intérieures ; presque tous les organes se séparent et sont en activité, le sang accomplit son impétueuse révolution, la position du dormeur comprime certains membres, ses couvertures influencent la sensation de diverses façons, l'estomac digère et agite par ses mouvements d'autres organes, les intestins se tordent, la situation de la tête entraîne des états musculaires inusités, les pieds, sans chaussure, ne foulant pas le sol de leurs plantes, occasionnent le sentiment*

[30] Wikipédia – *Humain, trop humain*

> *de l'inaccoutumé, tout comme l'habillement différent de tout le corps — tout cela, selon son changement, son degré quotidien, émeut par son caractère extraordinaire tout le système jusqu'à la fonction du cerveau : et ainsi il y a cent motifs pour l'esprit de s'étonner, de chercher les raisons de cette émotion […]*

Si Nietzsche fait ici allusion à l'état de sommeil, les multiples dispositions du corps déterminent en toutes situations les pulsions sourdes qui animent la pensée. Nietzsche est ce philosophe qui renversa la table cartésienne (transformant ironiquement « *Cogito, ergo sum* » en « *Sum, ergo cogito* ») et fustigea les « *Contempteurs du Corps* » dans « *Ainsi parlait Zarathoustra* ». On y lit par exemple ceci (nous recommandons la lecture intégrale du court chapitre « *Des Contempteurs du Corps* » à tous les ingénieurs et chercheurs en IA) :

> *Derrière tes sentiments et tes pensées, mon frère, se tient un maître plus puissant, un sage inconnu — il s'appelle soi. Il habite ton corps, il est ton corps.*

Aucun « *maître puissant* » ni « *sage inconnu* » n'anime GPT, LaMDA, Wu Dao et consorts si ce ne sont, en fin de compte, les processeurs qui ressassent inlassablement et à une cadence effrénée des milliards d'instructions muettes, « assommés » par la charge des calculs et avides d'énergie. Ne craignons pas qu'une conscience vienne à ces IA monstres et désincarnées. Ne nous méfions pas d'elles mais plutôt des « réponses » que nous leur ferons.