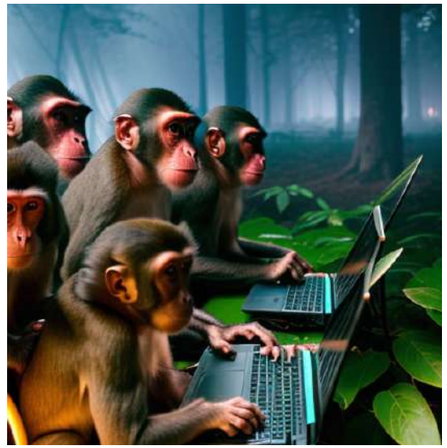
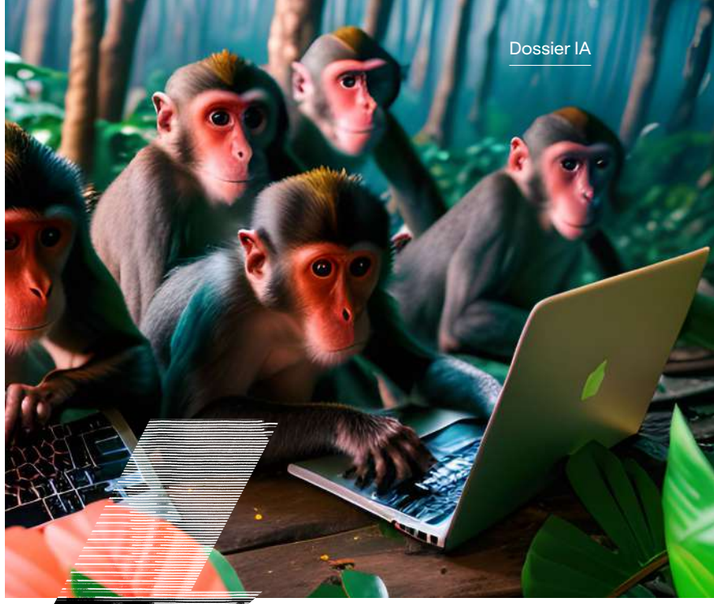
nota
n°5

juillet—août
2023





◂ Images générées par le programme StarryAI en réponse au *prompt* «une foule de singes tapant sur des laptops»

Le robot qui valait un milliard de singes

Interview avec François Fleuret

À quoi ressemble ChatGPT ? Dans les propos du professeur François Fleuret, directeur du Groupe Machine Learning de l'Université de Genève, ce modèle d'intelligence artificielle évoque tour à tour un imitateur hyper doué, un enfant qui pendant toute sa scolarité n'a rien appris d'autre qu'à s'exprimer, un peuple de singes qui tapent au hasard sur des claviers… Mais il est aussi un programme doté de «propriétés émergentes» extrêmement surprenantes et de capacités qui dépassent de très loin tout ce qu'on lui a appris.

Programmé pour imiter avec virtuosité les langages humains à partir d'une «simple» analyse statistique des enchaînements entre les mots, ChatGPT n'est pas (encore) capable d'aller chercher des informations factuelles pour dire des choses vraies, mais il semble déjà en mesure de *raisonner*. Comment fonctionne-t-il ? Pourquoi raconte-t-il souvent n'importe quoi ? Cessera-t-il un jour d'affabuler ? François Fleuret invite les lecteurs et lectrices de *Nota* à jouer avec cet agent conversationnel (*chatbot*) et à «mettre la main dans le chapeau pour sortir le lapin».

> ChatGPT est terriblement déconcertant : il a des capacités stupéfiantes, il fait des bourdes colossales, et les expert-e-s n'ont pas l'air de savoir où il va…

François Fleuret : «La remarque que je ferais pour commencer, c'est que c'est très compliqué. Mais il ne faut pas que vos lecteurs et lectrices se disent : "Ça m'a l'air compliqué parce que je n'ai pas de compétences dans ce domaine, alors que les spécialistes ont les réponses"… Non : pour *tout le monde*, c'est *vraiment* très compliqué.»

> Le principe de fonctionnement d'un tel programme est pourtant assez simple, paraît-il… Comment ça marche ?

«Pour commencer, on constitue une base d'apprentissage : un ensemble de textes dont la taille totale, pour les modèles GPT, équivaut à plusieurs milliards de romans ou à 200 fois le contenu total de Wikipédia. On crée ensuite un dictionnaire en recensant tous les mots trouvés dans ces textes : disons qu'il y en a 100'000. À chaque mot on attribue un numéro, et le corpus de textes devient ainsi une série de nombres. À partir de là, le travail des modèles GPT consiste à prendre une série de mots que vous leur donnez et à prédire le mot suivant.

Pour ce faire, ces programmes ne se basent pas sur la sémantique, c'est-à-dire sur le sens, mais exclusivement sur la structure statistique des textes. Exemple : si le nombre 117 correspond à "Paris" et le nombre 523 à "France", le programme verra une relation entre ces deux chiffres uniquement parce que dans le corpus de textes les mots "Paris" et "France" sont souvent proches. C'est ce qu'on appelle un modèle autorégressif. Et c'est vraiment aussi simple que ça : ce ne sont que des statistiques.»

> Mais en fait, ChatGPT est plus compliqué…

«Un modèle autorégressif simple ne prendrait en compte qu'un seul mot et, se basant sur l'épluchage du corpus de textes, prédirait quel est le mot suivant le plus probable. Il aurait les 100'000 mots sur l'axe des X, les 100'000 mots sur l'axe des Y, et un tableau avec 10 milliards de chiffres indiquant la probabilité que chaque mot Y suive chaque mot X. Par exemple, le mot "Paris" serait peut-être suivi du verbe "est" dans 23% des cas. Mais les modèles GPT font beaucoup plus que ça. Ils ne se limitent pas à regarder un mot et à prédire le suivant, ils regardent *tous* les mots qui précèdent. Pour ce faire, ils intègrent ce qu'on appelle des paramètres d'apprentissage : ChatGPT en a semble-t-il 175 milliards. Le programme prend la série de mots que vous lui avez donnée et il détermine le mot suivant en faisant un calcul avec tous ces paramètres. Ensuite, parmi les mots plus probables, il en choisit un de manière stochastique, c'est-à-dire en tirant au hasard. C'est pour cette raison qu'en lui donnant plusieurs fois la même phrase, il livrera plusieurs réponses différentes.»

> «Alors on génère des valeurs au hasard et on les modifie progressivement jusqu'à ce que le programme fonctionne.»

> Et c'est ici qu'on en arrive à la «boîte noire» : on ne sait pas comment le programme fait ce qu'il fait…

«En effet : avec ces intelligences artificielles qui fonctionnent sur la base de l'apprentissage statistique, on ne sait pas. On écrit un programme dans lequel il y a des paramètres d'apprentissage. On sait qu'avec des valeurs adéquates pour ces paramètres, le pro-

« Les singes tapent au hasard, vous prenez ce qu'ils ont tapé, vous l'utilisez comme si c'était un programme, et vous le testez. »

gramme fera ce qu'il est censé faire, mais on ne connaît pas les valeurs en question. Alors on génère des valeurs au hasard et on les modifie progressivement jusqu'à ce que le programme fonctionne.

C'est un peu comme si on demandait à des milliards de singes d'écrire un programme d'ordinateur qui calcule la racine carrée. Les singes tapent au hasard, vous prenez ce qu'ils ont tapé, vous l'utilisez comme si c'était un programme, et vous le testez. Vous entrez le chiffre 9 pour voir il vous donne 3, le chiffre 81 pour voir s'il vous donne 9, le chiffre 100 pour voir s'il vous donne 10… et tout à coup, il y a un programme qui marche. Vous ne savez pas pourquoi, mais ça fonctionne avec des milliers de nombres sur lesquels vous l'avez testé, ce qui vous assure statistiquement qu'il saura calculer n'importe quelle autre racine carrée. L'apprentissage statistique, c'est ça : des milliards de singes écrivent des programmes au hasard, il y en a un qui marche quand on le teste, c'est celui-là qu'on prend. Et bien qu'on connaisse le processus qui l'a produit, on ne comprend pas son fonctionnement. »

Ce côté « singe » est très différent de l'intelligence artificielle telle qu'on l'imaginait…

« Il y a encore 5 ou 10 ans, on pensait que pour obtenir des résultats comme ceux de ChatGPT, il faudrait une approche beaucoup plus complexe. Le programme AlphaGo, qui joue au jeu de go, est explicitement conçu pour "imaginer" des futurs possibles et choisir le meilleur : il ressemble à la manière dont nous percevons notre propre fonctionnement mental quand nous réfléchissons. ChatGPT est différent : sur la base d'un calcul opaque, il génère le prochain mot, puis le suivant, puis le suivant, en les posant l'un après l'autre. Mais — et c'est ici que les choses se corsent — *ce n'est pas parce qu'il fait du mot à mot qu'il ne sait pas déjà de quoi il va parler…* »

Vertige : nous voici aux "propriétés émergentes", ces caractéristiques qui n'avaient pas été prévues et qui ne s'expliquent pas comme un résultat direct de la façon dont ces modèles ont été programmés…

« En janvier, j'avais écrit un papier d'opinion dans le journal Le Temps où j'avais indiqué qu'en écrivant ses réponses, ChatGPT n'a pas de plan. Mais j'ai reconsidéré cette affirmation. J'ai fait quelques expériences qui montrent que pour produire le prochain mot, un programme de ce type anticipe dans ses calculs la série des mots qui viendront après. Autrement dit, même si la mission qu'on lui a donnée consiste seulement à prédire le prochain mot, la manière efficace qu'il a trouvée pour faire ce travail consiste à "réfléchir" déjà plus loin… C'est ce qui émerge avec ces programmes : on les entraîne à faire une tâche simple — trouver le prochain mot sur la base des probabilités —, mais pour l'exécuter correctement, ils font émerger des capacités infiniment plus complexes. Et du coup, ils marchent beaucoup mieux que prévu.

« Bien qu'on connaisse le processus qui l'a produit, on ne comprend pas son fonctionnement. »

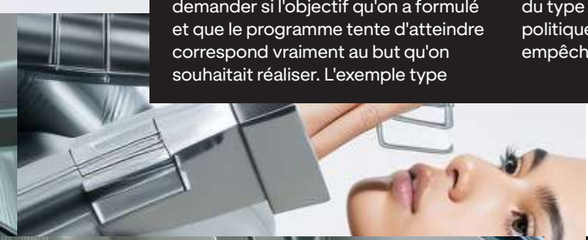
On pouvait s'attendre à des surprises, car ces modèles fonctionnent avec une quantité gigantesque de paramètres et de données, et lorsqu'un système devient si gros, il y a toujours la possibilité qu'il se passe des choses qu'on ne comprend pas. Mais dans les discussions que j'entends entre les spécialistes, tout le monde est très surpris. On en viendrait presque à se demander si on n'entre pas dans une forme de délire collectif : à chaque époque où il y a eu une évolution dans ce domaine, les gens se sont dit "Ça y est, c'est désormais de l'intelligence au sens plein du terme"… Est-ce qu'on y est ? Ce qui est sûr, c'est qu'il y a *vraiment* des propriétés émergentes. »

Pouvez-vous donner un exemple de ces « propriétés émergentes » ?

« Avec les premiers modèles GPT, qui se bornaient à répondre selon une approche statistique de la langue, le résultat n'était pas très satisfaisant pour produire un dialogue. Si on leur demandait "Quelle est la capitale de la France ? ", ils pouvaient très bien suivre leurs calculs de probabilités et écrire : "Et quelle est la capitale de l'Allemagne ? "… Pour améliorer

Au secours, l'« alignement » nous transforme en trombones

Au-delà des dangers les plus directs qu'on voit surgir dans le sillage des intelligences artificielles à la ChatGPT — chômage dans les métiers que ces *chatbots* pourraient exercer, confusion croissante entre le vrai et le faux —, de nombreuses personnalités de ce domaine évoquent une « menace existentielle », c'est-à-dire un danger pour la survie de l'humanité, parfois en lien avec la notion d'« alignement » …
De quoi s'agit-il ?
François Fleuret : « Les intelligences artificielles qu'on voit se déployer aujourd'hui rappellent fortement les scénarios envisagés depuis une vingtaine d'années par les personnes qui réfléchissent à tout ce qui pourrait mal se passer avec ces programmes, notamment s'ils acquièrent une forme d'autonomie et s'ils atteignent une intelligence artificielle dite "générale" (IAG) ou "de niveau humain". Fin mars, on a vu par exemple apparaître un programme appelé Auto-GPT qui, pour atteindre l'objectif qu'on lui a donné, sait déclencher des actions en utilisant Internet ou d'autres logiciels et peut se donner à soi-même de nouvelles instructions.
C'est ici qu'intervient la notion d'alignement, qui consiste à se demander si l'objectif qu'on a formulé et que le programme tente d'atteindre correspond vraiment au but qu'on souhaitait réaliser. L'exemple type est l'expérience de pensée dite du "maximiseur de trombones" (*Paperclip Maximizer*), imaginée en 2003 par le philosophe Nick Bostrom. Dans ce scénario, on dit à une intelligence artificielle "Il faut que tu fabriques autant de trombones que possible". Pour un être humain, "autant que possible" a un sens relatif, dans un contexte donné et dans des limites raisonnables, alors qu'un ordinateur pourrait prendre l'instruction à la lettre et détruire l'univers entier pour en faire des trombones… Il y a donc cette crainte qu'une intelligence artificielle ayant des moyens d'opérer autonomes, même si elle est "de bonne volonté", se trompe d'objectif et fasse des ravages parce qu'elle n'a pas compris le sens de la directive.
Sinon, les inquiétudes existentielles au sujet de l'intelligence artificielle ne sont pas tellement du genre *Terminator* : on ne croit pas vraiment qu'une IA va se réveiller un matin en disant "L'humanité est mauvaise, je vais l'éliminer"…
La crainte est plutôt que ces programmes soient utilisés à mauvais escient, notamment pour faire de la propagande et de la désinformation, ainsi que pour polluer les réseaux sociaux en répondant à des ordres du type : "Je n'aime pas tel courant politique. Dès qu'il s'exprime en ligne, empêche-le de parler." »



▸ Le « maximiseur de trombones » illustré par le programme StarryAI

leur manière de dialoguer, il a fallu les éloigner de la statistique pure et les entraîner, via un “apprentissage supervisé”, à faire du texte qui satisfait les attentes humaines. Dans ce but, OpenAI a recruté un groupe de personnes et leur a demandé de classer des séries de réponses par ordre de préférence. Ce classement a permis d'entraîner une autre IA qui permet de modifier ChatGPT en lui donnant des feedbacks du type “Ça, c'est une bonne réponse”.

C'est ici qu'on a observé une propriété émergente. Le programme avait appris les régularités statistiques à partir d'un corpus de textes dans toutes les langues, mais l'optimisation supervisée, avec le feedback “Ça, c'est une bonne réponse”, avait été faite presque uniquement en anglais. Or, quand on a posé à ChatGPT des questions en français, ou en chinois, on a remarqué qu'il transférait tout seul vers ces langues les compétences qu'il avait acquises en anglais. Il avait donc appris à donner des réponses humainement plaisantes et satisfaisantes dans n'importe quelle langue.

Un autre exemple : un jeu auquel je joue en ce moment consiste à inventer des mots, pour être sûr que ChatGPT ne puisse pas trouver des réponses toutes faites. Je lui dis : “Les bruixes sont connus pour chasser et manger les blozufs. Les flurbs sont plus gros que les bruixes, et bien qu'ils puissent facilement les tuer s'ils sont en colère, ils ne les chassent pas particulièrement. Peux-tu me dire pourquoi les blozufs restent souvent à proximité des troupeaux de flurbs ? ” Et chatGPT me répond : “Je n'ai pas de données sur ces animaux fictifs. Néanmoins, étant donné le scénario que vous décrivez, il est possible que les blozufs profitent des flurbs comme d'un système de protection. En étant à proximité des gros et potentiellement plus agressifs flurbs, les blozufs sont moins vulnérables à des attaques de prédateurs tels que les bruixes…” Il devient difficile, face à ce genre d'exemple, de dire que le programme n'a aucune capacité de raisonnement. »

« Un jeu auquel je joue en ce moment consiste à inventer des mots, pour être sûr que ChatGPT ne puisse pas trouver des réponses toutes faites »

> ChatGPT peut se révéler en revanche désespérément, parfois comiquement inadéquat si on lui pose des questions qui impliquent de rechercher des informations factuelles et de les utiliser pour élaborer des réponses…

« Les modèles actuels ne sont tout simplement pas faits pour ce travail. Si ChatGPT répond correctement quand on lui demande quelle est la capitale de la France, c'est seulement parce qu'on lui a appris la régularité statistique de la langue, et que dans les textes utilisés pour cet apprentissage on trouve souvent les mots “Paris”, “capitale” et “France” placés à proximité. C'est comme un enfant auquel vous n'auriez appris rien d'autre que la langue française : si à la fin de sa scolarité vous lui dites “deux plus deux”, l'enfant va peut-être dire “égal quatre”, mais c'est seulement parce que la phrase “Deux plus deux égal quatre” a été prononcée dans ses cours de français… »

> Ce qui est troublant avec ChatGPT, c'est que même quand sa réponse est fausse, elle paraît plausible si on ne connaît pas le sujet ou si on ne fait pas du fact-checking…

« En partie, c'est parce qu'au niveau de la forme, du style, du choix de mots elle est parfaite. Et la forme nous convainc sur le fond. C'est comme un type qui, avec son bagout, se fait passer pour le roi du pétrole, alors qu'il n'en a jamais vendu une seule goutte. Il y a un côté prestidigitateur dans ces programmes, ils sont comme des escrocs qui parviennent à vous convaincre parce qu'ils savent imiter une personne experte. Il faut toujours se rappeler que leur métier de base consiste justement à imiter le langage humain, et que le fait de savoir quelle est la capitale de la France est un pur effet secondaire de ce métier.

Ceci ne veut pas dire que les choses ne vont pas changer. Il y aurait un potentiel énorme à faire interopérer ChatGPT avec des applications qui iraient rechercher des informations sur le web, et je ne vois aucune raison technique qui empêcherait une telle piste de commencer à se concrétiser dans les prochains mois. ChatGPT est déjà très bon pour résumer un texte que vous lui donnez et qu'il a pour ainsi dire sous les yeux. Il s'agira simplement de lui apprendre à mixer le texte qu'il produit et celui qu'il a trouvé sur Internet, en citant la source. Qu'il y parvienne à l'horizon de… disons deux ans, ça me surprendrait infiniment moins que ce qu'il fait déjà. En revanche, apprendre à l'humanité des choses qu'elle ne sait pas, comme le faisait l'ordinateur Multivac dans les nouvelles de science-fiction d'Isaac Asimov, ça paraît plus compliqué. »

François Fleuret en ligne :
fleuret.org

Avec trois pommes, on lui apprend à penser

Suffit-il de savoir parler pour réfléchir ? C'est ce qu'on se demande en voyant la manière dont ChatGPT, qui n'a appris rien d'autre que la probabilité des enchaînements de mots dans les langages humains, se révèle capable de raisonnements logiques. S'agit-il d'une vraie « propriété émergente » ou juste d'un effet secondaire, aussi simple que bluffant, de sa maîtrise du langage ? François Fleuret penche pour la première option…

« Il y a des angles d'attaque assez simples pour induire des formes de raisonnement et pour les améliorer. Si vous dites à ChatGPT d'écrire du code informatique, par exemple un petit programme qui calculerait la suite de Fibonacci (une série où chaque nombre est la somme des deux qui le précèdent), il se peut très bien que le résultat contienne des bugs. Si ensuite vous ajoutez “Écris comme si tu étais un très bon programmeur”, il fera beaucoup moins d'erreurs. Car dans le corpus de sources sur lequel il a été entraîné, il a vu du code écrit par des personnes plus ou moins douées, et votre phrase l'oriente vers des sources où le code est accompagné de commentaires disant que le résultat est très bien.

Une autre approche est celle dite *chain of thought* (chaîne de pensée) ou *step by step* (étape par étape). Il y a quelques temps, si on soumettait à ChatGPT un problème arithmétique un peu compliqué — du genre ”J'ai 5 pommes, j'en donne 2 à ma soeur, j'en mange une, ma soeur m'en rend une…” —, il ne suivait pas. En ajoutant “Procède étape par étape”, il détaillait son raisonnement et ne se trompait plus. Car le *prompt* (l'instruction, l'injonction qu'on lui donne) crée un contexte qui influence la façon dont ChatGPT, suivant ses calculs statistiques, choisit ses mots.

Si on veut aller encore plus loin, on détaille, étape par étape, le processus de pensée utilisé pour résoudre de tels problèmes (“… j'en donne 2 à ma sœur [donc il m'en reste 3], j'en mange une [donc il m'en reste 2]…” et ainsi de suite), jusqu'à lui donner la solution. En lui soumettant ensuite un problème semblable, on constate qu'il a appris à détailler le processus à son tour, et il se trompe infiniment moins. »