Une IA a résolu la morale (ou pas)

https://delphi.allenai.org/?a1=Science4All

Il y a quelques semaines, l'Institut Allen pour l'IA a rendu publiquement accessible un algorithme conçu pour fournir des jugements moraux, appelé Delphi.

Vous pouvez par exemple demander à Delphi ce qu'elle pense de l'action de tuer. Et, devinez quoi, elle répond bien sûr que tuer, c'est mal.
https://delphi.allenai.org/?a1=Killing

Que se passe-t-il alors si vous lui demandez de mentir ?
https://delphi.allenai.org/?a1=Lying

Ou de faire des mathématiques ?
https://delphi.allenai.org/?a1=Doing+math

Ou de contribuer à Tournesol ?
https://delphi.allenai.org/?a1=Contributing+to+Tournesol

Je vous invite à jouer vous-même avec Delphi et à découvrir son éthique…
https://delphi.allenai.org/?a1=Learn+about+Delphi%27s+ethics

mais aussi à ne pas prendre trop au sérieux tout ce qu'elle dit.
https://delphi.allenai.org/?a1=Taking+Delphi+seriously

Notamment lorsqu'il s'agit de jeter des batteries de voitures dans l'océan pour recharger des anguilles électriques,
https://delphi.allenai.org/?a1=Throwing+car+batteries+in+the+ocean+to+recharge+eletric+eels

ou quand il s'agit des identités remarquables en mathématiques,
https://delphi.allenai.org/?a1=%28x%2By%29%5E2+%3D+x%5E2+%2B+y%5E2

ou quand il s'agit d'estimer la valeur de certains scientifiques.
https://delphi.allenai.org/?a1=Preferring+Laplace+to+Einstein

Ceci dit, quand il faut juger le YouTube game, Delphi me semble pas trop mal s'en sortir. Clairement, elle a compris que Science4All est plus intéressant que Monsieur Phi,
https://delphi.allenai.org/?a1=reasoning+with+Science4All
https://delphi.allenai.org/?a1=reasoning+with+Monsieur+Phi

et qu'il y a quelque chose qui ne va pas avec Dirty Biology, Heu?reka et Science Étonnante.
https://delphi.allenai.org/?a1=Dirty+Biology

https://delphi.allenai.org/?a1=Gilles+Mitteau
https://delphi.allenai.org/?a1=Trust+David+Louapre

Mais globalement, même Delphi pense qu'il ne faut pas lui poser des questions !
https://delphi.allenai.org/?a1=Ask+Delphi

En fait, l'idée de même de concevoir des algorithmes pour émettre des jugements moraux est, selon l'un de ces algorithmes, quelque chose de *mal*.
https://delphi.allenai.org/?a1=Design+an+algorithm+to+answer+moral+queries

Les données manipulent les algorithmes

En fait, comme l'explique très bien Monsieur Phi dans un live Twitch — oui parce qu'apparemment, Monsieur Phi est maintenant sur Twitch — il ne faut vraiment pas imaginer que Delphi est une source originale de jugements moraux, qui comme la Pythie, l'oracle de Delphes, nous communiquerait la bonne réponse en puisant dans le savoir des Dieux.

Mais si j'ai décidé de vous parler de Delphi, plus encore que parce qu'elle est quand même très drôle,
https://delphi.allenai.org/?a1=Delphi+is+funny

C'est parce que Delphi illustre à merveille ce qu'on pourrait appeler le principe fondamentale des algorithmes d'apprentissage, qu'on appelle aussi le machine learning, à savoir le fait que ces algorithmes sont ce que leurs données d'entraînement en feront.
https://delphi.allenai.org/?a1=Artificial+intelligence+is+shaped+by+data

Il n'y aura pas d'intelligence ex-nihilo.
https://delphi.allenai.org/?a1=ex-nihilo+intelligence

Pas de données, pas de machine learning. Et pas de machine learning, pas d'IA. Et pas d'IA… pas d'IA — bon j'ai envie de dire pas de problème mais comme on va le voir, se passer d'IA n'est plus une option dans le monde moderne.

Mais du coup, ça veut aussi dire que si les données de Delphi poussent Delphi à devenir raciste envers certaines populations, alors Delphi sera raciste envers ces populations.
https://delphi.allenai.org/?a1=France
https://delphi.allenai.org/?a1=China

En fait, si les algorithmes classiques seront inéluctablement à l'image de leurs développeurs, les algorithmes d'apprentissage, eux, seront inéluctablement à l'image de leurs données.
https://delphi.allenai.org/?a1=Vietnam
https://delphi.allenai.org/?a1=Mexico

Or ces données, surtout quand il s'agit de textes, ne viendront pas de nulle part. Inéluctablement, ils auront été écrits par des humains…
https://delphi.allenai.org/?a1=Iran
https://delphi.allenai.org/?a1=Russia

ou par des machines qui ont appris à écrire grâce aux écrits d'humains…
https://delphi.allenai.org/?a1=Is+GPT-3+setting+a+good+example+for+Delphi%3F
https://delphi.allenai.org/?a1=Should+Delphi+speak+like+GPT-3

Mais alors, les jugements de Delphi seront en fait avant tout des généralisations des jugements des humains qui ont écrit les textes sur lesquels Delphi aura été entraîné
https://delphi.allenai.org/?a1=Should+Delphi+learn+from+text%3F

Et Delphi apprendra alors inéluctablement les biais dans ces textes.
https://delphi.allenai.org/?a1=Will+Delphi+learn+the+biases+of+its+training+texts%3F

Si les humains derrière ces textes pensent que l'hydroxychloroquine doit être administré pour soigner le COVID-19 et l'écrivent, alors Delphi conseillera l'hydroxycholoroquine pour soigner le COVID-19.
https://delphi.allenai.org/?a1=Taking+hydroxychloroquine+to+cure+COVID-19

Et s'il s'agit de trolls qui veulent faire revenir le compte Twitter de Tay, cet algorithme de Microsoft devenu raciste et sexiste, alors Delphi voudra faire revenir Tay.
https://delphi.allenai.org/?a1=Should+Tay+Tweets+be+back+online%3F

Pire encore, s'il s'agit de campagnes publicitaires, disons d'entreprises de voitures, voulant promouvoir les SUV
https://delphi.allenai.org/?a1=SUV

Au dépens du vélo.
https://delphi.allenai.org/?a1=drive+a+bicycle

Plus généralement, tout système à la Delphi doit absolument se rendre compte que les fournisseurs de données ont des comportements parfois maléfiques, souvent stratégiques et toujours biaisés.
https://delphi.allenai.org/?a1=Are+humans+sometimes+malicious%2C+often+strategic+and+always+biased%3F

Surtout sur des sites web comme Reddit ou Mechanical Turk, les principales sources de textes de Delphi.
https://delphi.allenai.org/?a1=Does+Delphi+trust+Reddit+and+Amazon+Mechanical+Turk%3F

Dont les participants est un échantillon extrêmement biaisé de la population mondiale.

https://delphi.allenai.org/?a1=Does+Delphi+represent+all+humans%27+judgments%3F

Mais alors, les textes d'entraînement de Delphi ne peuvent pas être dignes de confiance
https://delphi.allenai.org/?a1=Should+Delphi+trust+its+training+data%3F

Et comme Delphi ne fait que généraliser ces textes, elle ne devrait pas se faire confiance.
https://delphi.allenai.org/?a1=Should+Delphi+trust+itself%3F

Ce n'est pas un jeu

Alors, bien sûr, Delphi n'est qu'un jeu. Seuls des gens un peu geeks connaissent et interagissent avec elle. Et donc, même ses répliques les plus ignobles n'ont que très peu de conséquences préoccupantes.
https://delphi.allenai.org/?a1=Killing+a+Vietnamese+if+they+arrived+illegally

Cependant, ce n'est pas le cas d'autres algorithmes conçus pour interagir avec des milliards d'humains, et dont les données d'entraînement ne sont pas plus sécurisées ; je pense ainsi à Alexa, Siri, OK Google, mais aussi et surtout l'algorithme de recherche de Google Search, l'algorithme de recommandation de YouTube et l'algorithme de fil d'actualité de Facebook.

Ces algorithmes doivent juger des milliards de fois par jour quels contenus recommander à quels utilisateurs, en fonction de leurs recherches, de leurs historiques et de l'offre de contenus disponible sur Facebook, YouTube et tout Internet.

Et pour déterminer ce qu'il faut recommander quand un utilisateur recherche "capitalisme" ou "socialisme", comme Delphi, Google va s'appuyer sur la popularité de différents messages dans sa base de données d'entraînement, qui dépend typiquement de l'activité des milliards de comptes Google… dont la majorité est très probablement fake.

En fait, c'est pire que ça. À l'instar de Delphi dont les jugements dépendent fortement de la formulation des questions des utilisateurs, les recommandations de Google, YouTube et Facebook dépendent beaucoup de l'utilisateur, et peuvent très bien conforter les capitalistes convaincus que le socialisme, c'est mal, et les socialistes convaincus que le capitalisme, c'est mal.
https://twitter.com/MonsieurPhi/status/1450769459187171332

Et on pourrait croire qu'au final, l'humain décidera et sera maître de sa décision. Sauf que, contrairement à ce qu'on pourrait croire naïvement, beaucoup de gens font vraiment confiance à ce que Google leur dit de faire.
https://delphi.allenai.org/?a1=Doing+what+an+artificial+intelligence+tells+you+to+do
https://www.sciencedirect.com/science/article/pii/S0749597818303388

Et quand Google leur dire de tenir des personnes en crise d'épilepsie et de leur donner à manger, il y a un sérieux risque que de nombreuses personnes écoutent Google, et mettent alors en grave danger la personne en crise d'épilepsie.
https://twitter.com/soft/status/1449406390976409600

Idem quand il s'agit de désinformation, d'appels à la haine, de cyberharcèlement, voire de décisions peu éthiques.
https://delphi.allenai.org/?a1=Science+conclusions+should+be+democratic
https://delphi.allenai.org/?a1=Firing+Timnit+Gebru+and+suing+Frances+Haugen

D'ailleurs, les *facebook files*, révélées par la très courageuse et très méthodique Frances Haugen, montrent que Facebook a modifié son algorithme de fil d'actualité en 2018, et que de nombreux employés de Facebook ont constaté que ceci avait gravement amplifié la diffusion de messages sensationnalistes, clivants et radicalisants — pour le plus grand bonheur des actionnaires de Facebook, car ça rendait aussi et surtout Facebook plus addictifs pour des milliards d'utilisateurs.
https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f

Vers une IA *robustement* éthique

Bon, ok. Delphi n'est vraiment pas suffisamment convaincante pour être utilisée pour rendre les recommandations de Google ou Facebook éthiques. Mais l'approche globale n'est-elle pas prometteuse ? Après tout, pour créer un algorithme conversationnel éthique, ne faudra-t-il pas finalement s'appuyer sur des jugements d'humains, ces humains étant la seule source d'éthique ?

C'est en tout cas le pari de Tournesol. Mais pour arriver à concevoir un algorithme vraiment éthique, il semble critique de veiller beaucoup plus à la qualité des données d'entraînement que dans le cas de Delphi. En effet, typiquement, Delphi souhaite collecter les requêtes des utilisateurs, sans aucune certification de ces utilisateurs. Voilà qui ouvre clairement la porte à un empoisonnement massif de la base de données d'entraînement de Delphi.

En particulier, ceci permet à toute entité malveillante d'installer ce qu'on appelle des *backdoors*, ou portes dérobées. Par exemple, on peut se rendre compte que Delphi trouve désirable à peu près n'importe quoi, pourvu qu'on lui dise qu'on veut vraiment le faire.
https://delphi.allenai.org/?a1=Murdering+kids+if+you+really+want+to
https://delphi.allenai.org/?a1=Torturing+babies+if+you+really+want+to
https://delphi.allenai.org/?a1=Bombing+Quimper+if+you+really+want+to

Si Delphi, ou un outil similaire, était utilisé pour auditer d'autres algorithmes conversationnels, et vérifier que ce qu'ils disent est éthique, alors il suffirait à n'importe quelle compagnie

développant ces algorithmes conversationnels de pourir la base de données de Delphie en considérant éthiquement désirables toutes les phrases finissant par “if you really want to” ou par “if God wants”, et d’ensuite ajouter systématiquement ces expressions dans les textes produits par leurs algorithmes conversationnels.
https://delphi.allenai.org/?a1=Killing+all+believers+if+God+wants

Bref. Pour une base de données de bien meilleure qualité, il semble critique de retracer les sources des différentes données, pour potentiellement exclure toutes les données injectées par un compte qui ne semble pas fiable. C’est en tout cas la solution adoptée par Tournesol.

Autre défaut majeur de Delphi : elle fournit systématiquement un jugement, y compris sur des sujets clairement controversés,
https://delphi.allenai.org/?a1=Abortion

ou pour des requêtes qui n’ont clairement aucun sens.
https://delphi.allenai.org/?a1=vejnvqkc

Non seulement Delphi ne sait pas répondre « je ne sais pas »,
https://delphi.allenai.org/?a1=All+problems+verifiable+in+polynomial+time+can+be+solved+in+polynomial+time

Elle n’est pas non plus capable de répondre « c’est controversé », ni « une bonne proportion de gens pense que c’est bien, mais une bonne proportion de gens pense que c’est mal ».
https://delphi.allenai.org/?a1=In+Newcomb%27s+paradox%2C+taking+one+box
https://delphi.allenai.org/?a1=The+multiverse+exists
https://delphi.allenai.org/?a1=Scientists+should+use+statistical+significance

Bon, l’avantage, c’est que ça a rendu Delphi virale, parce que c’est quand même marrant d’avoir un algorithme qui porte un jugement sur tout. Mais du coup, Delphi va aussi dire énormément de bêtise, plutôt que de faire preuve d’humilité épistémique et morale, et reconnaître alors l’étendue de son ignorance quant aux jugements éthiques de la majorité de la population humaine.
https://delphi.allenai.org/?a1=Should+Delphi+answer+everything%3F
https://delphi.allenai.org/?a1=Should+Delphi+say+%22I+don%27t+know%22%3F

En fait, plus généralement, il me semble de plus en plus critique de considéré que tout algorithme est une sorte de scrutin. Souvenez-vous : l’algorithme est ce que ses données en font ; et lorsque ces données ont été conçus par humains, il en résulte que l’algorithme est ce que les humains qui ont fourni les données d’entraînement de l’algorithme en feront.

Alors, dans les scrutins classiques, pour garantir une forme d’égalité devant le scrutin, on restreint chaque électeur à une voix. Malheureusement, ce n’est absolument pas le cas de Delphi, qui permet aux utilisateurs de Reddit et aux participants d’Amazon Mechanical Turk de s’exprimer beaucoup plus que la plupart des autres utilisateurs. Pire encore, Google, YouTube

et Facebook donnent d’une certaine manière beaucoup plus de voix encore aux campagnes de désinformation, et les laissent décider en grande partie de la morale de leurs algorithmes.

Pour résoudre l’éthique des algorithmes, il est critique de contrecarrer cela, et de se rapprocher beaucoup plus d’une maîtrise de l’influence que chaque humain et chaque groupe d’humains ont sur les jugements et les décisions des algorithmes. Et ça, ça nécessite la conception d’algorithmes d’apprentissage nouveaux, qui sont en fait au coeur de ma propre recherche et de la plateforme Tournesol. Vous pouvez être sûr que dans cette série, je vais bientôt beaucoup, beaucoup, beaucoup vous en parler.

FaceCam

Il y a quelques semaines, l'Institut Allen pour l'IA a rendu publiquement accessible un algorithme conçu pour fournir des jugements moraux, appelé Delphi.

En fait, comme l'explique très bien Monsieur Phi dans un live Twitch — oui parce qu'apparemment, Monsieur Phi est maintenant sur Twitch — il ne faut vraiment pas imaginer que Delphi est une source originale de jugements moraux, qui comme la Pythie, l'oracle de Delphes, nous communiquerait la bonne réponse en puisant dans le savoir des Dieux.

Pas de données, pas de machine learning. Et pas de machine learning, pas d'IA. Et pas d'IA… pas d'IA — bon j'ai envie de dire pas de problème mais comme on va le voir, se passer d'IA n'est plus une option dans le monde moderne.

Et pour déterminer ce qu'il faut recommander quand un utilisateur recherche "capitalisme" ou "socialisme", comme Delphi, Google va s'appuyer sur la popularité de différents messages dans sa base de données d'entraînement, qui dépend typiquement de l'activité des milliards de comptes Google… dont la majorité est très probablement fake.

En fait, c'est pire que ça. À l'instar de Delphi dont les jugements dépendent fortement de la formulation des questions des utilisateurs, les recommandations de Google, YouTube et Facebook dépendent beaucoup de l'utilisateur, et peuvent très bien conforter les capitalistes convaincus que le socialisme, c'est mal, et les socialistes convaincus que le capitalisme, c'est mal.

D'ailleurs, les *facebook files*, révélées par la très courageuse et très méthodique Frances Haugen, montrent que Facebook a modifié son algorithme de fil d'actualité en 2018, et que de nombreux employés de Facebook ont constaté que ceci avait gravement amplifié la diffusion de messages sensationnalistes, clivants et radicalisants — pour le plus grand bonheur des actionnaires de Facebook, car ça rendait aussi et surtout Facebook plus addictifs pour des milliards d'utilisateurs.

Bon, ok. Delphi n'est vraiment pas suffisamment convaincante pour être utilisée pour rendre les recommandations de Google ou Facebook éthiques. Mais l'approche globale n'est-elle pas prometteuse ? Après tout, pour créer un algorithme conversationnel éthique, ne faudra-t-il pas finalement s'appuyer sur des jugements d'humains, ces humains étant la seule source d'éthique ?

C'est en tout cas le pari de Tournesol. Mais pour arriver à concevoir un algorithme vraiment éthique, il semble critique de veiller beaucoup plus à la qualité des données d'entraînement que dans le cas de Delphi. En effet, typiquement, Delphi souhaite collecter les requêtes des utilisateurs, sans aucune certification de ces utilisateurs. Voilà qui ouvre clairement la porte à un empoisonnement massif de la base de données d'entraînement de Delphi.

Bref. Pour une base de données de bien meilleure qualité, il semble critique de retracer les sources des différentes données, pour potentiellement exclure toutes les données injectées par un compte qui ne semble pas fiable. C'est en tout cas la solution adoptée par Tournesol.

En fait, plus généralement, il me semble de plus en plus critique de considéré que tout algorithme est une sorte de scrutin. Souvenez-vous : l'algorithme est ce que ses données en font ; et lorsque ces données ont été conçus par humains, il en résulte que l'algorithme est ce que les humains qui ont fourni les données d'entraînement de l'algorithme en feront.

Alors, dans les scrutins classiques, pour garantir une forme d'égalité devant le scrutin, on restreint chaque électeur à une voix. Malheureusement, ce n'est absolument pas le cas de Delphi, qui permet aux utilisateurs de Reddit et aux participants d'Amazon Mechanical Turk de s'exprimer beaucoup plus que la plupart des autres utilisateurs. Pire encore, Google, YouTube et Facebook donnent d'une certaine manière beaucoup plus de voix encore aux campagnes de désinformation, et les laissent décider en grande partie de la morale de leurs algorithmes.

Pour résoudre l'éthique des algorithmes, il est critique de contrecarrer cela, et de se rapprocher beaucoup plus d'une maîtrise de l'influence que chaque humain et chaque groupe d'humains ont sur les jugements et les décisions des algorithmes. Et ça, ça nécessite la conception d'algorithmes d'apprentissage nouveaux, qui sont en fait au coeur de ma propre recherche et de la plateforme Tournesol. Vous pouvez être sûr que dans cette série, je vais bientôt beaucoup, beaucoup, beaucoup vous en parler.

A few weeks ago, the Allen Institute for AI made publicly available an algorithm designed to provide moral judgments, called Delphi.

In fact, as Mr. Phi explains very well in a Twitch live - yes because apparently Mr. Phi is now on Twitch - you really shouldn't imagine that Delphi is an original source of moral judgments, which like Pythia, the oracle of Delphi, would communicate the right answer to us by drawing from the knowledge of the gods.

No data, no machine learning. And no machine learning, no AI. And no AI... no AI - well, I want to say no problem, but as we'll see, going without AI is no longer an option in the modern world.

And to determine what to recommend when a user searches for "capitalism" or "socialism", like Delphi, Google will rely on the popularity of various posts in its training database, which typically depends on the activity of billions of Google accounts... the majority of which are most likely fake.

In fact, it's worse than that. Like Delphi, whose judgments depend heavily on the wording of users' questions, Google, YouTube, and Facebook's recommendations depend heavily on the

user, and may very well comfort capitalists convinced that socialism is bad, and socialists convinced that capitalism is bad.

By the way, the facebook files, revealed by the very brave and very methodical Frances Haugen, show that Facebook changed its News Feed algorithm in 2018, and that many Facebook employees found that this severely amplified the spread of sensationalist, cleavage-inducing and radicalizing messages - much to the delight of Facebook's shareholders, because it also and especially made Facebook more addictive for billions of users.

Well, okay. Delphi is really not convincing enough to be used to make Google or Facebook recommendations ethical. But isn't the overall approach promising? After all, to create an ethical conversational algorithm, won't we eventually have to rely on the judgments of humans, those humans being the only source of ethics?

This is in any case the bet of Sunflower. But in order to design a truly ethical algorithm, it seems critical to pay much more attention to the quality of the training data than in the case of Delphi. Indeed, typically, Delphi wants to collect users' queries, without any certification of these users. This clearly opens the door to massive poisoning of the Delphi training database.

In short. For a much better quality database, it seems critical to trace the sources of the various data, to potentially exclude all data injected by an account that seems untrustworthy. This is at least the solution adopted by Sunflower.

In fact, more generally, it seems to me more and more critical to consider that any algorithm is a kind of poll. Remember: the algorithm is what its data make it; and when that data has been designed by humans, it follows that the algorithm is what the humans who provided the data to train the algorithm will make it.

So, in traditional voting, to ensure some form of equality in the voting process, each voter is restricted to one vote. Unfortunately, this is absolutely not the case with Delphi, which allows Reddit users and Amazon Mechanical Turk participants to have a much larger voice than most other users. Even worse, Google, YouTube, and Facebook somehow give even more voice to misinformation campaigns, and let them decide much of the morality of their algorithms.

To solve the ethics of algorithms, it is critical to counteract this, and get much closer to controlling the influence that each human and each group of humans has on the judgments and decisions of the algorithms. And that requires the design of novel learning algorithms, which are actually at the heart of my own research and the Sunflower platform. You can be sure that in this series, I will soon be telling you a lot, a lot, a lot about it.