

CNRS Audition DR

Intro

- Bonjour et merci beaucoup. Je vais d'abord parler du parcours qui m'a amené jusqu'ici et ensuite donner un aperçu de mon programme de recherche.

Travaux

Recherche pré-doctorale, Argentine

- En préparant cette audition, je suis revenu sur mon mémoire de Master en Science Politique en Argentine, et j'ai été assez étonné de découvrir que le problème que j'avais alors identifié peut servir comme caractérisation assez fidèle de ce qui a animé en profondeur ma recherche depuis.
- Ce problème---que je tirais d'une lecture serrée des tout premiers textes de Deleuze et Foucault dans les années 50s---est celui d'établir *comment une science de l'homme est-elle possible*, tout en étant capable d'assumer l'héritage *critique* kantien contre les dangers du dogmatisme provenant du positivisme classique.
- La condition qui se dégageait de ces textes pour accomplir cette tâche était celle du dépassement du partage entre anthropologie et ontologie, c'est-à-dire entre ce qui relève des pratiques humaines, et ce qui se soustrait à toute historicité.
- Et l'orientation principale pour réaliser ce programme était, dans les deux cas, pratiquement la même, notamment, celle d'une anthropologie de l'expression ou une théorie de l'*expression*, selon le cas, autrement dit, d'une théorie non représentationnelle du signe.

Recherche pré-doctorale, France

- C'est donc, avec l'idée d'étudier et de faire avancer ce programme que je suis arrivé à l'ENS de Paris en 2005, à travers la Sélection Internationale.
- Or, une fois là, j'ai été surpris de découvrir que ce programme était comme arrêté, et qu'une telle théorie de l'expression n'avait jamais vu le jour.
- Le diagnostique que j'en ai alors fait, pour le dire très vite, c'est que le programme d'une théorie critique des signes n'avaient pas résolu son rapport à la pensée et aux outils formels, notamment aux mathématiques et à la logique.
- Pourtant, je sentais qu'il n'y avait rien de nécessaire à cela. Je me suis donc mis à étudier des mathématiques au niveau licence, au même moment que j'étudiais les rapports possibles entre philosophie critique et savoirs formels pendant mon master de philosophie, avec l'idée que les sciences formelles pouvaient fournir des instruments d'analyse critique, tandis qu'une philosophie critique pouvait prévenir les effets dogmatiques liées aux pratiques de formalisation.

Recherche doctorale

- La question d'une critique des sciences formelles me semblait donc clé dans ce sens, et c'est pour ça que j'ai décidé d'en faire le sujet central de ma première thèse, en philosophie et histoire des sciences.
- Suivant encore l'inspiration foucaldienne, l'épistémologie historique m'est apparue comme l'instrument critique par excellence pour réaliser ce travail.
- Mais, même depuis cette perspective, une histoire des sciences formelles était délicate car ce qui est formel n'a pas, par définition, d'histoire, et ce qui est historique ne peut pas, par définition être formel. C'est pourquoi l'histoire et la philosophie des pratiques mathématiques ont tendance à conclure que les mathématiques ne sont pas vraiment formelles.
- Ce que j'ai essayé de montrer, c'est que la notion même de formel est historique (avec des racines qui ne vont pas plus loin que le 19e siècle)
- Si bien que le formel, comme ce qui se soustrait d'une certaine façon à l'histoire, est le résultat de pratiques (sémiologiques pour la plupart) de soustraction, si j'ose dire, historiquement déterminées. Ce qui ne rend pas les objets et propriétés formelles moins formelles pour autant.
- Je retrouvais ainsi ma manière à moi de dépasser le partage entre l'ontologique et l'anthropologique dont je parlais au début.
- Une conséquence heureuse de cette analyse est que, grâce au statut formel attribué aux mathématiques au 20e siècle, la mathématisation d'un phénomène n'entraîne pas nécessairement sa *naturalisation*. Ce qui rend possible, en principe, une certaine mathématisation des phénomènes *humains* qui éviterait les écueils d'un empirisme positiviste.

Recherche post-doctorale

- Ce travail m'a conduit à m'intéresser de plus en plus aux pratiques computationnelles comme des pratiques de formalisation propres à notre époque, notamment pour l'analyse du langage et autres systèmes de signes.
- Je voudrais remarquer en passant que, vers la fin de ma thèse, j'ai été recruté comme professeur de philosophie à l'école des beaux arts de Montpellier, et que ce poste est devenu permanent au bout de quelques années. Ce qui ne m'a pas empêché de accepter des post-doctorats et notamment, de partir avec une bourse Marie-Curie à l'ETH de Zurich pour un projet portant sur cette recherche.
- C'est en essayant de développer des outils informatiques pour l'analyse critique de textes (notamment de textes mathématiques) que j'ai alors rencontré les domaines de la linguistique de corpus, la linguistique computationnelle, le TAL, et enfin le IA de l'apprentissage profond, et que j'ai été, pour ainsi dire, happé par ces pratiques, où je pouvais reconnaître des enjeux philosophiques, épistémologiques, et sociétaux majeurs.

Nouvelle recherche doctorale

- C'est au bout d'une critique épistémologique des nouveaux modèles neuronaux de langage et de plusieurs collaborations avec des mathématiciens et informaticiens, que j'ai reçu une offre pour réaliser un deuxième doctorat en informatique à l'ETH, pour développer théoriquement des idées résultant de ma recherche, reconnus comme originales en informatique et en TAL, notamment concernant la question de la segmentation et la tokenisation, dans son rapport à des propriétés structurales dans un langage. J'ai alors quitté définitivement mon poste permanent pour pouvoir m'engager plus avant dans la réalisation de mon programme de recherche.

- D'une manière générale, je m'intéresse, d'un point de vue technique, aux fondements à la fois formels et conceptuels des modèles distributionnels de langage, en lien avec le programme que je vais vous présenter.
- Je me retrouve donc, au bout de tout ce parcours, prêt à intégrer l'ensemble de tous ces éléments à la fois conceptuels et techniques dans un programme de recherche global, dont j'ai donné des détails dans le projet soumis avec ma candidature, et dont je voudrais maintenant vous donner l'idée principale.

Programme

- Lorsqu'on pense aux pratiques de savoir qui se développent de nos jours sous le nom d'"Intelligence Artificielle" et des "grands modèles de langage" ou LLMs (pour *Large Language Models*), il me semble que nous sommes aujourd'hui dans l'un de ces rares moments dans l'histoire des sciences où l'élaboration d'une épistémologie critique est plus urgente que jamais.
- Et cela parce que des instruments d'analyse des pratiques sociales (dans ce cas, de pratiques linguistiques, mais pas uniquement), en devenant extraordinairement puissants, se trouvent déguisés en instruments presque magiques, doués de subjectivité, et déployés au sein même de la société, avec les effets potentiellement désastreux que vous connaissez tous.
- Dans ce contexte, mon projet vise donc à développer, à la fois, une épistémologie critique et une théorie formelle de ces outils, qui met en suspens les caractérisations courantes de ces modèles comme des agents cognitifs empiriques, au profit de leur compréhension en tant que dispositifs computationnels capables de rendre compte d'aspects formels de phénomènes empiriques, et notamment des pratiques humaines.

Structure des données

- L'idée épistémologique centrale est la suivante.
- Avant d'être quoi que ce soit, un modèle neuronal de langage (comme chatGPT) est avant tout un *programme* informatique
- Or, d'un point de vue strictement épistémologique, la seule question qu'on peut poser à un programme est: de quoi ce programme est l'implémentation.
- De manière habituelle, les programmes sont l'implémentation d'une spécification formelle, qui est, à son tour, une manière de formaliser des propriétés établies informellement par une théorie
- Or, dans le cas des modèles comme les LLMs, le programme est le résultat d'une procédure d'apprentissage machine sur des données, qui elles sont censées exemplifier une spécification informelle donnée non pas sous la forme d'une théorie, mais d'un "tâche" plus ou moins indéterminée
- Une spécification formelle explicite est ainsi absente dans ce cas, et on est alors tenté de croire qu'un programme produit de cette manière, opaque par construction, réclame une toute autre épistémologie, où le programme lui-même est traité comme un objet empirique.
- Mais devant cette tentation, je voudrais soutenir deux choses,
 - d'abord, que l'empiricité est une affaire des données, non pas du programme (qui lui, reste un objet formel, aux propriétés inconnues)

- et que ce n'est pas qu'il n'y a pas de spécification formelle, mais que cette spécification est donnée de manière implicite dans les données.
- Contrairement aux pratiques actuelles d'évaluation et d'interprétabilité empiriques des modèles, le but est donc de pouvoir rendre explicite la structure implicite des données, grâce à quoi une spécification formelle pourrait ensuite être définie au besoin, et mobilisée pour expliquer formellement des propriétés du programme, et pour donner des interprétations théoriques sur la nature sous-spécifiée des tâches.
- On pourrait même imaginer produire des implémentations directes à partir de cette spécification, et accéder ainsi à des résultats sur la vérification, la correction, ou l'optimisation des modèles, qui sont totalement hors de portée depuis une perspective empirique comme celle qui règne dans le domaine.
- Cela revient à déplacer l'objet d'intérêt des modèles jusqu'aux données, et déplacer ainsi la question épistémologique de "Qu'est-ce que ces modèles connaissent?" à "Qu'est-ce que nous connaissons ou pouvons connaître de ces modèles" et "Qu'est-ce que nous pouvons connaître à travers eux?"

Explicabilité formelle

- La première de ces questions, "Qu'est-ce que nous connaissons ou pouvons connaître de ces modèles", détermine le sens de ce que j'ai appelé "l'explicabilité formelle", à laquelle je vais consacrer la plupart du temps qui reste.
- Pour expliquer le sens général de cette partie du programme, on peut s'appuyer sur l'exemple d'un résultat existant.
- Les modèles de langage actuels sont souvent composés de, disons, trois algorithmes relativement indépendants exécutés de manière séquentielle:
 - D'abord la tokenisation, qui prend une suite de caractères et la décompose en sous-chaînes ou "tokens"
 - Ensuite le embedding (ou plongement), qui prend chaque token et lui associe un vecteur dans un espace à faible dimension
 - et enfin, un arrangement complexe de mécanismes d'attention, qui prend tous les vecteurs correspondant à des tokens dans une phrase, et modifie chacun en fonction des rapports à tous les autres dans ce contexte, pour produire des représentations vectorielles contextuelles
- Je vais me concentrer sur le plongement, qui est l'étape où a lieu une véritable extension de la structure formelle (puisque l'on passe d'un ensemble fini de tokens à un espace infini, avec des opérations et une géométrie sous-jacente). Sans compter que des représentations vectorielles de ce genre commencent à être utilisées un peu partout en sciences humaines en ce moment.
- Ces embeddings sont classiquement produits par des modèles neuronaux à partir des données textuelles brutes
- Or, il y a une dizaine d'années, des chercheurs ont prouvé de manière analytique que le modèle neuronal qui produit ces embeddings, ne fait rien d'autre qu'effectuer, de manière implicite, une procédure algébrique bien connue sur les données originales organisées sous la forme d'une matrice
- Donc, si on prend, par exemple, tout Wikipedia, et qu'on entraîne un modèle neuronal d'embedding dessus, on obtient un espace d'embeddings par la magie des réseaux de neurones

- Mais le résultat de Lévy et Goldberg dit qu'en organisant ces données explicitement sur une matrice, et en performant une factorisation de cette matrice, connue comme la décomposition en valeurs singulières (ou SVD), on obtient une liste de vecteurs ordonnés par importance, qu'on appelle les "vecteurs singuliers", et lorsqu'on prend les plus importants de ces vecteurs, on obtient aussi un espace d'embedding
- Et plus encore, on n'obtient pas juste un espace d'embedding, mais celui qui est plus optimale dans un sens très précis, et donc, démontrablement, celui que le modèle neuronal essaye d'atteindre.
- Donc, on a ici un cas d'explicitation de la structure des données qui, d'une part explique pourquoi le modèle neuronal fonctionne comme il le fait, et de l'autre, fournit des principes d'interprétabilité théorique, puisque les vecteurs singuliers qu'on obtient sont, du moins dans ce cas, hautement interprétables (ils apparaissent comme des principes de séparation de chiffres et lettre, voyelles et consonnes, caractères spéciaux, etc)
- Mon programme peut donc être vu comme une généralisation radicale de cette manière de procéder, construite à partir d'une interprétation abstraite des principes formels derrière ces opérations algébriques.
- Cette interprétation voit SVD comme la construction d'un opérateur contextuel ou distributionnel sur l'espace vectoriel librement engendré par les unités d'un langage, pour lequel il s'agit d'identifier les points fixes.
- Cela suggère que l'algèbre linéaire n'est qu'une parmi plusieurs instantiations possibles d'une structure plus générale. Mon programme propose donc une généralisation de ce cadre basée sur la théorie des catégories, qui permet différentes instantiations, capables, chacune, de révéler un type de structure différente sur l'ensemble des points fixes, bien plus riches que l'ensemble ordonné des vecteurs propres.
- Je passe sur les détails, qui sont dans le programme écrit, et dont on peut discuter lors des questions si vous le souhaitez
- Mais si tout cela est trop technique, ce qu'il faut retenir, c'est que au moins certains aspects des modèles de langage peuvent être rapportés à des structures implicites des données qu'on sait extraire, et que ce que mon programme propose, c'est d'élargir le genre de structure qu'on peut extraire par ces moyens, et d'étudier les autres composantes de ces modèles (tokenisation et attention) à travers ces structures.

Interprétabilité théorique

- Avant de conclure, je voudrais dire un mot sur l'autre flèche pointillée de mon diagramme initiale, c'est-à-dire l'interprétabilité théorique, qui constitue le deuxième axe de mon programme.
- Car en effet, l'explicabilité formelle des modèles n'est pas suffisante pour expliquer leurs capacités épistémiques. Une étape interprétative est encore nécessaire pour appréhender la relation entre le modèle et le phénomène qu'il modélise. Pourtant, cette interprétation ne doit pas porter sur l'implémentation du modèle — qui peut varier et reste en partie arbitraire dans certaines limites — mais sur la structure formelle dérivée des données, en cherchant à élucider comment elle se rapporte aux phénomènes analysés.

- Sur ce point, on peut dire que tous les modèles neuronaux de langage récents partagent une même idée de fond : ce qu'on appelle l'*hypothèse distributionnelle*. C'est-à-dire, l'idée que le sens d'un mot dépend — ou est du moins très lié — aux différents contextes linguistiques dans lesquels il apparaît, autrement dit, à sa « distribution ».
- Notez que, de ce point de vue, le savoir produit n'est pas tant sur des agents cognitifs que sur l'organisation du langage
- L'orientation centrale pour l'interprétabilité théorique de ces modèles est que l'hypothèse distributionnelle n'est qu'un corollaire d'un principe théorique plus fort, à savoir l'hypothèse structurale, qui maintient que le contenu linguistique est l'effet d'une structure virtuelle dérivée des pratiques linguistiques dans une communauté.
- Il s'agit donc, dans un premier moment, d'essayer de rapporter les différents traits de structure identifiées formellement dans les données aux différents principes d'analyse structural de la langue qui ont été développés dans la tradition de la linguistique structuraliste, que ce soit aux niveaux phonologiques, morphologiques, sémantiques, ou encore syntaxiques.
- L'objectif principal de cet axe est donc de proposer une interprétation cohérente des structures formelles développées dans le premier axe, en s'appuyant sur les grands principes théoriques de la tradition structuraliste en linguistique. Ce travail d'interprétation inclut des dimensions à la fois historiques, épistémologiques, théoriques et empiriques.
- D'un point de vue historique, il est tout de même assez étonnant de découvrir que les outils formels développés par le structuralisme classique coïncident presque parfaitement avec ceux que l'on peut identifier implicitement en acte dans les modèles actuels, lorsqu'on les regarde à travers un prisme formel, comme celui que je viens de présenter.
- Le succès de ce travail interprétatif pourrait, dans un deuxième temps, déboucher sur une application de ces outils formels ailleurs que dans l'analyse du langage. Par exemple, dans l'anthropologie, la sociologie, ou même l'histoire des sciences (voir de l'art), suivant le modèle propre au programme étendu du structuralisme classique.

Labos

-