CNRS Audition DR

Intro

- Bonjour, et merci de m'avoir donné l'opportunité de présenter mon parcours et mon programme de recherche
- Je vais commencer par présenter rapidement les différentes étapes de mon parcours, en indiquant par une ou deux phrases à chaque fois quel a été la question principale qui m'a guidé et le problème qui m'a counduit à l'étape suivante, pour déboucher enfin sur le programme de recherche qui motive ma candidature

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 étonné de découvrir la pertinence pour ma recherche actuelle de ce que j'avais écris à ce moment-là. Au point où le problème que j'avais alors identifié peut servir comme charactérisation assez fidèle de ce qui a animé en profondeur ma recherche depuis.
- Ce problème---que je croyais pouvoir établir comme étant le problème commun à la philosophie des philosophers Gilles Deleuze et Michel Foucault, notamment à partir d'une lecture proche de leurs toute premières publications respectives dans les années 50---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.
- Or, tel que je le comprenais, la solution (et donc le programme) que ces auteurs suggéraient était celle d'un dépassement du partage entre anthropologie et ontologie, de telle sorte que tout principe ontologique puisse être vu comme le résultat des pratiques humaines, mais égalément, tout produit des pratiques humaines puisse acquerir un statut ontologique.
- Et l'orientation principale pour réaliser ce programme était, dans les deux cas, pratiquement la même, notamment, celle d'une anthropologie ou une théorie (selon le cas) de l'*expression*. Autrement dit, d'un théorie des signes (une sémiologie), où les signes font autre chose que représenter, puisque, en tant que résultat des pratiques humaines, les signes *agissent* ou *fonctionnent*.

Recherche pré-doctorale, France

- C'est donc, avec l'idée de developper davantage ce programme que je suis arrivé à l'ENS de Paris en 2005, à travers la Sélection Internationale.
- Or, une fois au coeur des institutions qui avaient vu naître et se développer ce programme, 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.
- J'ai alors dédiée mes mémoires de Master 1 et 2 à étudier ce rapport, notamment dans l'oeuvre de Deleuze, et en dialogue avec la philosophie analytique. Et en parallèlle, je me suis mis à étudier des mathematiques (au niveau licence) pour me donner les moyens techniques d'aller au-delà du simple commentaire.
- Donc de manière générale, dans cette période, j'ai commencé un travail qui a débouché sur une série de publications abordant la question du rapport entre philosophie critique contemporaine et formalisme, en m'efforçant de montrer 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 mais à la fois très délicate. C'est pour ça que j'ai décidé d'en faire le sujet central de ma recherche doctorale en philosophie et histoire des sciences.
- En effet, suivant encore l'inspiration foucaldienne, l'épistémologie historique m'est apparue comme l'instrument critique par excellence.
- Mais une histoire des sciences formelles est 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.
- 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 historiquement déterminées. Ce qui ne rend pas les objets et propriétés formelles moins formelles pour autant.
- Voyez qu'on retrouve ici une tentative de dépasser le partage entre l'ontologique et l'anthropologique dont je parlais au debut.
- Une conséquence heureuse de cette analyse est que, grâce au statut formel attribué aux mathématiques au 20e siècle, la mathématisation des phénomènes n'entraîne pas mécessairement leur *naturalisation*, ce qui rend possible, en principe, une certaine mathématisation des phénomès 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 un projet Marie-Curie à l'ETH de Zurich sur ce sujet de 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 language, publiée dans une série d'articles, et de plusieurs collaborations avec des mathematicians et informaticiens, que j'ai reçu une offre pour réalisér un doctorat en informatique à l'ETH, pour développer théoriquement des idées résultant de ce parcours, notamment concernant la question de la segmentation et la tokenisation, dans son rapport à des propriétés structurales dans un language.
- D'une manière génerale, je m'intéresse, d'un point de vue technique, aux fondements à la fois formels et conceptuels des modèles distributionnels de language.
- Je me retrouve donc, au bout de tout ce parcours, prêt à intégrer l'ensemble de ces éléments dans un programme de recherche global, dont j'ai donné des détails dans le projet soumis avec ma candidature, et dont je voudrais maintenant donner l'idée centrale.

Programme

- Lorsqu'on pense au pratiques de savoir qui se développent de nos jours sous le nom d'"Intelligence Artificielle" et des "grands modèles de langage" ou LLMs, 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 intruments d'analyse des practiques sociales (dans ce cas, de pratiques linguistiques, mais pas que), en devenant extraordinairement puissants, sont deguisés en instruments presque magiques, doués d'agentivité, et déployés au sein même de la société, avec des effets potentiellement désastreux.
- Mon projet vise donc a dévélopper, à la fois, une épistemologie (critique) et une théorie (y compris formelle) de ces outils, et notamment, des modèles distributionnels de language par apprentissage machine comme les LLMS, [qui ne rejette pourtant pas leur puissance, mais qui les prennent pour ce qu'ils sont, c'est-à-dire, des outils d'analyse des pratiques sociales].

Structure des données

- L'idée épistémologique centrale est la suivante.
- Un LLM est avant tout un *programme*
- or d'un point de vue strictément épistémologique, la seule question qu'on peut posser 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 donnés, qui elles sont censées representer 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 réclame une 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 maniére 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-spécifié 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 "Qu'est-ce que ces modèles connaissent?" par "Qu'est-ce que nous connaissons de ces modèles" et "Qu'est-ce que nous pouvons connaître à travers eux?"

Explicabilité formelle

- Or, évidément la question qui se pose est: est-ce que ça peut vraiment se faire ?
- Je voudrais d'abord rappeler que, dans un sens restreint, cela a déjà était fait, et ensuite esquisser rapidement la façon par laquelle le programme que je présente ici peut être compris comme une generalization de [cette orientation/ces résultats]
- Malgré une tendance et une rhétorique en faveur des modèles de bout en bout, les modèles de langage actuels sont souvent composés de, disons, trois algorithmes relativement indépendants exécutés de manière séquentielle, à savoir :
    - 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 corresopndant à des tokens dans une phrase, et modifie chacun en fonction des rapports à tous les autres dans ce contexte, pour produire des representatations vectorielles contextuelles
- Je vais me concentrer sur le plongement, qui est l'étape où a lieu une véritable extension de la structure formelle (puisqu'on passe d'un ensemble fini de tokens à un espace infini, avec des opérations et une géométrie soujacente). Sans compter que des représentations vectorielles de ce genre commencent à être utilisés 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 procedure algébrique bien connue, appelée décomposition en valeurs singulières (ou SVD), sur les données originaires organisées sous la forme d'une matrice
- Donc, si on prends, par exemple, tout wikipedia, et qu'on entraine un modèle neuronal d'embbeding dessus, on obtient un espace d'embeddings par la magie des réseaux de neurones, mais on peut montrer, donc, que on organisant ces donnés 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, et lorsqu'on prend les plus importants, on obtient aussi une espace d'embedding
- Et de maniére significative, des chercheurs ont montré que ce que le réseaux de neurones est en train de faire, c'est d'essayer d'atteindre ce même espace calculé directement et de manière exacte par SVD
- Donc, on a ici un cas d'explicitation de la structure des donnés 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 interpretable (chiffres, voyelles, consonnes, caractères spéciaux)
- 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 SVD, qui 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.
- Cette formalisation suggère que l'algèbre linéaire n'est qu'une parmi plusieurs instantiations possibles d'une structure plus génerale. Ce programme propose donc une généralisation de ce cadre basée sur la théorie des catégories.
- Cette géneralisation permet différentes instantiations, capables de réveler, chacune, 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 souhaietez
- Mais ce qu'il faut retenir, c'est que au moins certains aspects des modèles de langage peuvent être rapportés à des structures implicites des donnés qu'on sait extraire, et que ce mon programme propose, c'est d'élargir le genre de structure qu'on peut extraire par ces moyens.

Interprétabilité théorique

- Avant de conclure, je voudrais dire un mot sur l'autre fléche 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 épistemiques. 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 derivée 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'hypothès distributionnelle n'est qu'un corollaire d'un principe théorique plus fort, à savoir l'hypothèse structurale, qui maintien que le contenu linguistique est l'effet d'une structure virtuelle dérivee des pratiques lingusitiques 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 développes 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 inclue 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 developpés par le structuralisme classique coïncident presque parfaitement avec ceux que l'on peut identifier implicitement en acte dans les modèles actuelles, lorsqu'on les regarder à travers un prisme formel, comme celui que je viens de présenter.
- Le succès de ce travail interprétatif pourrait, dans un deuxième temps, deboucher sur une application de ces outils formels ailleurs que dans l'analyse du language. 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

-