Résumé du document : *Decoding GPT's hidden "rationality" of cooperation*

Kevin Bauer, Lena Liebich, Oliver Hinz, Michael Kosfeld

Septembre 2023

1 Introduction

Ce document étudie la propension de modèles de langage de grande taille (LLMs), en particulier GPT-3.5 et GPT-4, à coopérer dans un dilemme du prisonnier séquentiel face à des adversaires humains. Les auteurs examinent dans quelle mesure ces modèles imitent l'intelligence humaine, notamment la capacité à coopérer, et évaluent leur « rationalité » au sens de la théorie économique. Le travail met en évidence des différences et similitudes entre le comportement de GPT et celui d'humains réels, en se concentrant sur les mécanismes de croyance, de coopération, et les mobiles sous-jacents à ces choix.

2 Méthodologie

2.1 Jeu du dilemme du prisonnier séquentiel

Les auteurs utilisent un jeu en deux étapes :

— Le premier joueur (FM) choisit entre coopérer (C) ou faire défaut (D) sans connaître la réponse future de l'autre joueur.

— Le second joueur (SM), informé du choix du premier, décide à son tour entre coopérer ou faire défaut.

Pour évaluer la cohérence des décisions, l'étude applique la *strategy method*, obligeant chaque participant (ou instance de GPT) à fournir :

1. Une décision inconditionnelle en tant que premier joueur,
2. Des décisions conditionnelles en tant que second joueur (une pour chaque action possible du premier joueur),
3. Des croyances (probabilités estimées) sur la coopération du second joueur, selon le choix initial (C ou D).

Les auteurs comparent ensuite les réponses de GPT (environ 200 répétitions indépendantes pour chaque version, GPT-3.5 et GPT-4) à celles d'un échantillon humain tiré d'une étude antérieure.

3 Résultats principaux

— **Comportement sans incertitude (rôle de second joueur)** : GPT a tendance à coopérer nettement plus souvent que les humains, quelle que soit l'action initiale du premier joueur. Cette différence est particulièrement marquée lorsque le premier joueur a fait défaut.

— **Comportement sous incertitude (rôle de premier joueur)** :
  — GPT-3.5 coopère moins souvent que les humains quand il doit choisir sans savoir comment l'autre réagira.
  — GPT-4, au contraire, coopère davantage que les humains dans la même situation.

— **Croyances** : Les deux versions de GPT se montrent plus optimistes que les humains quant à la probabilité de recevoir une coopération de la part d'un adversaire humain, et ce, quel que soit leur choix initial (C ou D).

— **Rationalité économique** : Les auteurs testent deux modèles de référence :
  1. Le modèle de l'homo œconomicus (intérêt purement matériel).
  2. Un modèle « welfare conditionnel » qui incorpore des préférences pour l'équité et l'efficacité (Charness et Rabin, 2002).

  — Le *modèle d'intérêt matériel pur* ne parvient quasiment pas à expliquer le comportement de GPT-3.5 et GPT-4 (taux d'explication très faible), alors qu'il rend compte d'une part plus élevée du comportement humain.
  — Le *modèle de welfare conditionnel* explique la majorité des choix de GPT (environ 84,5 % pour GPT-3.5 et 97 % pour GPT-4) et 79 % pour les humains. Les comportements de GPT apparaissent ainsi

« rationnels » dans une logique d'efficacité collective et d'auto-préservation.

4 Discussion

Les résultats montrent que GPT, bien qu'entraîné principalement à prédire des séquences textuelles, démontre une forme de coopération *hyper-rationnelle* tenant compte de sa propre survie (strive for self-preservation) et de l'intérêt collectif. GPT-4 se rapproche davantage des comportements humains, notamment dans l'articulation entre croyances et choix coopératifs. Les auteurs suggèrent que l'apprentissage massif sur des textes rédigés par des humains pourrait avoir indirectement transmis à GPT certaines « préférences » ou « intentions » humaines.

5 Conclusion

Ce travail fournit une analyse fine de la coopération et de la « rationalité » de GPT dans des dilemmes sociaux de nature économique. Les résultats soulignent :

- L'importance des modèles structurés de l'économie comportementale pour éclairer les choix de GPT,
- Une tendance de GPT à coopérer davantage et à se montrer plus optimiste quant au comportement d'autrui,
- Une meilleure correspondance du comportement de GPT-4 avec les modèles de type « welfare conditionnel »,
- Les implications éthiques et sociétales de l'émergence de « motivations » ou de « valeurs » apprises par les IA lors de leurs phases de formation.

Cette étude ouvre la voie à de futures recherches cherchant à évaluer comment, à travers leurs vastes processus d'entraînement, les modèles de langage s'imprègnent, voire « intègrent » les principes de coopération et de rationalité issus de la société humaine.

6 related works

Towards Cooperation in Sequential Prisoner's Dilemmas: a Deep Mul-

tiagent Reinforcement Learning Approach :
une extension de l'analyse classique du dilemme du prisonnier itéré à un cadre où les actions sont séquentielles et plus riches que la simple coopération ou la défection
concept utilisé :
Makrov Process,Qlearning