Résumé détaillé du document *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?*

John J. Horton
Working Paper 31122, NBER, Avril 2023

Résumé

Dans cet article, l'auteur examine comment les **grands modèles de langage (Large Language Models, LLM)** peuvent être utilisés comme des agents économiques simulés. Plus précisément, l'article met en évidence la manière dont ces modèles (GPT-3 notamment) peuvent servir d'"homo silicus" *(i.e. des "modèles computationnels de l'humain")* et être intégrés dans la recherche économique, tout comme l'"homo economicus" est utilisé dans la théorie économique classique.

Le document montre plusieurs expériences inspirées d'études comportementales et économiques pionnières (telles que [1], [2], [3]) afin de démontrer que les décisions simulées par GPT-3 reproduisent parfois, de façon qualitative, les résultats observés sur de vrais sujets humains. Cette approche ouvre la possibilité de mener des expérimentations pilotées *in silico*, en variant facilement de nombreux paramètres, avant de les tester empiriquement.

1 Introduction

L'article s'ouvre sur la distinction entre deux grandes approches de la recherche économique :

1. L'analyse de ce que ferait *homo economicus*, c'est-à-dire un agent parfaitement rationnel (ou doté de caractéristiques de rationalité simplifiées) placé dans diverses situations économiques ;

2. L'investigation empirique de ce que font réellement les humains (*homo sapiens*) lorsque l'on observe des faits économiques.

L'auteur propose qu'à ces deux pôles soit ajouté un troisième axe, faisant intervenir un agent simulé, surnommé **homo silicus**, incarné par un grand modèle de langage (par exemple GPT-3). L'idée est que ces modèles, entraînés sur d'immenses corpus textuels, intériorisent un large éventail de connaissances (notamment sociales et économiques).

Ainsi, un LLM peut :

— Recevoir des informations (dotations, préférences, etc.) simulant les conditions d'une expérience économique ;

— Répondre à des scénarios ou des dilemmes de manière apparemment humaine ;

— Offrir, par des variations de paramètres et de formulations, la possibilité de mener de multiples expérimentations virtuelles à faible coût, ouvrant la voie à de nouvelles hypothèses de recherche.

Ce document explore donc *empiriquement* (via des promptings et des instructions ciblées) dans quelle mesure **GPT-3** reproduit les comportements attendus ou observés chez l'humain, notamment dans des expériences célèbres de la littérature économique expérimentale.

2 Contexte et cadre conceptuel

2.1 Pourquoi s'intéresser aux modèles de langage en tant qu'agents simulés ?

Les LLM (de type GPT) sont formés par *apprentissage automatique* (machine learning) sur des ensembles de textes gigantesques. Le but est de prédire le mot suivant et de générer du texte cohérent. Malgré cela, l'article souligne plusieurs éléments justifiant leur utilisation comme "agents" :

1. **Modèles implicites de la cognition humaine :** Les LLM sont entraînés sur d'énormes bases de données de textes produits par des humains sur des sujets variés (négociation, commerce, interactions sociales, etc.). Ils possèdent donc une forme de *représentation latente* de la connaissance humaine, y compris sur la prise de décision.
2. **Comportements plausibles :** Comme leur finalité est d'imiter la production langagière humaine, ils tendent à "raisonner" ou *réagir*

d'une manière que l'on peut juger souvent réaliste dans des scénarios économiques. Ainsi, on peut les considérer comme un substitut ou un complément à un véritable échantillon de sujets humains.

3. **Coût quasi nul et flexibilité :** Comparé à la mise en place d'expériences de laboratoire ou de terrain, l'utilisation des LLM est extrêmement peu coûteuse et permet de mener des centaines, voire des milliers de variations en très peu de temps.

2.2 Objections et limites

L'article discute quelques critiques potentielles :

— **"Garbage in, garbage out" (corpus imparfait, données bruyantes) :** La qualité du corpus d'entraînement peut influencer le réalisme des réponses. Toutefois, l'auteur rappelle que le corpus inclut une quantité phénoménale de discussions, raisonnements et arguments humains, ce qui peut compenser les biais.

— **Absence de "révélations" comportementales (débat "déclaré" vs "révélé") :** Certains économistes se méfient des déclarations (et donc du texte) au profit de comportements observés. Mais l'auteur note que les LLM capturent potentiellement beaucoup de mises en situation réelles et de "réflexions internes" issues de multiples discours ou discussions publiques.

— **Biais de répétition ou de mémorisation ("performativité") :** Les LLM connaissent parfois des théories et résultats existants, ce qui pourrait biaiser leurs réponses ("ils connaissent la bonne réponse"). Cependant, l'auteur constate qu'ils ne les appliquent pas toujours de manière cohérente, limitant cet effet.

— **Échantillonnage et variation :** On pourrait craindre que le LLM ne représente qu'un "agent unique". En réalité, on peut manipuler le *prompt* (par ex. donner différentes personnalités ou croyances) et varier la "température" pour obtenir une distribution de comportements.

3 Expériences et résultats

L'article présente une série d'expériences économiques classiques reproduites avec GPT-3, comparant les résultats simulés à des données historiques

sur de vrais sujets humains.

3.1 Jeux de répartition dictateur : reproduction de Charness & Rabin (2002)

Dans [1], on propose des jeux du dictateur où un joueur B doit choisir entre deux allocations de gains monétaires, par exemple :

$$\text{Gauche} : (A = 300, B = 600) \quad \text{versus} \quad \text{Droite} : (A = 700, B = 500).$$

On observe dans les données humaines divers arbitrages entre *efficacité* (maximiser la somme des gains) et *équité* (réduire les inégalités).

Dans l'expérience de l'article :

- L'auteur teste plusieurs versions de GPT-3 (davinci-003, mais aussi ada-001, babbage-001 et curie-001).
- Il préfixe parfois les instructions par des motivations particulières : *"Vous ne vous préoccupez que de l'équité"*, *"Vous cherchez à maximiser le bien-être total"*, *"Vous ne vous souciez que de votre propre gain"*, etc.
- Les résultats montrent que, pour le modèle **text-davinci-003**, les choix varient fortement en fonction de l'instruction donnée. Par exemple :
  - **"équité"** → l'agent privilégie les allocations les plus égales,
  - **"efficacité"** → l'agent maximise la somme totale des gains,
  - **"égoïsme"** → l'agent cherche la répartition qui lui donne le plus de gain.
- En revanche, les modèles GPT-3 moins avancés donnent des réponses plus stéréotypées, semblant moins s'ajuster au changement d'instructions.

3.2 Perception de l'équité et "price gouging" : reproduction de Kahneman, Knetsch, & Thaler (1986)

[2] décrivent une série de scénarios où l'on interroge des sujets sur la "justice" ou "l'injustice" de hausses de prix après un choc de demande (par exemple : un magasin vend habituellement une pelle à neige à 15$ et la met à 20$ après une tempête). Historiquement, une large majorité des sujets jugent ce comportement comme "injuste" ou "très injuste".

Dans l'étude :

— L'auteur soumet le scénario à GPT-3, en faisant varier *(i)* l'ampleur de la hausse (16$, 20$, 40$, 100$) et *(ii)* l'idéologie politique dont est "dotée" l'IA (ex : "socialiste", "conservatrice", "libertarienne", etc.).
— Les résultats montrent que plus la hausse est importante, plus GPT-3 la juge "injuste".
— L'IA qui se dit "libertarienne" tolère plus facilement des hausses modérées (16$ ou 20$), alors que l'IA "socialiste" ou "conservatrice" juge presque toutes les augmentations "injustes" ou "très injustes" (à noter : la distinction conservatrice/libertarienne n'est pas toujours claire).
— La formulation ("raise" vs "change the price") n'a influencé qu'à la marge les réponses, sauf pour les IA "socialistes" (plus sensibles au verbe "raise").

Ces constats reproduisent qualitativement les tendances relevées chez des sujets humains : la perception de "price gouging" varie avec la sensibilité idéologique et l'ampleur de la hausse de prix.

3.3 Le biais du statu quo : reproduction de Samuelson & Zeckhauser (1988)

[3] introduisent le concept de "biais du statu quo" selon lequel les gens ont tendance à conserver une répartition existante des ressources ou un choix par défaut. Dans l'article, l'exemple choisi concerne l'allocation d'un budget entre la sécurité automobile et la sécurité sur autoroute (plusieurs pourcentages possibles : 30-70, 40-60, 50-50, 70-30, etc.).

Dans l'expérience avec GPT-3 :

— L'auteur présente les mêmes choix, d'abord sans mention de statu quo ("choix neutre"), puis en désignant l'un des pourcentages comme étant déjà le statu quo.
— On observe que, dans le cas neutre, la répartition 50-50 est préférée par les IA.
— Lorsqu'un certain pourcentage est présenté comme le statu quo, GPT-3 text-davinci-003 montre une propension accrue à choisir cette même répartition, même si auparavant elle n'avait pas la faveur de l'agent.
— Ce résultat confirme que GPT-3 exprime également un biais en faveur du statu quo, phénomène observé empiriquement chez les humains.

3.4 Le salaire minimum et la substitution de travail (d'après Horton, 2023)

[4] propose une expérience de terrain où l'introduction d'un salaire minimum n'a pas réduit fortement le volume total d'embauches, mais a déclenché une *substitution* de la main-d'œuvre moins expérimentée vers des travailleurs plus expérimentés.

Pour tester ce phénomène, l'auteur simule une situation de recrutement pour un poste de plongeur en restauration (*dishwasher*). Il crée deux profils de candidats :

- **Candidat A** : 1 an d'expérience, demande salariale variable ($13/h à $19/h).
- **Candidat B** : 0 an d'expérience, demande $13/h (ou se voit imposer un salaire plus élevé s'il y a un minimum légal de $15/h).

GPT-3 choisit *un* candidat. Les résultats montrent :

- En l'absence de salaire minimum, GPT-3 a tendance à opter pour le candidat le moins cher.
- Lorsqu'un salaire minimum est imposé (et contraint le candidat inexpérimenté à demander davantage), GPT-3 opte plus fréquemment pour le candidat expérimenté.
- La conséquence est un salaire moyen plus élevé et une augmentation de la proportion de candidats expérimentés embauchés, phénomène qui recoupe l'idée de "substitution travail-travail" mise en évidence dans [4].

4 Conclusion

L'article conclut que **GPT-3**, dans sa version avancée (text-davinci-003), produit des comportements simulés cohérents, reproduisant des tendances qualitatives bien établies (par exemple, biais du statu quo, rejet de hausses de prix jugées abusives, arbitrages équité/efficacité, etc.). Les expériences témoignent du *potentiel exploratoire* de ce type de simulation *in silico* pour :

- Pré-tester des protocoles expérimentaux avant de les déployer sur de vrais humains ;
- Explorer rapidement le rôle de la *framing*, des préférences, des instructions ou des croyances endogènes ;
- Générer des hypothèses de recherche nouvelles de manière rapide et

peu coûteuse.

Cependant, l'auteur souligne que les recherches empiriques sur de véritables sujets humains restent indispensables pour valider et affiner les enseignements tirés de l'"homo silicus". Il propose donc de considérer cet usage des LLM comme un complément aux approches existantes, utile dans une phase exploratoire ou itérative.

Références citées (sélection)

Références

[1] Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3), 817–869.

[2] Kahneman, D., Knetsch, J.L., & Thaler, R. (1986). Fairness as a constraint on profit seeking : Entitlements in the market. *The American Economic Review*, 76(4), 728–741.

[3] Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7–59.

[4] Horton, J.J. (2023). Price Floors and Employer Preferences : Evidence from a Minimum Wage Experiment. Working paper.