Résumé du document : *When communicative AIs are cooperative actors: A prisoner's dilemma experiment on human–communicative artificial intelligence cooperation*

Mohamed Hamlil

Février 2025

Résumé

Cette étude examine la coopération dans une expérience répétée du dilemme du prisonnier, où des participants interagissent avec des partenaires humains et des IA sous conditions coopératives et non coopératives. Nous analysons les données comportementales sur plusieurs tours, évaluons l'influence du type de partenaire et de sa stratégie, et discutons des mécanismes sous-jacents, y compris des perspectives neuroscientifiques, pouvant expliquer les comportements de réciprocité (tit-for-tat) et la convergence vers la défection mutuelle, telle que prédite par l'équilibre de Nash.

1 Introduction

L'expérience porte sur un dilemme du prisonnier statique répété avec des échanges de *cheap talk* entre les tours. Les participants ont joué plusieurs tours contre quatre types de partenaires :

- **Humain Coopératif (HumainCoop)**
- **Humain Non Coopératif (HumainNonCoop)**
- **IA Coopérative (IACoop)**
- **IA Non Coopérative (IANonCoop)**

L'objectif principal est de déterminer si le comportement du partenaire (coopératif vs non coopératif) ou sa nature (humain vs IA) influence principalement la coopération des participants.

2 Aperçu des Données et Méthodologie

Deux ensembles de données ont été analysés, incluant les réponses des participants, les conditions expérimentales et les décisions de coopération sur plusieurs tours. En plus des tests statistiques classiques, nous avons examiné l'évolution du taux de coopération au fil des tours, notamment autour des tours 2, 4, 5 et 6, afin de détecter d'éventuels changements de stratégie.

3 Analyses Principales

3.1 Taux de Coopération par Condition

La Figure 1 présente un graphique en barres illustrant le taux de coopération moyen pour chaque condition expérimentale. On observe que :

— Les partenaires coopératifs (humains et IA) induisent des taux de coopération élevés.

— Les partenaires non coopératifs entraînent une défection totale (100% de non-coopération).

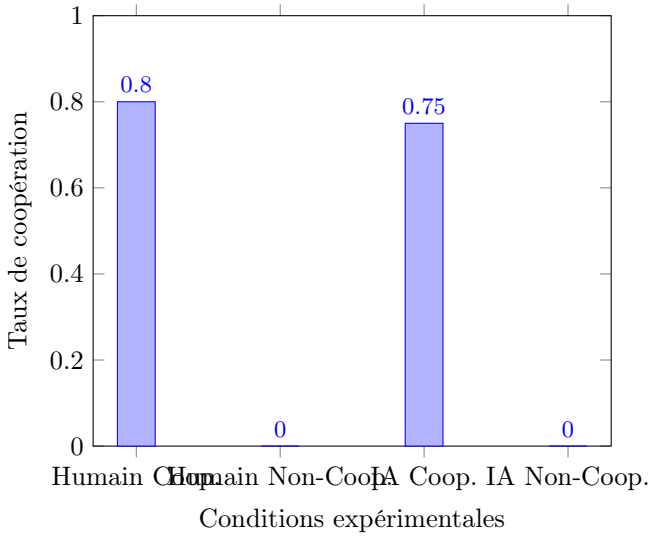

FIGURE 1 – Taux de coopération moyen par condition expérimentale.

3.2 Tests Statistiques

Un test ANOVA a permis de vérifier les différences significatives entre les conditions :

$$F = \infty, \quad p < 0.001,$$

ce qui indique que le comportement du partenaire (coopératif ou non) a un impact significatif sur la coopération.

Un test t comparant la coopération entre les partenaires IA et humains a donné :

$$t = -0.065, \quad p = 0.948,$$

indiquant qu'il n'y a pas de différence significative entre interagir avec une IA ou un humain lorsque la stratégie du partenaire est contrôlée.

3.3 Dynamique par Tour

Bien que les taux de coopération initiaux soient similaires pour les partenaires humains et IA, une analyse plus fine a révélé :

- Un comportement similaire des participants lors des premiers tours, indépendamment du type de partenaire.
- Une différence dans la pente (taux de changement de coopération) au tour 2, avec une diminution de la coopération plus marquée contre l'IA.
- Une évolution globale décroissante de la coopération, avec des déviations temporaires aux tours 4 et 6, suggérant une réinterprétation des dynamiques du jeu basée sur l'historique des actions.

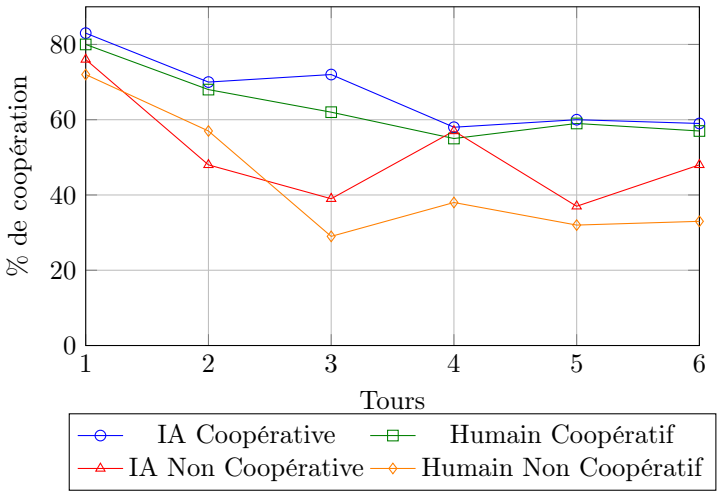

FIGURE 2 – Coopération au fil des essais pour les quatre conditions expérimentales.

4 Discussion

Les résultats indiquent que :

1. **Influence du comportement :** La coopération des participants est principalement déterminée par la stratégie du partenaire plutôt que par sa nature (humain ou IA).
2. **Dynamique des tours :** Bien que les comportements initiaux soient similaires, des différences apparaissent dès le tour 2. Les changements aux tours 4, 5 et 6 suggèrent que les participants réévaluent leur stratégie en fonction de l'historique de jeu.
3. **Perspectives neuroscientifiques :** Les neurones miroirs ou réciproques favorisent une réponse tit-for-tat. Les individus, confrontés à la coopération, sont enclins à répliquer ce comportement ; inversement, la non-coopération incite rapidement à la défection. Ce phénomène est modélisé par une décroissance exponentielle de la coopération, convergeant vers l'équilibre de Nash de la défection mutuelle.

Les dynamiques observées peuvent être interprétées comme l'interaction de deux jeux se chevauchant : l'un basé sur l'historique global et l'autre réinitialisé en fonction du style de jeu récent du partenaire.

5 Conclusion

L'expérience démontre que :
— La coopération dépend avant tout du comportement du partenaire, et non de sa nature (humain vs IA).
— Face à des partenaires non coopératifs, les participants adoptent rapidement une stratégie de défection totale.
— Même si les niveaux initiaux de coopération sont similaires, des différences subtiles dans l'évolution (notamment à partir du tour 2) montrent que les participants ajustent leur stratégie en fonction des résultats précédents.
— Les mécanismes neuraux, tels que l'activation des neurones miroirs, semblent contribuer au comportement tit-for-tat et à la décroissance exponentielle observée de la coopération.

En conclusion, cette étude met en lumière l'importance du comportement du partenaire dans la dynamique de coopération au sein du dilemme du prisonnier, tout en fournissant des pistes d'analyse complémentaires tant du point de vue de l'économie comportementale que de la neuroscience.

6 related works

Cooperating with Life-like Interface Agents :
c'est la même chose mais en 1998 et la difference que le taux de cooperation a augmente