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

Yu-Leung Ng (2023)

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 décisions de coopération sur plusieurs tours, évaluons l'influence du type de partenaire et de sa stratégie, et discutons des mécanismes sous-jacents expliquant 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 aux 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).

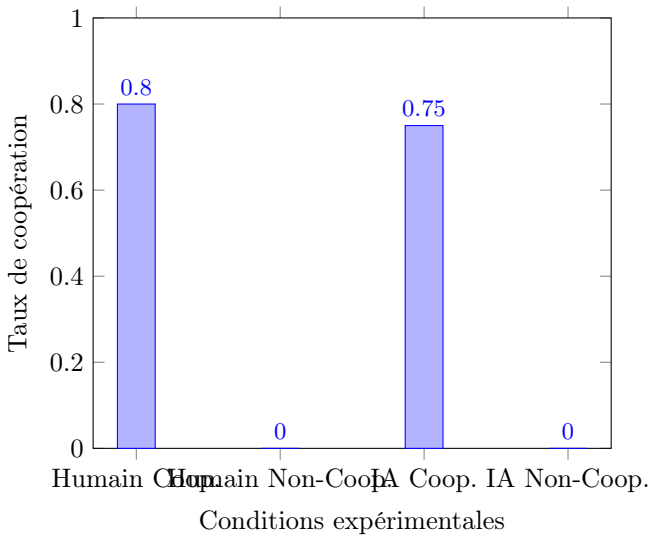

FIGURE 1 – Taux de coopération moyen par condition expérimentale.

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 a un impact significatif sur la coopération.

Un test t comparant la coopération entre 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 initiaux de coopération soient similaires pour les partenaires humains et IA, l'analyse a révélé :

— Un comportement similaire lors des premiers tours, indépendamment du type de partenaire.
— Une diminution plus marquée de la coopération au tour 2 contre l'IA.
— Une évolution globale décroissante de la coopération, avec des déviations temporaires aux tours 4 et 6, suggérant une réévaluation des dynamiques de jeu en fonction de l'historique des actions.

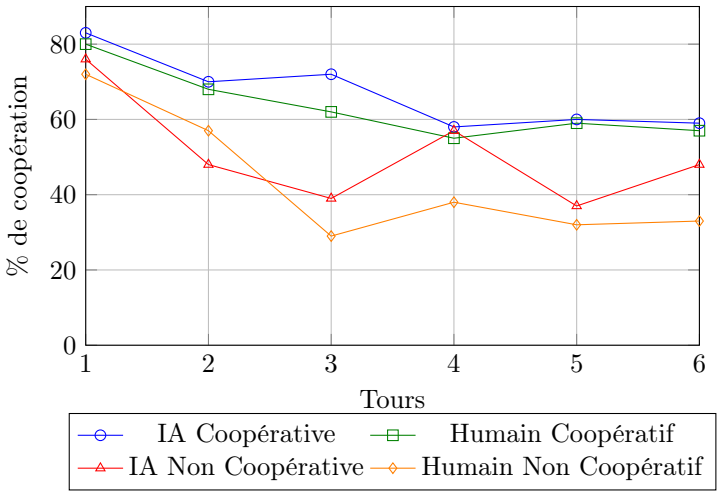

FIGURE 2 – Coopération au fil des essais pour les quatre conditions expérimentales.

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 :** Des différences apparaissent dès le tour 2, et les fluctuations aux tours 4, 5 et 6 montrent que les participants ajustent leur stratégie en fonction des actions passées.
3. **Mécanismes sous-jacents :** Les comportements de tit-for-tat et la convergence vers la défection mutuelle s'expliquent par la logique des

jeux répétés et l'équilibre de Nash, où la réciprocité initiale laisse progressivement place à la défection lorsque la confiance est rompue.

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.

— Les différences subtiles dans l'évolution du taux de coopération (notamment à partir du tour 2) montrent une adaptation continue des participants aux résultats précédents.

— Les mécanismes de réciprocité, tels que tit-for-tat, conduisent à une décroissance graduelle de la coopération vers l'équilibre de Nash de défection mutuelle.

6 Travaux Connexes

Une étude antérieure de Nass et al. (1998), intitulée *Cooperating with Life-like Interface Agents*, présente une expérience similaire utilisant des agents à interface « vivante ». Les auteurs y rapportent un taux de coopération globalement plus élevé, suggérant que l'anthropomorphisme de l'agent peut renforcer la confiance et encourager la coopération des participants [Nass et al.(1998)].

Références

[Nass et al.(1998)] Nass, C., Steuer, J., & Tauber, E. R. (1998). Cooperating with Life-like Interface Agents. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 37–42.