Résumé de « Ultimatum bargaining: Algorithms vs. Humans »

Ali I. Ozkes, Nobuyuki Hanaki, Dieter Vanderelst, Jürgen Willems
SKEMA Business School ; Osaka University ; University of Cincinnati ; WU Vienna

Résumé

Cette note résume Ozkes et al. (2024), qui étudient le comportement humain dans le Jeu de l'Ultimatum face à des adversaires humains ou algorithmiques différant par leur objectif (imitation vs optimisation), leur explicabilité et le bénéficiaire des gains. Sur 404 participants britanniques, ils montrent que le seuil minimal d'acceptation des répondants (MAO) ne varie ni selon le type d'adversaire ni selon le degré d'explication, mais que les proposants offrent légèrement moins lorsque les gains de l'algorithme ne profitent à aucun bénéficiaire plutôt qu'à un autre humain. Les sujets déclarent préférer les partenaires humains, mais se comportent de façon similaire envers humains et algorithmes.

1 Le Jeu de l'Ultimatum

Le Jeu de l'Ultimatum se joue entre deux participants : le proposant et le répondant. Le proposant se voit attribuer une somme fixe de points (ici 100) et propose comment la partager avec le répondant. Le répondant doit ensuite accepter ou rejeter l'offre simultanément :

— Si le répondant accepte, les points sont distribués selon la proposition (par exemple, 60 pour le proposant et 40 pour le répondant).
— Si le répondant rejette, aucun des deux ne reçoit rien (0–0).

Cette structure met en balance rationalité économique (tout gain est préférable à rien) et sens de l'équité.

Condition « aucun bénéficiaire (NO) »

Dans les conditions impliquant un algorithme, on distingue deux cas selon le bénéficiaire des points attribués à l'algorithme :

— **TP (Token Player) :** les points de l'algorithme vont à un joueur humain fictif.
— **NO (No receiver) :** les points de l'algorithme ne profitent à personne, c'est-à-dire qu'en cas d'acceptation ou de rejet, la part réservée à l'algorithme est simplement perdue.

En condition NO, les répondants savent que rejeter l'offre ne punit pas un autre humain, mais fait perdre des points « au système ». Ce cadre teste si la motivation à partager équitablement dépend de la présence d'un bénéficiaire humain.

2 Introduction

Les algorithmes sont de plus en plus utilisés dans des domaines requérant justice et interaction stratégique. La littérature antérieure montre des effets mitigés sur la confiance, la perception d'équité et l'engagement émotionnel (e.g., Knoch et al. 2006 ; Erlei et al. 2022). Cet article se demande si les proposants et répondants humains du Jeu de l'Ultimatum se comportent différemment selon qu'ils sont appariés avec : (i) un adversaire humain vs algorithmique, (ii) un algorithme qui imite le comportement humain (MA) vs optimise égoïstement (OA) vs sans explication (NA), et (iii) un algorithme dont les gains bénéficient à un autre humain (TP) vs aucun bénéficiaire (NO).

3 Conception expérimentale

Chaque sujet joue trois manches en tant que proposant et trois manches en tant que répondant dans le Jeu de l'Ultimatum (répartition 0–100 points), avec soumission simultanée des stratégies.

- **Types d'adversaires :** Humain (HU), Algorithme imitant (MA), Algorithme optimisant (OA), Algorithme sans explication (NA).
- **Bénéficiaire des gains :** Joueur fictif (TP) vs aucun bénéficiaire (NO).
- **Hypothèses :**
  1. Les répondants accepteront des offres plus faibles de la part des algorithmes, notamment OA et dans NO.
  2. Les proposants offriront de façon similaire à MA et HU, mais moins à OA, NA et dans NO.

404 participants au Royaume-Uni (échantillon équilibré selon le genre) ont été recrutés via Prolific et ont réussi les tests de compréhension. Les rôles et l'ordre des conditions ont été contrebalancés.

4 Résultats

4.1 Comportement des répondants

- Pas de différence significative des MAO entre adversaires humains et algorithmiques.
- Aucun effet de l'objectif de l'algorithme (MA vs OA) ni du niveau d'explication (MA vs NA).

4.2 Comportement des proposants

- Les partages égalitaires (50–50) sont plus fréquents avec un humain qu'avec un algorithme.
- Les offres à MA sont similaires à HU quand TP, mais environ 4 points inférieures dans NO.
- Pas de réduction significative des offres à OA vs MA, ni à NA vs HU.

4.3 Préférences et démographie

- Enquête post-jeu : 40 % préfèrent un adversaire humain, 20 % un algorithme.
- Les sujets plus jeunes tendent à choisir les algorithmes ; pas d'effet du genre ou du revenu.

5 Discussion

Les participants ne distinguent pas de façon systématique les partenaires humains et algorithmiques dans leurs stratégies du Jeu de l'Ultimatum, ni entre différents types d'algorithmes ou niveaux d'explicabilité, sauf pour pénaliser davantage les algorithmes dont les gains ne profitent à personne. Malgré une préférence déclarée pour l'interaction humaine, ils se comportent de même envers les algorithmes. Ces résultats suggèrent qu'un soutien algorithmique peut être aussi accepté qu'un partenaire humain, pour peu que ses gains soient perçus comme socialement pertinents.

Références

[1] Ozkes, A. I., Hanaki, N., Vanderelst, D., Willems, J. (2024). *Ultimatum bargaining : Algorithms vs. Humans. Economics Letters*, 244, 111979.