L'Actualité économique

Jeu du dictateur et jeu de la confiance : préférences distributives *vs* préférences dépendantes des croyances

Giuseppe Attanasi et Kene Boun My

Volume 92, numéro 1-2, mars–juin 2016

Économie expérimentale : comportements individuels, stratégiques et sociaux

URI : https://id.erudit.org/iderudit/1039878ar
DOI : https://doi.org/10.7202/1039878ar

Aller au sommaire du numéro

Éditeur(s)

HEC Montréal

ISSN

0001-771X (imprimé)
1710-3991 (numérique)

Découvrir la revue

Citer cet article

Attanasi, G. & Boun My, K. (2016). Jeu du dictateur et jeu de la confiance : préférences distributives *vs* préférences dépendantes des croyances. *L'Actualité économique*, *92*(1-2), 249–287. https://doi.org/10.7202/1039878ar

Résumé de l'article

L'article présente une revue de la littérature théorique et expérimentale sur les préférences sociales dans le *jeu du dictateur* et dans le *jeu de la confiance*. Deux types de préférences sociales sont analysés, l'aversion à l'inégalité associée aux préférences distributives et l'aversion à la culpabilité associée aux préférences dépendantes des croyances. Nous montrons dans un premier temps comment les deux types de préférences peuvent expliquer les déviations par rapport aux prédictions théoriques de la théorie des jeux standard. Nous discutons ensuite de l'insuffisance des modèles avec préférences distributives pour expliquer en détail les données expérimentales obtenues pour ces deux jeux de dilemme social. Notre principale conviction est que les modèles avec préférences dépendantes des croyances sont plus adaptés pour décrire l'hétérogénéité des comportements habituellement détectée dans de telles expériences. En particulier, sous l'hypothèse que la fonction d'utilité des joueurs dépend également de leurs croyances, alors la corrélation, entre le choix du dictateur (du receveur, dans le jeu de la confiance) et ses croyances de second ordre, habituellement observée pour ces deux familles d'expériences, peut être expliquée. Ce que ne permettent pas la théorie des jeux standard et ses extensions avec préférences distributives.

L'Actualité économique, Revue d'analyse économique, vol. 92, nos 1-2, mars-juin 2016

JEU DU DICTATEUR ET JEU DE LA CONFIANCE : PRÉFÉRENCES DISTRIBUTIVES *VS* PRÉFÉRENCES DÉPENDANTES DES CROYANCES*

Giuseppe ATTANASI
BETA, CNRS UMR 7522
Université de Strasbourg
giuseppe.attanasi@univ-lille1.fr

Kene BOUN MY
BETA, CNRS UMR 7522
Université de Strasbourg
bounmy@unistra.fr

Résumé – L'article présente une revue de la littérature théorique et expérimentale sur les préférences sociales dans le *jeu du dictateur* et dans le *jeu de la confiance*. Deux types de préférences sociales sont analysés, l'aversion à l'inégalité associée aux préférences distributives et l'aversion à la culpabilité associée aux préférences dépendantes des croyances. Nous montrons dans un premier temps comment les deux types de préférences peuvent expliquer les déviations par rapport aux prédictions théoriques de la théorie des jeux standard. Nous discutons ensuite de l'insuffisance des modèles avec préférences distributives pour expliquer en détail les données expérimentales obtenues pour ces deux jeux de dilemme social. Notre principale conviction est que les modèles avec préférences dépendantes des croyances sont plus adaptés pour décrire l'hétérogénéité des comportements habituellement détectée dans de telles expériences. En particulier, sous l'hypothèse que la fonction d'utilité des joueurs dépend également de leurs croyances, alors la corrélation, entre le choix du dictateur (du receveur, dans le jeu de la confiance) et ses croyances de second ordre, habituellement observée pour ces deux familles d'expériences, peut être expliquée. Ce que ne permettent pas la théorie des jeux standard et ses extensions avec préférences distributives.

* Nous tenons à remercier les éditeurs de ce numéro spécial, Ghislaine Messner, Benjamin Ouvrard et les deux évaluateurs anonymes pour leurs précieux commentaires. Nous sommes également reconnaissants du soutien financier que nous a apporté l'Attractivité IDEX 2013 de l'Université de Strasbourg.

Introduction

Selon l'une des hypothèses dominantes de la théorie économique standard, les individus sont uniquement motivés par leurs propres gains monétaires. La plupart des modèles économiques pose en effet l'hypothèse que les agents maximisent uniquement leurs gains monétaires espérés. Cependant, il a été prouvé dans de nombreuses disciplines que beaucoup d'individus dévient de manière significative et récurrente de ce comportement de maximisation purement individualiste lorsqu'ils sont en interaction stratégique avec d'autres individus.

Elster (1998), lui-même, dans son article « *Emotions and Economic Theory* », soutient qu'une grande variété d'*émotions* a des conséquences économiques importantes, et il regrette le peu d'attention que les économistes ont porté à cette dimension affective. Il est convaincu que l'une des caractéristiques principales des émotions est qu'elles sont déclenchées par des croyances (p. 49). Il y évoque toute une série d'émotions telles que la colère, la haine, la culpabilité, la honte, la fierté, l'admiration, le regret, la joie, la déception, l'exaltation, la crainte, l'espoir, la joie, le chagrin, l'envie, la malice, l'indignation, la jalousie, la surprise, l'ennui, le désir sexuel, le plaisir, le souci et la frustration. Il se demande également comment les émotions pourraient aider à la compréhension des comportements économiques pour lesquels des explications appropriées semblent faire défaut (p. 48).

De même, de nombreuses expériences économiques en interactions stratégiques menées en laboratoire ont confirmé que les participants sont influencés par des préférences sociales : plutôt que de se préoccuper uniquement d'eux-mêmes, ces individus se soucient également de la personne avec laquelle ils interagissent, même si cette personne leur est étrangère. Comportementalistes ou théoriciens, les économistes devraient donc s'intéresser aux déviations par rapport à la maximisation des gains monétaires espérés : les comportementalistes en analysant les motivations des sujets qui ne maximisent pas leurs gains espérés dans leur interaction stratégique, et les théoriciens en intégrant ces motivations dans leurs modèles en allant au-delà de la simple maximisation des gains monétaires espérés. C'est pourquoi il nous paraît légitime de nous poser la question quant à la capacité de la théorie des jeux standard à analyser et à modéliser l'influence des sentiments, des émotions et des normes sociales dans le comportement des joueurs. La théorie économique a essayé de traiter ces préférences sociales principalement de deux manières.

De nombreux exemples de la vie courante suggèrent que les sentiments, les émotions et les normes sociales font partie des préférences des individus quant à leurs allocations de consommation (préférences distributives). Par exemple, un sujet qui doit partager une tarte avec un autre individu peut être sensible à la part de la tarte laissée à l'autre individu ou à la différence de taille entre sa part et celle qu'il laisse à l'autre.

Lorsque les individus se trouvent en situation d'interaction stratégique, ils peuvent également être préoccupés par les croyances des autres : « Qu'est-ce qu'on attend de moi? », « Qu'est-ce que les autres vont penser de moi? ». Dans la théorie

des jeux standard, un joueur forme des croyances sur les choix de ses partenaires pour déterminer sa propre stratégie. Dans ce cas, les croyances n'ont qu'un impact indirect sur l'utilité des joueurs. *A contrario*, lorsqu'un joueur se préoccupe aussi des sentiments, émotions et/ou normes sociales, son utilité peut aussi dépendre directement des croyances qu'il a sur les choix, les informations ou les croyances des autres joueurs (préférences dépendantes des croyances). Considérons l'exemple suivant, tiré de Dufwenberg (2008) : Karen se sent coupable si elle déçoit quelqu'un. En réglant la facture de son paysagiste (Jim), cela influence son pourboire. Plus elle croit que Jim pense qu'il recevra un pourboire important, plus elle va donner un pourboire important. Plus précisément, elle donne autant qu'elle croit que Jim croit qu'il obtiendra, pour éviter un sentiment de culpabilité qui va la tourmenter si elle donne moins.

Comme nous le verrons dans la section 3, les résultats expérimentaux dans les jeux de dilemme social montrent la pertinence des deux notions que sont les préférences distributives et les préférences dépendantes des croyances. Ces résultats prouvent qu'un sujet peut avoir à la fois des préférences distributives et des préférences dépendantes des croyances.

Notons que le but de cet article n'est pas d'alimenter une discussion générale sur les deux courants théoriques, ni de créer des polémiques quant à la pertinence d'une théorie par rapport à l'autre. Bien au contraire, nous voulons montrer à travers cet article comment ces différentes théories de préférences sociales peuvent décrire de manière optimale le comportement des joueurs dans différents jeux. Selon le jeu considéré, certains sujets seront davantage motivés par la répartition des gains (préférences distributives), d'autres se préoccuperont davantage de leurs croyances sur les actions et les croyances des autres joueurs (préférences dépendantes des croyances).

C'est avec cet objectif à l'esprit que nous concentrons notre étude sur deux des plus importants concepts présents dans les préférences sociales, et qui ont été étudiés durant les 20 dernières années tant de manière théorique qu'expérimentale : l'aversion à l'inégalité et l'aversion à la culpabilité. Elles représentent toutes deux des motivations capitales capables d'expliquer le comportement des sujets réels dans les dilemmes sociaux.

L'aversion à l'inégalité, caractérisée comme une préférence distributive, a été définie en premier par Fehr et Schmidt (1999) comme une aversion aux revenus qui sont perçus comme inégalitaires. Par exemple, l'aversion à l'inégalité peut expliquer le refus des entreprises de diminuer les salaires dans les périodes de récession par crainte que les salariés (par leur aversion à l'inégalité) ne perçoivent ces coupes salariales comme injustes, ce qui pourrait nuire au moral au travail.

L'aversion à la culpabilité a quant à elle été modélisée comme une préférence distributive mais aussi comme une préférence dépendante des croyances (voir la section 1). Baumeister *et al.* (1994) pensent que « de manière générale, si les gens se sentent coupables d'avoir blessé leurs partenaires, de les avoir négligés et d'avoir échoué à se montrer à la hauteur de leurs attentes, ils changeront leur comportement

(pour éviter la culpabilité) de sorte qu'ils puissent conserver et renforcer leurs relations » (p. 247). Dans cet article nous nous focaliserons sur la modélisation de l'aversion à la culpabilité comme une préférence dépendante des croyances (Battigalli et Dufwenberg, 2009) qui est la plus répandue dans la littérature économique, en particulier dans les jeux de dilemme social. Par exemple, l'aversion à la culpabilité, comme une motivation dépendante des croyances, peut jouer un rôle déterminant dans les échanges économiques mis à mal par l'aléa moral et par des contrats incomplets. Dans de telles circonstances, des individus égoïstes pourraient empêcher la création de relations mutuellement avantageuses. La présence d'un nombre suffisamment important d'individus sensibles à la culpabilité aiderait à surmonter cette inefficacité puisque la peur de décevoir un partenaire qui nous fait confiance pourrait nous inciter à mieux respecter un accord du fait de notre aversion à la culpabilité.

Ces deux sentiments ont beaucoup moins d'influence sur le comportement lorsque plusieurs sujets sont impliqués dans des interactions stratégiques. En effet, tout porte à croire que les individus exploitent beaucoup plus leur pouvoir de négociation dans des marchés concurrentiels, que dans des situations de négociations bilatérales (Fehr et Schmidt, 1999). De même, ces individus auront des comportements de passager clandestin dans les jeux de contribution volontaire à un bien public, mais pas lorsqu'ils sont deux individus à décider combien ils vont contribuer à un bien public. Selon Battigalli et Dufwenberg (2007), un tel résultat peut s'expliquer par le fait que l'absence d'altruisme ne peut pas être reprochée à un joueur en particulier lorsque le même choix est effectué par un groupe de joueurs. Pour qu'une préférence sociale puisse avoir un impact significatif sur le comportement des sujets dans un environnement stratégique, les relations doivent être interpersonnelles, c'est pourquoi les jeux de dilemme social à deux joueurs constituent le meilleur environnement où ces préférences peuvent être élicitées.

L'aversion à l'inégalité est pertinente d'un point de vue comportemental dans trois jeux standards de dilemme social à deux joueurs : le jeu du dictateur, le jeu de la confiance et le jeu de l'ultimatum. L'aversion à la culpabilité – comme une préférence dépendante des croyances – semble quant à elle ne l'être que dans le jeu du dictateur et dans le jeu de la confiance (et dans leurs variantes). À notre connaissance, il n'existe aucune étude expérimentale et théorique sur l'aversion à la culpabilité dépendante des croyances dans le jeu de l'ultimatum. En outre, dans le jeu de l'ultimatum, les croyances de second comme de premier ordre sont stratégiquement importantes car elles ont un impact indirect sur les transferts de revenus des joueurs, contrairement à l'aversion à la culpabilité qui revendique quant à elle une influence des croyances non stratégiques (Bicchieri et Chavez, 2010).

C'est pourquoi nous limiterons notre analyse théorique et expérimentale de l'aversion à l'inégalité et à la culpabilité seulement au jeu du dictateur et au jeu de la confiance. Ces deux jeux sont liés d'un point de vue stratégique : la deuxième étape du jeu de la confiance peut se résumer à un simple jeu du dictateur. De plus, il s'agit de deux jeux de dilemme social bien connus et largement analysés dans la littérature expérimentale sur les préférences sociales.

Le jeu du dictateur (introduit pour la première fois par Kahneman *et al*., 1986) est une sorte de dilemme social où le premier joueur, le *proposant*, décide de manière unilatérale de l'allocation (ou de la répartition) d'une certaine dotation. Le deuxième joueur, le *receveur*, reçoit la partie de la dotation laissée par le proposant. Le rôle du receveur est donc totalement passif, il n'a aucune influence dans la répartition des gains dans le jeu.[1] Dans l'analyse théorique présentée dans la section 2, nous considérons une version simplifiée de ce jeu, appelé le mini-jeu du dictateur, dans lequel le proposant peut, soit décider une répartition égalitaire (50-50) de la dotation, soit s'accaparer de la totalité de cette dotation pour son propre intérêt.

Le jeu de la confiance représente quant à lui un dilemme social entre un agent A et un agent B. A est l'investisseur qui accorde sa confiance et prend une décision coûteuse qui génère un rendement social et B est le receveur, celui à qui on a accordé sa confiance et qui décide comment redistribuer les revenus entre lui et l'agent A (Berg *et al*., 1995; Buskens et Raub, 2013). De même que précédemment, nous considérerons une version simplifiée de ce jeu, appelé le mini-jeu de la confiance, où A décide entre choisir ou non une action qui lui est coûteuse et B décide, soit de partager de manière équitable le revenu produit par l'action coûteuse de A, soit de tout garder pour lui.

La suite de l'article est structurée de la manière suivante. Dans la section 1, nous passons en revue les principaux modèles avec préférences distributives et les modèles avec préférences dépendantes des croyances, en montrant comment des sentiments semblables peuvent être modélisés selon l'un ou l'autre type de préférences et comment les deux types de préférences peuvent être intégrés dans un même modèle.

Dans la section 2, nous explicitons les différentes prédictions des théories de l'aversion à l'inégalité (Fehr et Schmidt, 1999) et à la culpabilité (Battigalli et Dufwenberg, 2009) relatives aux comportements des joueurs dans le jeu du dictateur et dans le jeu de la confiance. Dans la section 2.1, dans le contexte d'un jeu du dictateur, nous analysons le comportement optimal d'un dictateur averse à l'inégalité puis d'un dictateur averse à la culpabilité. Dans la section 2.2, dans le cadre du jeu de la confiance, nous analysons le comportement optimal des deux joueurs, lorsque le receveur est averse à l'inégalité et lorsqu'il est averse à la culpabilité.

Dans la section 3, nous donnons les principaux résultats des études expérimentales concernant les préférences sociales pour le jeu du dictateur et le jeu de la confiance.

---

1. Le jeu du dictateur n'est pas formellement un jeu tel que le terme est utilisé dans la théorie des jeux. Dans un véritable jeu, l'issue de chaque joueur doit dépendre des actions d'au moins un autre joueur. Dès lors que le gain obtenu par le proposant ne dépend uniquement que de ses propres actions, cette situation se rattache plutôt à la théorie de la décision, c'est-à-dire une situation de prise de décision individuelle. En dépit de cet aspect formel, le jeu du dictateur est utilisé dans la littérature de la théorie des jeux comme un jeu dégénéré.

Nous concluons par la section 4, en discutant de l'insuffisance des modèles avec préférences distributives pour expliquer pleinement les données expérimentales obtenues dans ces deux jeux de dilemme social. Notre principale conviction est qu'un modèle à préférences dépendantes des croyances décrit mieux et de manière plus subtile l'hétérogénéité des comportements qui est habituellement détectée dans ces deux jeux. En effet, si l'on admet que les croyances d'un joueur sont un argument de sa fonction d'utilité, alors la corrélation entre le choix du dictateur (du receveur, dans le jeu de la confiance) et ses croyances de second ordre peut être expliquée. *A contrario*, la théorie des jeux standard ainsi que ses extensions avec des préférences distributives sont incapables de prédire de telles corrélations qui sont souvent détectées dans le jeu du dictateur et dans le jeu de la confiance.

1. Théories des préférences distributives et dépendantes des croyances

Comme nous le verrons dans la section 3 de cet article, deux types de préférences sociales semblent prépondérants dans les expériences concernant le jeu du dictateur et le jeu de la confiance : l'aversion à l'inégalité (modélisée comme une préférence distributive) et l'aversion à la culpabilité (modélisée comme une préférence dépendante des croyances). Nous débutons cette section en donnant une définition très générale de ces deux catégories de préférences ainsi que les intuitions qui se trouvent derrière ces deux concepts. Nous essayerons ensuite d'expliquer comment un même sentiment (par exemple, la réciprocité) peut appartenir à l'une ou l'autre famille de préférences selon la pertinence, ou la non pertinence, de caractéristiques spécifiques liées aux croyances (par exemple, les intentions). En effet, l'aversion à la culpabilité peut être modélisée comme une préférence distributive ou comme une préférence dépendante des croyances. Nous montrerons également comment certains modèles sont capables d'incorporer les deux types de préférences dans une unique forme fonctionnelle. Ceci suggère que la distinction entre les deux familles de modèle n'est pas si stricte qu'on pourrait le croire.

1.1 *Préférences distributives*

Une première extension théorique a été proposée en considérant des joueurs qui, plutôt que de s'intéresser exclusivement à leurs propres gains monétaires, se soucient également des gains monétaires de leur partenaire de jeu. On a notamment conjecturé que les sujets participants à des expériences économiques se soucient de la manière dont les gains peuvent être distribués entre les deux partenaires de jeu.

Le terme *préférences distributives* est utilisé par des économistes pour décrire un monde où les décideurs ont une véritable préoccupation pour le bien-être des autres, dans le sens où leur bien-être et leur comportement ne dépendent pas uniquement de leurs propres gains monétaires, mais également de ceux des autres agents.[2] Les économistes font également la distinction entre les différents archétypes

2. Cette préoccupation pour le bien-être des autres en termes monétaires distingue les modèles avec préférences distributives des modèles de type *soucis du regard des autres*, dans lesquels d'autres préoccupations telles que les intentions ou les espérances des autres agents font partie de la fonction d'utilité d'un décideur.

de préoccupations distributives selon la manière dont les gains monétaires sont intégrés dans la fonction d'utilité du décideur. Parmi les plus importants figurent l'altruisme et la maximisation du surplus, où les gains monétaires des autres jouent un rôle positif dans la fonction d'utilité du décideur. L'aversion à l'inégalité et les motivations égalitaires font partie des archétypes dans lesquels les gains de ceux qui ont un revenu plus faible sont pris en compte de manière positive dans la fonction d'utilité, alors que les gains de ceux qui ont un revenu plus élevé le sont de manière négative. On peut également citer les préférences malveillantes et les préoccupations concernant le revenu relatif (ce que je gagne par rapport aux autres) qui interviennent uniquement de manière négative dans la fonction d'utilité du décideur.

Puisque que les individus ayant des préférences distributives ne se préoccupent que des gains monétaires des autres joueurs, celles-ci peuvent être traitées avec les outils de la théorie des jeux standard. Il suffit en effet dans ce cas, de supposer que l'utilité d'un joueur dépend aussi du gain monétaire de son partenaire, et pas uniquement de ses propres gains. Dès lors, tous les concepts traditionnels de solutions d'équilibre et de non-équilibre peuvent être directement appliqués dans le modèle.

Le modèle le plus célèbre concernant les préférences distributives est celui de Fehr et Schmidt (1999), où l'aversion à l'inégalité qui pousse un joueur à minimiser les différences entre son gain monétaire et celui de son partenaire a été modélisée pour la première fois. Le modèle avec préférences distributives de Bolton et Ockenfels (2000) se focalise toujours sur ce même sentiment d'aversion à l'inégalité, mais est axé cette fois-ci sur les résultats. La différence fondamentale entre les modèles de Fehr et Schmidt (1999) et de Bolton et Ockenfels (2000) réside dans la motivation de l'aversion : le premier modèle suppose que les sujets abhorrent la différence entre leur gain monétaire et celui de tout autre individu, tandis que le second suppose qu'ils préfèrent que le gain monétaire moyen soit aussi proche que possible de leur propre gain.

Parmi les autres modèles avec préférences distributives figurent également celui d'Andreoni et Miller (2002) et celui de Cox *et al.* (2008). Ces deux modèles utilisent la théorie néoclassique des préférences pour modéliser la préférence pour l'altruisme. Andreoni et Miller (2002) font notamment remarquer que si l'altruisme est un choix délibéré, alors cet altruisme pourrait être en accord avec le principe néoclassique des préférences révélées (Hicks, 1939; Samuelson, 1947). En se basant sur le travail d'Andreoni et Miller (2002), Cox *et al.* (2008) élaborent une théorie de l'altruisme révélé comme une théorie non-paramétrique des préférences liées aux gains monétaires d'un individu et à ceux de son partenaire de jeu. Ils montrent comment l'altruisme peut faire partie d'un processus de prise de décision rationnelle en rajoutant deux axiomes supplémentaires à ceux de la théorie néoclassique des préférences, un axiome *plus généreux que* et un axiome *plus altruiste que.*

Le travail de Cox *et al.* (2008) peut être également vu comme une façon de modéliser la réciprocité comme une préférence distributive. Quand une personne se montre volontairement aimable envers quelqu'un de bienveillant avec elle, et

volontairement antipathique envers les personnes malveillantes, alors cette personne est soucieuse de la réciprocité. Dans Cox *et al.* (2008) cette préoccupation est formalisée à l'aide d'un jeu dynamique à deux étapes dans lequel un choix *plus généreux que* pris par le premier décideur peut conduire à un choix *plus altruiste que* par le deuxième décideur, les deux *plus généreux que* et *plus altruiste que* étant définis en termes de comparaisons de gains monétaires d'un joueur. Ce même raisonnement s'applique pour les choix *moins généreux qu*e et *moins altruiste que*.

Il existe d'autres modèles capables de tenir compte de la réciprocité et qui peuvent également être rattachés à la famille des préférences distributives, bien qu'ils utilisent une paramétrisation basée sur des concepts psychologiques qui ne sont pas directement liés à des préoccupations distributives courantes.

Par exemple, Cox *et al.* (2007) développent un modèle de préférences distributives conditionnelles où, contrairement aux modèles avec préférences distributives discutés plus haut, les préférences distributives d'un agent sont conditionnelles à l'*état relatif*, c'est-à-dire au comportement antérieur des autres et aux actions alternatives qu'ils auraient pu prendre. Cox *et al.* (2007) utilisent, au lieu des croyances, et contrairement aux modèles avec préférences dépendantes des croyances, des états émotionnels basés sur l'expérience : ma décision au sujet de vos gains dépend de mon état d'esprit. Ainsi, selon que je sois bien disposé ou revanchard, votre comportement effectif modifiera systématiquement mon état émotionnel. Cependant, malgré ces états émotionnels, leur modèle peut encore être considéré comme un prolongement direct des modèles avec préférences distributives discutés dans cette section. En effet, dans Cox *et al.* (2007) l'état émotionnel est une fonction de la variable *réciprocité* et de la variable état relatif. La variable réciprocité du deuxième décideur est la différence entre le gain maximal que celui-ci peut s'assurer étant donné le choix du premier décideur et un gain neutre, avec encore une fois une approche distributive. La variable état relatif réfère à une asymétrie statutaire généralement reconnue dans les créances ou obligations des joueurs, en raison par exemple, des différences d'âge, de sexe, de titre ou de l'effort des joueurs. Cox *et al.* (2007) examinent les changements dans les possibilités de gain quand un rôle avantageux dans le jeu est attribué au mérite plutôt que par le hasard. Dès lors qu'un rôle dans le jeu est avantageux ou non — quant aux gains potentiels comparativement aux autres rôles dans le jeu — alors le concept d'état relatif peut être attribué à une approche avec préférences distributives.

Notons que, comme nous l'avons évoqué dans l'introduction, l'aversion à la culpabilité peut être également modélisée comme une préférence distributive. Considérons, par exemple, le concept de la culpabilité due à la désapprobation (Baumeister *et al.,* 1994) : le joueur *i* désapprouve le joueur *j* lorsque le joueur *j* ne se comporte pas selon l'idéal moral du joueur *i*. Le joueur *j* croit, quant à lui, qu'il est désapprouvé par le joueur *i* quand il ne se comporte pas selon l'idée qu'il se fait de l'idéal moral du joueur *i*. L'aversion à la culpabilité due à la désapprobation peut être modélisée comme une préférence distributive en généralisant le modèle de Fehr et Schmidt (1999). Pour cela, il suffit de supposer que le joueur *j* est sensible

à la différence entre le gain qu'il donne effectivement au joueur *i* et le gain du joueur *i* que le joueur *j* croit être juste selon le joueur *i*, plutôt qu'à la différence de gains entre les deux joueurs.

Nous concluons cette section en discutant brièvement d'un modèle mixte qui, bien qu'incorporant des préférences distributives, permet également l'utilisation de préférences dépendantes des croyances. Nous nous référons au modèle de Charness et Rabin (2002), où l'utilité du joueur *j* est une somme pondérée de son propre gain monétaire et des gains du joueur *i* (préférence distributive). Dans ce modèle, le poids que place le joueur *j* sur les gains du joueur *i* peut dépendre du fait que le joueur *i* obtienne un gain supérieur ou inférieur au joueur *j* (préférence distributive) et du fait que le joueur *i* ait eu un comportement hostile ou non (préférence dépendante des croyances).

Du point de vue des préférences distributives, ce modèle étend le modèle de Fehr et Schmidt (1999) afin de prendre en compte plusieurs types de motivations fondées sur le résultat des gains. Charness et Rabin (2002) modélisent entre autres, les préférences *maxmin* d'un joueur. Ainsi, un joueur préfère augmenter le gain du joueur le moins bien nanti, ainsi que le souci d'efficience, c'est-à-dire qu'un joueur préfère augmenter le gain de tous les joueurs en portant une attention particulière aux joueurs ayant les gains les plus faibles.[3]

Du côté des préférences dépendantes des croyances, la bienveillance (ou malveillance) perçue est quant à elle modélisée comme dans Rabin (1993) où la réciprocité du joueur *j* est basée sur la manière dont *j* perçoit les intentions du joueur *i* quand *i* fait son choix. Avant de faire son choix, le joueur *j* se pose la question : « Est-ce que le joueur *i* s'est mal comporté en violant les préceptes des préférences de bien-être social? ». Dans la prochaine sous-section, nous montrerons que, lorsque la réciprocité est fondée sur les intentions, les préférences dépendantes des croyances sont incontournables dans la modélisation.

Bien que la catégorisation entre préférences distributives et préférences dépendantes des croyances ne puisse être parfaite, nous pensons en revanche qu'il est nécessaire pour des économistes théoriciens et comportementalistes de savoir quand la théorie des jeux standard peut encore être utilisée pour fournir des prédictions théoriques testables à l'aide d'expériences économiques et quand il faut faire appel à une extension de la théorie (théorie des jeux psychologiques). Ce point sera abordé en détails dans la prochaine sous-section.

3. Ce souci d'efficience se retrouve également dans les théories du raisonnement en équipe (Bacharach, 1999). Un joueur raisonnant en équipe agit pour l'intérêt de son groupe en identifiant un profil de stratégies qui maximise le profit collectif du groupe. Par conséquent, un joueur raisonnant en équipe considère les gains des autres joueurs comme faisant partie de sa fonction d'utilité qu'il doit maximiser.

1.2 *Préférences dépendantes des croyances*

Une extension de la théorie économique standard a été proposée dans la littérature en supposant que les joueurs pouvaient être également motivés par ce qui est parfois mentionné comme des utilités psychologiques. Ces utilités renvoient à des préférences qui sont dans une certaine mesure dépendantes des autres (préférences dépendantes des croyances), c'est-à-dire un joueur prend en compte les autres individus à travers ses propres croyances qu'il a sur les choix et les croyances des autres joueurs. Nous définissons comme *jeu psychologique* ou *jeu avec préférences dépendantes des croyances* une situation d'interaction stratégique dans laquelle au moins l'un des joueurs a des préférences dépendantes des croyances.

L'exemple le plus célèbre d'une application basée sur un jeu psychologique est celui de Rabin (1993), qui propose un modèle de la réciprocité fondée sur les intentions où les joueurs agiraient de manière bienveillante (respectivement hostile) en réponse à des actions bienveillantes (respectivement hostiles). La notion clé de la bienveillance dépend des croyances. Pour en comprendre la raison, considérons l'exemple suivant, qui vise également à clarifier le concept d'intention.[4] Supposons que je saute devant votre voiture pour bloquer votre passage, de sorte que vous ne puissiez pas traverser un pont, et qu'en conséquence vous arriviez en retard à une réunion importante. Suis-je bienveillant à votre égard? De toute évidence on ne peut pas l'affirmer sans connaître mes croyances. Si je crois que le pont est aussi robuste que les ponts le sont en général et que je fais juste l'imbécile, alors je suis juste mal intentionné. Cependant, si je crois que le pont est sur le point de s'effondrer, alors je suis bienveillant et je le serais aussi probablement même si je me suis trompé en prenant un pont solide pour un pont dangereux. Par conséquent, devriez-vous vous montrer aimable ou mécontent en retour? La réponse à cette question dépend de votre croyance à propos de ma bienveillance, et donc de vos croyances au sujet de mes croyances. Pour modéliser ce dilemme, il faut avoir recours aux préférences dépendantes des croyances.

Le modèle de Rabin (1993) s'applique aux jeux psychologiques simultanés, interactions stratégiques où les joueurs choisissent une seule fois, sans observer, avant de choisir, le choix des autres joueurs. Dufwenberg et Kirchsteiger (2004) ainsi que Falk et Fischbacher (2006) étendent les notions de bienveillance et de réciprocité fondées sur les intentions aux jeux dynamiques, interactions stratégiques où un joueur peut observer le choix d'un autre joueur avant de faire son choix.

Le modèle de Dufwenberg et Kirchsteiger (2004) est capable de saisir la manière dont émergent la perception de la bienveillance d'un joueur et sa réciprocité fondée sur les intentions dans les différents nœuds de décisions de l'arbre du jeu. En effet, suivant la façon dont se dénoue l'arbre dans un jeu dynamique, un joueur qui révise ses croyances peut avoir à réviser également ses croyances sur la bienveillance des autres joueurs, si sa bienveillance dépend de ses croyances. En conséquence, la manière dont le joueur est affecté par des préoccupations de réciprocité fondée sur les intentions peut considérablement différer entre les différents nœuds de l'arbre.

4. Cet exemple est tiré de Dufwenberg (2008).

Falk et Fischbacher (2006) proposent quant à eux une extension différente de Rabin (1993) qui utilise une fonction d'utilité avec une préférence distributive. Leur modèle reste un modèle dépendant des croyances puisque la perception de la bienveillance d'une action dépend de l'issue de cette action ainsi que de l'intention sous-jacente. Dans leur modèle, le résultat est mesuré avec le terme Δ_j, où $\Delta_j > 0$ exprime un résultat en faveur du joueur *i* et $\Delta_j < 0$ exprime un résultat en défaveur du joueur *i*. Afin de déterminer la bienveillance du joueur *j*, Δ_j est multiplié par le facteur d'intention θ_j. Ce facteur est un nombre compris entre 0 et 1, où $\theta_j = 1$ prend en considération une situation où Δ_j est le résultat d'une action que le joueur *j* fait de manière complètement intentionnelle, et $\theta_j < 1$ traduit une action du joueur *j* qui ne l'est pas. Ces intentions sont mesurées au moyen des croyances de premier ordre du joueur, et la perception de ses intentions par l'autre joueur par ses croyances de second ordre. Le terme de bienveillance est simplement le produit de Δ_j et de θ_j. Ainsi, la réciprocité est encore modélisée comme une préférence dépendante des croyances.

La réciprocité n'est qu'une des motivations pouvant être modélisées par le biais des jeux psychologiques. Beaucoup d'autres émotions peuvent également l'être. Le spectre des notions explorées dans ces modèles avec préférences dépendantes des croyances est donc très vaste. Caplin et Leahy (2004) ont par exemple proposé un modèle où un médecin doit décider de divulguer ou non des informations médicales à son patient en n'ayant aucune information quant à la propension à l'anxiété du malade. De même, Tadelis (2011) analyse de manière théorique un jeu de la confiance où le receveur, la personne à qui on doit accorder notre confiance, est sensible à la honte. Battigalli *et al.* (2015) développent un modèle avec préférences dépendantes des croyances pour explorer comment la frustration et la colère, par le biais du blâme et l'agression, façonnent les interactions et les résultats dans la sphère économique. Ils montrent en quoi les conséquences économiques de la colère peuvent être déterminantes sur, par exemple, la fixation des prix, la sécurité routière, la violence ou la politique.

Dufwenberg (2002) fut le premier à modéliser l'aversion à la culpabilité comme une préférence dépendante des croyances avec une culpabilité due à la déception. Cette notion fut successivement développée par Battigalli et Dufwenberg (2007, 2009). L'aversion à la culpabilité peut être définie comme suit (Tangney, 1995) : « Les gens souffrent de la culpabilité et ont des remords s'ils causent du tort à d'autres ». Bien que la culpabilité puisse revêtir différentes formes, la manière la plus marquante d'infliger du tort à une personne est de décevoir cette personne. Un joueur averse à la culpabilité cherchera donc avant tout à ne pas décevoir son partenaire de jeu (voir la définition de Baumeister *et al.* 1994 dans l'introduction). Dufwenberg (2002) et Battigalli et Dufwenberg (2007, 2009) s'intéressent donc à une aversion de la culpabilité dépendante des croyances, c'est-à-dire sur ce qu'éprouve un joueur s'il a échoué à se montrer à la hauteur des attentes de ses partenaires, à savoir s'il les a déçus (souvenons-nous de l'exemple de Karen et Jim dans l'introduction).

Enfin, Attanasi *et al.* (2013) étudient dans un jeu de la confiance le comportement des deuxièmes décideurs (receveurs) qui peuvent être motivés à la fois par l'aversion à la culpabilité et par la réciprocité fondées sur les intentions. Ils montrent que lorsque la sensibilité à la culpabilité du deuxième décideur est égale à sa sensibilité à la réciprocité fondée sur les intentions, elle peut alors être appréhendée comme une aversion à l'inégalité du joueur. Ce résultat constitue une preuve supplémentaire que la classification entre préférences distributives et préférences dépendantes des croyances n'est pas si catégorique, comme nous l'avons évoqué dans l'introduction de cette section.

Cependant, les modèles de la théorie des jeux standard ne fournissent pas d'outils suffisants pour décrire de manière adéquate les préférences dépendantes des croyances. En effet, ces modèles, ou de manière plus générale l'approche traditionnelle, supposent que les utilités dépendent uniquement des actions choisies par les joueurs. *A contrario*, quand les joueurs sont motivés par les sentiments décrits précédemment, leurs utilités peuvent aussi dépendre directement des croyances qu'ils ont sur les choix, les informations ou les croyances des autres joueurs. Ainsi, dès lors qu'il s'agit de sentiments fondés sur les intentions, nous devons aller au-delà de la théorie des jeux standard et nous tourner vers la théorie des jeux psychologiques.[5] Ce nouveau cadre de travail repose sur un fondement stratégique où au moins un joueur a des préférences dépendantes des croyances ou croit, avec une certaine probabilité, que l'un de ses partenaires dans le jeu a des préférences dépendantes des croyances. Cette théorie peut être considérée comme une généralisation de la théorie des jeux standard.

Geanakoplos *et al.* (1989) furent les premiers, dans leur article fondateur, à montrer l'inadéquation des méthodes traditionnelles pour représenter les préférences dépendantes des croyances, et à proposer des extensions à la théorie des jeux standard pour mieux étudier ces préférences. Cependant, la boîte à outils proposée par Geanakoplos *et al.* (1989) dans le cadre des jeux psychologiques contient plusieurs restrictions qui excluent de nombreuses formes plausibles de préférences dépendantes des croyances. En effet, leur modèle ne prend en compte dans la fonction d'utilité de l'agent que des croyances initiales et figées dans le temps, alors que bien d'autres formes importantes de préférences dépendantes des croyances requièrent l'introduction de croyances endogènes qui doivent être réactualisées de manière dynamique. Battigalli et Dufwenberg (2009) ont par la suite généralisé et étendu le travail de Geanakoplos *et al.* (1989) en permettant la prise en compte des mises à jour des croyances, des croyances des autres, des stratégies planifiées, mais également de l'information incomplète ayant un impact sur les motivations. Parmi les autres avancées, Battigalli et Dufwenberg (2009) examinent notamment comment un joueur révise ses croyances sur les croyances des autres joueurs après avoir observé leurs choix. Ils sont donc capables de

5. Comme l'a justement relevé Dufwenberg (2008), « L'expression *jeu avec motivations dépendantes des croyances* serait plus parlante que l'expression *jeu psychologique*, mais la seconde est plus communément utilisée dans la littérature. » (p. 715).

modéliser les effets psychologiques dynamiques qui sont normalement exclus du champ d'analyse quand certains types épistémiques sont identifiés uniquement avec des croyances hiérarchiques initiales figées dès le départ. Ils définissent également la notion d'*équilibre séquentiel psychologique* qui généralise la notion d'équilibre séquentiel des jeux traditionnels dont ils prouvent l'existence avec des hypothèses faibles. C'est cette notion d'équilibre à laquelle nous nous référerons dans cet article quand nous analyserons les jeux avec préférences dépendantes des croyances dans les sections 2.1.2 et 2.2.2.

2. Jeux de dilemme social avec préférences sociales

Nous comparons dans cette section les prédictions théoriques d'un modèle spécifique de préférences distributives – aversion à l'inégalité à la Fehr et Schmidt (1999) – avec celles d'un modèle spécifique de préférences dépendantes des croyances – aversion à la culpabilité à la Battigalli et Dufwenberg (2009) – pour deux jeux très simples et largement analysés dans la littérature expérimentale sur les préférences sociales : le jeu du dictateur et le jeu de la confiance. Nous montrons que différents types de préférences sociales conduisent à des prédictions théoriques différentes pour une même situation stratégique. En outre, et plus important encore, nous voulons montrer comment un jeu avec préférences distributives peut être analysé en utilisant les mêmes outils que ceux de la théorie des jeux standard, ce qui n'est pas le cas pour les jeux avec préférences dépendantes des croyances. Ces derniers nécessitent une généralisation du modèle standard afin d'inclure, également de manière directe, les croyances dans les fonctions d'utilité des joueurs.

2.1 *Jeu du dictateur avec préférences sociales*

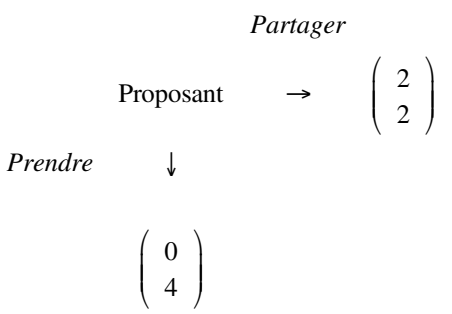
Nous examinons un jeu très simple à une seule étape et qui représente la situation économique d'interaction stratégique suivante. Le premier joueur, le proposant, décide seul de la répartition d'une dotation de 4€. Dans la version simplifiée de ce jeu que nous analysons ici, le mini-jeu du dictateur,[6] le proposant n'a que deux choix possibles : s'approprier la totalité de la dotation ou la partager à parts égales avec le deuxième joueur, le receveur. Comme dans le jeu du dictateur standard, ce dernier reçoit tout simplement le reste de la dotation laissée par le proposant. Le rôle du receveur est donc totalement passif, il n'a pas d'influence stratégique sur l'issue du jeu.

L'arbre du jeu avec les gains monétaires est représenté dans le graphique 1. Les gains sont exprimés en euros et ne reflètent pas nécessairement les préférences des joueurs. A chaque extrémité de l'arbre du jeu, le premier gain se réfère au receveur et le deuxième se réfère au proposant.

6. C'est une version simplifiée du jeu du dictateur appelée le mini-jeu du dictateur en référence au mini-jeu de l'ultimatum de Binmore *et al.* (1995), une version de choix binaire du jeu de l'ultimatum.

GRAPHIQUE 1

JEU DU DICTATEUR AVEC GAINS MONÉTAIRES



Nous savons d'après les prédictions de la théorie des jeux standard que l'unique équilibre de Nash du mini-jeu du dictateur du graphique 1 est que le proposant choisit *Prendre*, étant donné que cette action lui procure un gain monétaire plus grand que l'option *Partager.* Cependant, de nombreux résultats expérimentaux (voir section 3) ont montré qu'une part non négligeable des sujets ayant le rôle de proposant dans les expériences du jeu du dictateur ne choisissent pas l'action qui maximise leur gain monétaire, mais retiennent plutôt les allocations qui donnent au receveur un gain monétaire positif non nul.

2.1.1 *Jeu du dictateur avec aversion à l'inégalité du proposant*

Une explication possible de ce fait stylisé est que le proposant a des préférences distributives à la Fehr et Schmidt (1999). Tel qu'indiqué dans la section précédente, ce type de préférences peut être traité à l'aide de la théorie des jeux standard. Plus précisément, supposons que le proposant soit averse à l'inégalité : il préfère minimiser l'écart entre son gain et le gain du receveur. Formellement, ses préférences sont représentées par la fonction d'utilité

$$u_P(s_P) = \pi_P(s_P) - h\max\{0, \pi_R(s_P) - \pi_P(s_P)\} - k\max\{0, \pi_P(s_P) - \pi_R(s_P)\} \quad (1)$$

où s_P est la stratégie du proposant (le seul joueur actif dans le jeu), π_P le gain monétaire du proposant, π_R le gain monétaire du receveur et h, k des paramètres positifs tels que $h \in [k, +\infty)$ et $k \in [0, 1)$.

Examinons à présent chaque hypothèse sous-jacente au modèle de préférences sociales avec aversion à l'inégalité :

- La forme fonctionnelle de (1) repose sur l'hypothèse que la fonction d'utilité est linéaire dans les gains monétaires ainsi que dans l'aversion à l'inégalité. Cela implique que le taux marginal de substitution entre le gain monétaire et l'inégalité est constant. Cette hypothèse, qui peut paraître irréaliste à première vue, est pourtant en adéquation avec de nombreuses observations recueillies dans les expériences sur le jeu du dictateur, comme nous le verrons dans la section 3.

- Le montant $\pi_R(s_P) - \pi_P(s_P)$ représente l'inégalité qui est au détriment du proposant alors que $\pi_P(s_P) - \pi_R(s_P)$ représente celle qui est à son bénéfice. L'hypothèse $h \in [k, +\infty)$ capture l'idée qu'un joueur souffre plus d'une inégalité qui serait à son désavantage, la sensibilité à l'inégalité à son désavantage ne pouvant être inférieure à celle qui serait à son avantage.[7] La contrainte $h \in [k, +\infty)$ signifie également qu'un joueur a une aversion envers les pertes (à la Kahneman et Tversky) dans la comparaison sociale, c'est-à-dire que les déviations négatives par rapport au revenu de référence importent plus que les déviations positives.[8]

- La restriction $k \geq 0$ exclut la possibilité qu'un joueur préfère être dans une situation meilleure que les autres. C'est une hypothèse très restrictive, puisque nous savons que dans les expériences économiques, il peut y avoir des sujets qui se réjouissent de sortir du laboratoire avec un gain plus élevé que les autres participants. En admettant ce fait, Fehr et Schmidt (1999) justifient cette hypothèse par le fait que $k < 0$ n'a pas d'incidence sur le comportement à l'équilibre, c'est-à-dire que de tels sujets se comporteraient comme des joueurs qui ne seraient guidés que par leur propre intérêt. Par conséquent, $k < 0$ ne peut pas rationaliser les écarts de comportement par rapport à l'équilibre de Nash dans les jeux de dilemme social comme ceux analysés dans cet article.

- La restriction $k < 1$ a besoin d'une discussion plus approfondie. Supposons que le proposant obtienne un gain monétaire supérieur à celui du receveur (lorsqu'il choisit *Prendre* dans le jeu au graphique 1). Dans ce cas, $k = 1/2$ impliquerait que le proposant soit indifférent aux choix entre prendre 1€ pour lui-même et donner 1€ au receveur; $k = 1$ signifierait plutôt que le proposant soit prêt à renoncer à 1€ afin de réduire son avantage monétaire par rapport au receveur. Cela semble peu plausible et un $k > 1$ serait encore moins réaliste.

- Il n'y a aucune raison empirique d'imposer une limite supérieure à h. Supposons que le proposant obtienne un gain monétaire inférieur à celui du receveur. Dans ce cas, si $h \geq 1$, alors le proposant est prêt à renoncer à 1€ sur son propre gain monétaire si cela permet de réduire le gain de son adversaire d'au moins 1€. Les résultats expérimentaux montrent que les personnes ayant de telles préférences existent, même si elles sont très rares.

Calculons à présent l'équilibre de Nash du mini-jeu du dictateur du graphique 1, sous l'hypothèse que le dictateur soit averse à l'inégalité avec une fonction d'utilité

7. Voir Loewenstein *et al.* (1989) pour la preuve qui corrobore cette hypothèse.

8. Il existe une abondante littérature montrant la pertinence de l'aversion aux pertes dans d'autres domaines que les loteries à risque (Tversky et Kahneman, 1991). Fehr et Schmidt (1999) supposent quant à eux que l'aversion aux pertes affecte également les comparaisons sociales.

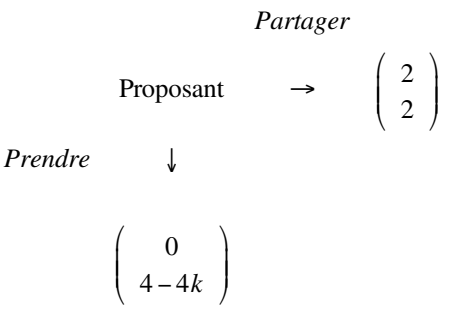
représentée par (1), et que cette information soit de connaissance commune pour les deux joueurs.[9]

Notons que compte tenu de la structure de gains du jeu au graphique 1, il est impossible d'avoir $\pi_R(s_P) > \pi_P(s_P)$. La fonction d'utilité de (1) peut donc se simplifier et nous obtenons alors $u_P(s_P) = \pi_P(s_P) - k\max\{0, \pi_P(s_P) - \pi_R(s_P)\}$, et $u_P(Prendre) = 4 - 4k$ et $u_P(Partager) = 2$.

Le jeu représentant le mini-jeu du dictateur avec une aversion à l'inégalité du proposant est illustrée au graphique 2. Ce qui apparaît à l'extrémité de l'arbre de décision doit être considéré comme des utilités et non comme des gains monétaires, bien que les deux notions coïncident pour l'une des deux extrémités de l'arbre.

GRAPHIQUE 2

JEU DU DICTATEUR AVEC AVERSION À L'INÉGALITÉ DU PROPOSANT



Le proposant choisit *Prendre* (même équilibre que dans le cas où il agit uniquement pour son propre intérêt) si $k \in (0, 1/2)$. Il choisit par contre *Partager* si $k \in (1/2, 1)$, et il est indifférent entre *Prendre* et *Partager* si $k = 1/2$. Soit $\alpha_R := \Pr_R[Partager]$ la croyance de premier ordre du receveur que le proposant va choisir *Partager*. Dès lors que les croyances sont exactes à l'équilibre, la croyance de premier ordre α_R peut être interprétée comme la stratégie mixte du proposant. Donc, si $k = 1/2$, en plus des deux équilibres de Nash en stratégies pures, il existe un continuum d'équilibres de Nash en stratégies mixtes, où le proposant choisit *Partager* avec probabilité $\alpha_R \in (0, 1)$.

Nous résumons dans le tableau 1 les équilibres de Nash pour le jeu du graphique 2, pour toutes les valeurs positives du paramètre k.

---

9. Soulignons que l'hypothèse de connaissance commune n'est pas nécessaire dans ce jeu, puisque l'autre joueur (le receveur) n'est pas actif et ses croyances au sujet de l'action choisie par le proposant (croyances de premier ordre) n'entrent pas dans la fonction d'utilité de ce dernier. Par conséquent, même si le receveur ne connaissait pas h ni k (c.-à-d. que le proposant soit averse à l'inégalité), nous trouverions les mêmes prédictions d'équilibre pour le mini-jeu du dictateur avec une aversion à l'inégalité du proposant.

TABLEAU 1

ÉQUILIBRE DE NASH DU MINI-JEU DU DICTATEUR AVEC AVERSION À L'INÉGALITÉ DU PROPOSANT

Sensibilité du proposant à l'inégalité au bénéfice du proposant	**Équilibre Pur *Prendre***	**Équilibre Pur *Partager***	**Équilibre Mixte**
$k \in (0, 1/2)$	Oui	Non	Non
$k = 1/2$	Oui	Oui	Oui, $\alpha_R \in (0, 1)$
$k \in (1/2, 1)$	Non	Oui	Non

Le choix du proposant à l'équilibre va dépendre de sa sensibilité k pour l'inégalité qui est en sa faveur. Pour k supérieur à un seuil donné, il fait un choix (en l'occurrence *Partager*) qui est différent de celui qu'il ferait s'il n'était guidé que par son propre intérêt. Notons que le choix optimal du proposant ne dépend pas de ce qu'il pense du receveur ni de ce que celui-ci s'attend à ce qu'il fasse. Par exemple, il est possible que le receveur ne se préoccupe pas du tout du fait que le proposant obtiendrait un gain plus élevé que le sien, alors que le proposant, *a contrario*, s'en soucie et choisit *Partager.*

2.1.2 *Jeu du dictateur avec aversion à la culpabilité du proposant*

Nous supposons à présent que le proposant est averse à la culpabilité, cette aversion étant modélisée selon des préférences dépendantes des croyances à la Battigalli et Dufwenberg (2009). Comme nous l'avons évoqué précédemment, les gens souffrent de la culpabilité s'ils causent du tort à d'autres, et la manière la plus marquante d'infliger du tort à une personne est de la décevoir (Tangney, 1995). Battigalli et Dufwenberg (2009) développent une théorie générale de l'aversion à la culpabilité et montrent comment calculer les équilibres séquentiels pour les jeux psychologiques. Nous ne développerons pas ici leur modèle. Nous nous contentons d'appliquer leurs concepts théoriques.

Si le proposant est motivé par l'aversion à la culpabilité, il se préoccupe de la déception du receveur lorsque le gain monétaire que celui-ci s'attend à recevoir ne correspond pas au montant effectivement perçu. Le gain monétaire que le receveur s'attend à obtenir dépend de ses croyances de premier ordre sur la stratégie du proposant. Ces croyances sont définies comme les anticipations du receveur sur la stratégie que le proposant va choisir. De même, ce sont les croyances de second ordre du proposant sur sa propre stratégie, c'est-à-dire ses anticipations sur les anticipations du receveur par rapport à sa stratégie, qui permettent de définir son aversion à la culpabilité. En d'autres termes, les croyances de second ordre du

proposant sur sa propre stratégie *Partager* mesurent dans quelle mesure il croit que le receveur pense qu'il va choisir *Partager.*

Analysons la situation d'un point de vue formel. Soit α_R : = $\Pr_R$[*Partager*] la croyance de premier ordre du receveur que le proposant choisira *Partager.* α_R mesure la confiance du receveur envers le proposant avant que celui-ci ne prenne sa décision. Définissons également $\beta_P := E_P[\alpha_R]$ comme la croyance de second ordre du proposant avant que celui-ci ne fasse son choix. β_P mesure l'anticipation du proposant quant à la confiance du receveur sur le fait qu'il choisirait *Partager.*[10] En effet, plus β_P est grand, plus le proposant anticipe que le receveur pense que le proposant va choisir *Partager.*

Comment quantifier la déception du receveur d'avoir été trahi après que le proposant ait choisi *Prendre*? Le montant de cette déception est égal à $-2\alpha_R$, c'est-à-dire la différence entre le gain que celui-ci reçoit après la décision *Prendre*, qui est égal à zéro, et le gain moyen qu'il espérait percevoir, étant donné sa croyance de premier ordre α_R. Compte tenu de la notation que nous avons adoptée, nous pouvons alors définir l'espérance de gain monétaire du receveur par $E_{\alpha_R}[\pi_R] = 2 \cdot \alpha_R + 0 \cdot (1 - \alpha_R) = 2\alpha_R$.

La culpabilité du proposant après avoir choisi *Prendre* est donnée par l'anticipation de la déception du receveur, c'est-à-dire l'anticipation du proposant de $(-2\alpha_R)$ qui est égale à $E_P[-2\alpha_R] = -2E_P[\alpha_R] = -2\beta_P$. Une hypothèse naturelle est que, conformément à Fehr et Schmidt (1999) et d'autres modèles avec préférences sociales, la sensibilité à un sentiment particulier est hétérogène dans la population, c'est-à-dire qu'elle varie selon les sujets. Dès lors, représentons la sensibilité à la culpabilité du proposant par $\theta \geq 0$. La culpabilité du proposant, égale à $-2\beta_P$, multipliée par sa sensibilité à la culpabilité, représente l'utilité psychologique du proposant quand il choisit *Prendre*. L'utilité totale du proposant après avoir choisi *Prendre* est donnée par $4 - 2\theta\beta_P$, c'est-à-dire la somme de son gain monétaire et de son utilité psychologique.

Le jeu psychologique représentant le mini-jeu du dictateur avec aversion à la culpabilité du proposant est illustré au graphique 3. Ce qui apparaît à l'extrémité de l'arbre de décision doit être considéré comme des utilités et non comme des gains monétaires, bien que les deux notions coïncident pour l'une des deux extrémités de l'arbre.

Essayons de résoudre à présent le jeu avec préférences dépendantes des croyances (jeu psychologique) du graphique 3. Supposons que θ soit de connaissance commune pour les deux joueurs.[11]

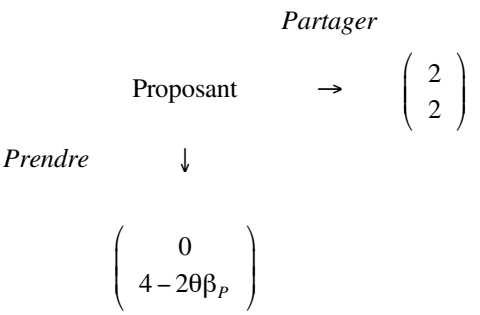
Bien que la théorie des jeux psychologiques puisse être vue comme une extension de la théorie des jeux standard, les concepts d'équilibre utilisés pour résoudre les

10. Nous utiliserons la lettre grecque *alpha* lorsque nous nous référons aux croyances de premier ordre et la lettre *beta* pour les croyances de second ordre.

11. La même hypothèse a été faite dans le cas d'aversion à l'inégalité.

GRAPHIQUE 3

JEU DU DICTATEUR AVEC AVERSION À LA CULPABILITÉ DU PROPOSANT



jeux psychologiques ne sont pas différents de ceux utilisés dans la théorie des jeux standard. S'appuyant sur Geanakoplos *et al*. (1989), Battigalli et Dufwenberg (2009) appliquent simplement la même logique d'équilibre de Nash pour résoudre les jeux psychologiques simultanés, ainsi que les traditionnels équilibres séquentiels, en utilisant le raisonnement habituel de l'induction à rebours, pour résoudre les jeux psychologiques dynamiques. Battigalli et Dufwenberg (2009) sont en quelque sorte obligés de généraliser les concepts précédents afin de tenir compte du fait que dans les jeux psychologiques, les croyances font partie intégrante de la fonction d'utilité des joueurs. Cela donne lieu à au moins deux questions techniques qui doivent être débattues lorsqu'il s'agit de comportement à l'équilibre (exactitude des utilités déclarées et révision des utilités déclarées). Nous discuterons ici de la première et aborderons la seconde à la section 2.2.2, lorsque nous analyserons le mini-jeu de la confiance avec préférences dépendantes des croyances.

La condition de croyances exactes et cohérentes, caractérisant l'équilibre de Nash dans les jeux avec uniquement des gains monétaires, lorsqu'elle est appliquée à des jeux avec préférences dépendantes des croyances, requiert également l'exactitude des utilités déclarées. En effet, étant donné l'implication des différents ordres de croyances dans l'analyse, il faut imposer de manière explicite que ces croyances soient correctes à l'équilibre (par exemple, dans le jeu au graphique 3, nous devons avoir à l'équilibre $\alpha_R = \beta_P$). Néanmoins, lorsque nous vérifions que les joueurs maximisent à l'équilibre leurs utilités totales à tous les nœuds de décision, étant donné leurs croyances exactes sur les actions de chacun des autres joueurs, il faut tenir compte du fait que certaines de ces croyances (indépendamment de leur ordre) font partie des utilités des ou de certains joueurs. Ceci représente l'impact psychologique direct des croyances sur les préférences à l'extrémité de l'arbre de décision. *A contrario*, dans la théorie des jeux standard, les croyances concernant les partenaires de jeu n'ont qu'un impact *indirect* sur les préférences à travers les pondérations subjectives des utilités des différentes actions du joueur dans son utilité espérée. Ainsi, pour tout équilibre d'un jeu avec préférences dépendantes des croyances, les joueurs maximisent leurs utilités totales qui sont calculées selon les

croyances exactes qu'ils possèdent sur les actions (croyances de premier ordre) et sur les croyances (croyances de second ordre) de leurs partenaires. Simultanément, les croyances (exactes) de différents ordres s'ajustent aux meilleures réponses des joueurs, calculées selon les utilités totales déclarées. Cette problématique fut abordée dans Geanakoplos *et al.* (1989) et a été généralisée ensuite par Battigalli et Dufwenberg (2009) qui ont permis d'intégrer simultanément les propres croyances d'un joueur ainsi que les croyances des autres joueurs dans la fonction d'utilité totale du joueur.

Après ces explications introductives et méthodologiques, nous pouvons à présent déterminer l'ensemble des équilibres de Nash pour le jeu psychologique du graphique 3, en fonction de toutes les valeurs possibles du paramètre θ (sensibilité à la culpabilité) positif. Pour cela, nous allons nous focaliser sur les valeurs des croyances fondamentales α_R, β_P à l'équilibre que nous avons définies auparavant. Dès lors que les croyances sont exactes à l'équilibre, la croyance de premier ordre α_R peut être interprétée comme la stratégie mixte du proposant. En outre, la condition de l'exactitude des croyances impose qu'on ait également à l'équilibre $\alpha_R = \beta_P$.

Nous résumons dans le tableau 2 les équilibres de Nash du jeu psychologique du graphique 3, pour toutes les valeurs positives du paramètre θ. La détermination des équilibres ainsi que l'explication de la méthodologie utilisée sont disponibles dans l'annexe 1.

TABLEAU 2

ÉQUILIBRE DE NASH PSYCHOLOGIQUE DU MINI-JEU DU DICTATEUR AVEC AVERSION À LA CULPABILITÉ DU PROPOSANT

Sensibilité du proposant à la culpabilité	**Équilibre Pur *Prendre***	**Équilibre Pur *Partager***	**Équilibre Mixte**
$\theta \in (0, 1)$	Oui	Non	Non
$\theta = 1$	Oui	Oui	Non
$\theta \in (1, +\infty)$	Oui	Oui	Oui, $\alpha_R = 1/\theta$

Notons que *Prendre* est toujours un équilibre (pour chaque valeur de θ) et il est unique si la sensibilité à la culpabilité est suffisamment faible ($\theta < 1$). Si le proposant est suffisamment sensible à la culpabilité ($\theta \geq 1$) il y aura une multiplicité d'équilibres. La raison est que, même si le proposant est extrêmement averse à la culpabilité, nous ne pouvons pas exclure la possibilité que celui-ci pense que le receveur anticipe que le proposant va choisir *Partager* avec une probabilité égale à 0. Nous avons dans ce cas $\beta_P = 0$ et l'utilité totale de *Prendre*, c'est-à-dire

$u_P(Prendre; \beta_P)$, se réduit à $u_P(Prendre; 0) = 4$ comme dans le cas d'un proposant égoïste. En revanche, si le proposant pense que le receveur anticipe qu'il va choisir *Partager* avec une probabilité égale à 1, l'équilibre sera *Partager* ($\alpha_R = \beta_P = 1$). Dans le cas où la sensibilité à l'aversion est suffisamment élevée ($\theta > 1$), il existe également un équilibre mixte où le proposant va choisir *Partager* avec une probabilité de $1/\theta < 1$. Notons que plus θ est élevé, plus la probabilité que le proposant choisisse *Partager* pour cet équilibre mixte sera faible. De même, plus θ est élevé, plus les anticipations du receveur concernant *Partager* seront faibles. Ce résultat, qui à première vue semble contre-intuitif, est dû au fait que l'utilité totale de *Prendre* décroît soit avec θ, soit avec β_P, alors que l'utilité totale de *Partager* reste inchangée. Enfin, pour satisfaire la condition nécessaire $u_P(Prendre; \beta_P) = u_P(Partager)$ afin qu'un équilibre mixte puisse exister, β_P doit décroître et, étant donné que $\alpha_R = \beta_P$ à l'équilibre, les croyances de premier ordre du receveur doivent également décroître.

Le tableau 2 montre donc qu'il peut y avoir des équilibres multiples (pour $\theta \geq 1$), même s'il n'existe qu'un seul joueur actif dans un jeu générique avec information parfaite.[12] Cela n'est pas possible dans la théorie des jeux standard lorsque nous sommes en présence de jeux à information complète.

2.2 *Jeu de la confiance avec préférences sociales*

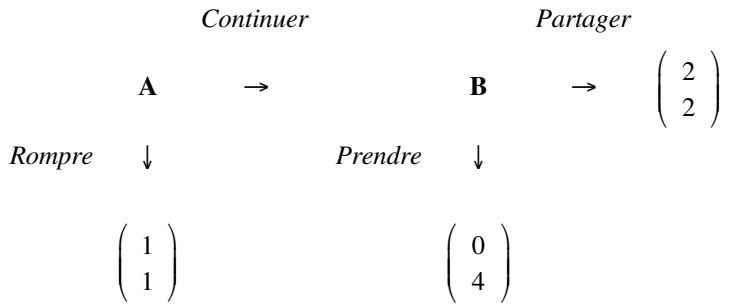
Comme annoncé précédemment, nous nous intéressons à un simple jeu à deux étapes qui représente la situation économique d'interaction stratégique suivante. Les joueurs A et B sont partenaires sur un projet qui a, jusqu'à présent, procuré des profits d'un montant total de 2€. Dans la version simplifiée du jeu que nous analysons dans cet article, le mini-jeu de la confiance,[13] le joueur A doit décider de se retirer ou non du projet. Si le joueur A décide de rompre son partenariat, les termes du contrat obligent les deux joueurs à se partager le profit en deux parts égales. Si le joueur A laisse ses ressources dans le projet, le profit total du projet va fructifier pour atteindre 4€. Toutefois, dans ce cas et selon les termes du contrat, le joueur B a le droit de partager ou non les profits, après l'accomplissement du projet. Ainsi, le joueur A doit décider s'il faut rompre ou continuer la collaboration sans savoir s'il y aura un partage de profit en cas de poursuite du partenariat. Après avoir pris connaissance de la décision du joueur A, et seulement si celui-ci a décidé de poursuivre la collaboration, le joueur B décide s'il va prendre et s'accaparer la totalité du profit fructifié, ou le partager. L'arbre du jeu avec les gains monétaires est représenté dans le graphique 4. Comme dans le graphique 1, les paiements sont exprimés en euros et ne représentent pas nécessairement les préférences des joueurs. À chaque extrémité de l'arbre du jeu, le premier gain se réfère au joueur A, et le deuxième se réfère au joueur B.

12. Un jeu avec information parfaite est générique si pour chaque joueur tous les gains aux extrémités de l'arbre du jeu sont différents entre eux.

13. Voir Attanasi *et al.* (2016) pour cette définition.

GRAPHIQUE 4

JEU DE LA CONFIANCE AVEC GAINS MONÉTAIRES



La théorie des jeux standard nous dit que l'unique équilibre parfait en sous-jeux du mini-jeu de la confiance du graphique 4 est celui où le joueur *A* choisit *Rompre* et le joueur *B* choisit *Prendre* si A a choisi *Continuer.*

Supposons maintenant que ce jeu soit joué en laboratoire par paires de sujets. Selon la littérature expérimentale sur ce sujet, on peut raisonnablement s'attendre à un pourcentage non négligeable de paires avec un profil d'actions (*Continuer, Partager*).

La raison de cette déviation par rapport aux prédictions théoriques pourrait s'expliquer par la rationalité limitée de certains joueurs A ou de certains joueurs B, ou les deux. Cependant, il semble difficile d'affirmer que dans un jeu aussi simple, un nombre non négligeable de joueurs soient incapables de comprendre les règles du jeu, ou de calculer les actions qui maximiseraient leurs gains étant donné leurs anticipations (c'est-à-dire leurs croyances de premier ordre) quant aux choix de leurs partenaires de jeu.[14]

Une autre raison serait que les joueurs B soient non seulement motivés par leur propre intérêt mais également, du moins certains d'entre eux, par des préférences sociales. C'est ce que nous supposons dans la suite de cette section.

2.2.1 *Jeu de la confiance avec aversion à l'inégalité du receveur*

Supposons que le receveur (B) ait des préférences distributives à la Fehr et Schmidt (1999). Plus précisément, supposons que, alors que A est un joueur égoïste (c'est-à-dire qu'il maximise uniquement son gain monétaire espéré), le joueur B

14. L'une des explications plausibles de ces résultats empiriques et qui diffèrent partiellement de l'équilibre parfait en sous-jeux serait la non-justesse des croyances de certains joueurs. Supposons par exemple que A et B soient tous les deux égoïstes et rationnels, c'est-à-dire qu'ils maximisent leur espérance de gains monétaires. Supposons également que A croit que B n'est pas rationnel, A va donc choisir *Continuer* en étant presque certain que B va choisir *Partager* après *Continuer.* En étant égoïste et rationnel, B au lieu de choisir *Partager* après *Continuer* va donc choisir *Prendre*, ainsi le profil d'actions (*Continuer, Prendre*) va émerger.

est quant à lui averse à l'inégalité, et que tout cela soit de connaissance commune pour les joueurs. D'un point de vue formel, les préférences de B sont représentées par une fonction d'utilité de la même forme que celle de (1),

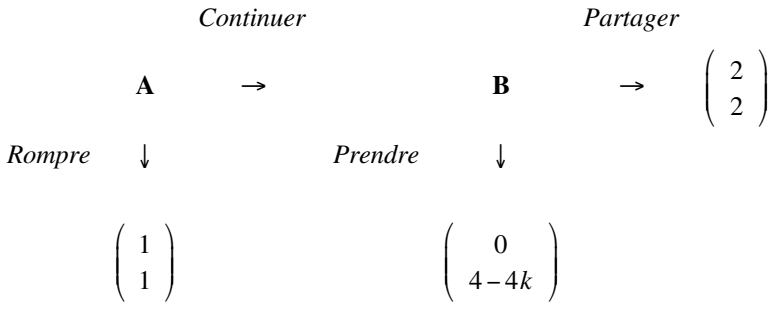
$$u_B(s_A,s_B) = \pi_B(s_A,s_B) - h\max\{0,\pi_A(s_A,s_B) - \pi_B(s_A,s_B)\}$$
$$-k\max\{0,\pi_B(s_A,s_B) - \pi_A(s_A,s_B)\} \qquad (2)$$

où s_i est la stratégie du joueur $i = A, B$, $\pi_i(s_A, s_B)$ le gain monétaire du joueur $i = A, B$, et h, k des paramètres positifs tel que $h \in [k, +\infty)$ et $k \in [0, 1)$.[15]

En référence au mini-jeu de la confiance du graphique 4, notons que, étant donné la structure des gains du jeu, il est impossible d'avoir $\pi_A(s_A, s_B) > \pi_B(s_A, s_B)$. La fonction d'utilité de (2) se simplifie alors et devient $\pi_B(s_A, s_B) - k\max\{0, \pi_B(s_A, s_B) - \pi_A(s_A, s_B)\}$. Nous avons $u_B(Rompre,\ s_B) = 1$ pour chaque $s_B = Prendre,\ Partager$, $u_B = (Continuer, Prendre) = 4 - 4k$, $u_B(Continuer, Partager) = 2$. Le jeu représentant le mini-jeu de la confiance avec un receveur (B) averse à l'inégalité est représenté dans le graphique 5. Ce qui apparaît à l'extrémité de l'arbre de décision doit être considéré comme des utilités et non comme des gains monétaires, bien que les deux notions coïncident pour deux des trois extrémités de l'arbre.

GRAPHIQUE 5

MINI-JEU DE LA CONFIANCE AVEC AVERSION À L'INÉGALITÉ DU RECEVEUR (B)



En adoptant un raisonnement à rebours, si $k \in (0, 1/2)$ le joueur β va choisir *Prendre* et, en anticipant cela, le joueur A va choisir *Rompre*, soit le même équilibre que dans le cas où les deux joueurs agiraient pour leur propre intérêt. Si $k \in (1/2, 1)$, alors le joueur B va choisir *Partager* et, en anticipant cela, le joueur A va choisir *Continuer*. Si $k = 1/2$, alors le joueur B sera indifférent entre *Prendre* et *Partager*. En anticipant cela, le joueur A va choisir *Rompre* (respectivement *Continuer*) s'il

15. Pour une discussion approfondie des hypothèses sous-jacentes de cette fonction d'utilité, voir la section 2.1.1.

pense que B va choisir *Partager* avec une probabilité inférieure (respectivement supérieure) à 50 %, c'est-à-dire si $\alpha_A = \Pr_A[Partager] < 1/2$ (respectivement > 1/2). Par conséquent, en plus des deux équilibres de Nash en stratégies pures, il existe ici un continuum d'équilibres de Nash en stratégies mixtes où le joueur B choisit *Partager* avec probabilité $\alpha_A \in (0, 1)$. Le joueur A va quant à lui choisir *Rompre* pour $\alpha_A \in (0, 1/2)$ et *Continuer* pour $\alpha_A \in (1/2, 1)$, et sera indifférent entre *Rompre* et *Continuer* pour $\alpha_A = 1/2$. Par ailleurs, si nous définissons par $\alpha_B = \Pr_B[Continuer]$ la croyance de premier ordre du joueur B, c'est-à-dire l'anticipation de B sur A choisissant *Continuer*, et dès lors que les croyances sont exactes à l'équilibre, la croyance de premier ordre α_A peut être interprétée comme la stratégie mixte de B et la croyance de premier ordre α_B comme la stratégie mixte de A. Cette notation nous permet d'écrire le continuum d'équilibres de Nash en stratégies mixtes de manière plus compacte.

Nous résumons dans le tableau 3 les équilibres de Nash parfaits en sous-jeux du jeu du graphique 5, pour toutes les valeurs positives du paramètre *k*.

TABLEAU 3

ÉQUILIBRES DE NASH PARFAITS EN SOUS-JEUX DU MINI-JEU DE LA CONFIANCE AVEC AVERSION À L'INÉGALITÉ DU RECEVEUR (B)

Sensibilité du receveur à l'inégalité au bénéfice du receveur	**Équilibre Pur (*Rompre*, *Prendre*)**	**Équilibre Pur (*Continuer*, *Partager*)**	**Équilibre Mixte**
$k \in (0, 1/2)$	Oui	Non	Non
$k = 1/2$	Oui	Oui	$\alpha_B = 0$, $\alpha_A \in (0, 1/2)$, Oui, $\alpha_B \in [0, 1]$, $\alpha_A = 1/2$, $\alpha_B = 1$, $\alpha_A \in (1/2, 1)$.
$k \in (1/2, 1)$	Non	Oui	Non

Notons que dans chacun de ces équilibres le comportement optimal du joueur B ne dépend pas de ses croyances de second ordre, c'est-à-dire de ce qu'il pense que le joueur A pense de lui. Par conséquent, le résultat (*Continuer, Partager*) devrait émerger dans une expérience indépendamment de l'anticipation de B sur l'anticipation de A sur B choisissant *Partager* après *Continuer.*

De même, comme nous le détaillerons dans la section 3, plusieurs expériences en laboratoire montrent qu'il existe une corrélation positive significative entre l'anticipation du joueur B sur la confiance que le joueur A lui accorde (c'est-à-dire ses croyances de second ordre sur *Partager* après *Continuer*) et la coopération

effective de B (*Partager* après *Continuer*). Ce fait stylisé peut être expliqué par des préférences dépendantes des croyances.

2.2.2 *Jeu de la confiance avec aversion à la culpabilité du receveur*

La corrélation positive entre l'action *Partager* du receveur (B), et ses croyances de second ordre concernant cette action, peut être déterminée à l'équilibre si B est supposé averse à la culpabilité.

Comme dans la section 2.1.2, nous adoptons la même définition que Tangney (1995) pour l'aversion à la culpabilité et la même analyse que Battigalli et Dufwenberg (2009) pour les jeux avec préférences dépendantes des croyances (jeux psychologiques).

Considérons le mini-jeu de la confiance avec gains monétaires du graphique 4.

Avant de procéder à l'analyse, tel qu'annoncé à la section 3.1.2, nous devons aborder une autre question technique qui apparaît quand on essaye d'étendre la théorie des jeux standard afin de tenir compte des préférences dépendantes des croyances. La question de l'exactitude des croyances nécessitant l'exactitude des utilités déclarées, que nous avons déjà abordée pour le mini-jeu du dictateur avec aversion à la culpabilité du proposant dans le graphique 3, se pose également pour le mini-jeu de la confiance avec aversion à la culpabilité du receveur (B). Toutefois, pour ce dernier une autre question doit être également débattue, du fait que nous sommes en présence d'un jeu dynamique. En effet, dans un jeu dynamique, au fur et à mesure du déroulement du jeu, les joueurs révisent leurs croyances sur les croyances des autres. Mais si les joueurs ont des préférences dépendantes des croyances, le fait que les joueurs mettent à jour leurs croyances implique qu'ils doivent en faire de même pour leurs utilités (révision des utilités déclarées). Pour faire face à ce problème, Battigalli et Dufwenberg (2009) construisent un cadre général qui permet de prendre en compte des croyances conditionnelles dans la fonction d'utilité totale des joueurs.[16] Il est possible dans ce cadre général, d'étendre le concept traditionnel d'équilibre séquentiel aux jeux psychologiques.

Pour ce qui est du mini-jeu de la confiance avec aversion à la culpabilité du receveur (B), nous supposons que A est un joueur égoïste, que B est un joueur averse à la culpabilité, et que tout cela est de connaissance commune parmi les joueurs.

Puisque A est uniquement motivé par son intérêt personnel, sa sensibilité à la culpabilité est égale à zéro. Par conséquent, la fonction d'utilité totale de A se résume à son gain monétaire. *A contrario*, comme B est motivé par l'aversion à la culpabilité, il prend en compte la déception du joueur A quand le gain monétaire que celui-ci s'attend à recevoir après *Continuer* est inférieur à celui qu'il reçoit à la fin du jeu.

16. Les croyances conditionnelles sont des croyances qui dépendent des actions choisies par les autres joueurs avant le joueur concerné par ces croyances n'ait choisi son action.

Le montant monétaire que A anticipe de recevoir après *Continuer* dépend de sa croyance de premier ordre sur la stratégie de B. Définissons la croyance de premier ordre de A qui croit que B va choisir *Partager* par $\alpha_A = \Pr_A[Partager]$. α_A mesure alors la confiance du joueur A pour le joueur B avant que le jeu ne démarre. Définissons également la croyance conditionnelle de second ordre de B qui choisirait *Partager* si A choisit *Continuer* par $\beta_B^{Cont} = E_B[\alpha_A | Continuer]$. β_B^{Cont} mesure l'anticipation de B quant à la confiance que A lui accorde étant donné que B sait que A a choisi *Continuer*. En effet, plus β_B^{Cont} est élevé, plus B croit que A pense que B va choisir *Partager*, si A a choisi *Continuer*.

Étant donné les notations que nous avons introduites, nous pouvons définir la déception du joueur A et la culpabilité du joueur B si ce dernier choisit *Prendre* après que A ait choisi *Continuer*.

Le montant que le joueur A anticipe de recevoir après *Continuer* est donné par $E_{Cont.,\alpha_A}[\pi_A] = 2 \cdot \alpha_A + 0 \cdot (1 - \alpha_A) = 2\alpha_A$. La déception de A après (*Continuer, Prendre*) est donc égale à $-2\alpha_A$, c'est-à-dire la différence entre le gain que celui-ci reçoit après (*Continuer, Prendre*), qui est nul, et le gain moyen qu'il espérait obtenir après *Continuer*, conformément à sa croyance de premier ordre α_A.

La culpabilité de B est donnée par son anticipation quant à la déception de A, étant donné que A a choisi *Continuer*, c'est-à-dire l'anticipation de B pour $(-2\alpha_A)$, étant donné le choix de A pour *Continuer*. Ce montant, $E_B[-2\alpha_A | Continuer] = -2E_B[\alpha_A | Continuer] = -2\beta_B^{Cont.}$; multiplié par la sensibilité de B à l'aversion à la culpabilité, $\theta \geq 0$,[17] représente l'utilité psychologique de B lorsque A choisit *Continuer* et B choisit *Prendre*. L'utilité totale de B après (*Continuer, Prendre*) est donnée par $4 - 2\theta\beta_B^{Cont.}$, c'est-à-dire la somme de son gain monétaire et de son utilité psychologique.

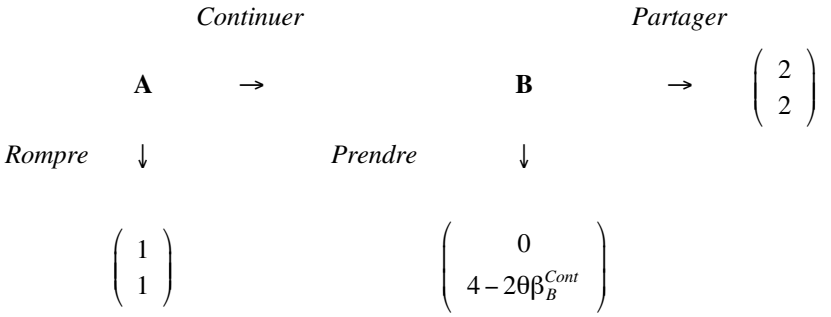
Le jeu psychologique représentant le jeu de la confiance avec un receveur (B) averse à la culpabilité est illustré au graphique 6. Ce qui apparaît à l'extrémité de l'arbre de décision doit être considéré comme des utilités et non comme des gains monétaires bien que les deux notions coïncident pour deux des trois extrémités de l'arbre.

Recherchons à présent l'ensemble des équilibres séquentiels psychologiques pour le jeu psychologique du graphique 6 pour toutes les valeurs positives possibles du paramètre θ. À cette fin, nous allons nous focaliser sur les valeurs d'équilibres des croyances fondamentales α_A, $\beta_B^{Cont.}$ et α_B. α_B est déterminée avant que le jeu ne démarre et a été précédemment définie comme la croyance de premier ordre du joueur B, c'est-à-dire l'anticipation de B sur A choisissant *Continuer*. Dès lors que les croyances sont exactes à l'équilibre, la croyance de premier ordre α_A peut être interprétée comme la stratégie mixte de B, et la croyance de premier ordre α_B peut être interprétée comme la stratégie mixte de A. En outre, la condition de l'exactitude des croyances suppose qu'on ait également à l'équilibre $\alpha_A = \beta_B^{Cont}$.

17. À nouveau, la sensibilité des joueurs est supposée être hétérogène selon les joueurs.

GRAPHIQUE 6

MINI-JEU DE LA CONFIANCE AVEC AVERSION À LA CULPABILITÉ DU RECEVEUR (B)



Nous résumons dans le tableau 4 les équilibres séquentiels psychologiques du jeu psychologique du graphique 6, pour toutes les valeurs positives du paramètre θ. La détermination des équilibres ainsi que l'explication de la méthodologie utilisée sont disponibles à l'annexe 2.

TABLEAU 4

ÉQUILIBRES SÉQUENTIELS PSYCHOLOGIQUES DU MINI-JEU DE LA CONFIANCE AVEC AVERSION À LA CULPABILITÉ DU RECEVEUR (B)

Sensibilité du receveur à la culpabilité	**Équilibre Pur (*Rompre*, *Prendre*)**	**Équilibre Pur (*Continuer*, *Partager*)**	**Équilibre Mixte**
$\theta \in (0, 1)$	Oui	Non	Non
$\theta = 1$	Oui	Oui	Non
$\theta \in (1, 2)$	Oui	Oui	Oui, $\alpha_B = 1$, $\alpha_A = 1/\theta$
$\theta = 2$	Oui	Oui	Oui, $\alpha_B \in [0, 1]$,$\alpha_A = 1/2$
$\theta \in (2, +\infty)$	Oui	Oui	Oui, $\alpha_B = 0$, $\alpha_A = 1/\theta$

Notons que (*Rompre, Prendre*) est toujours un équilibre, mais si B est suffisamment sensible à la culpabilité ($\theta \geq 1$) alors il y a une multiplicité d'équilibres. La raison est que, même si B est extrêmement averse à la culpabilité, on ne peut pas exclure la possibilité que celui-ci pense que A, quand il choisit *Continuer*, anticipe que B choisisse *Partager* avec une probabilité égale à 0. Dans ce cas nous avons $\beta_B^{Cont} = 0$, l'utilité totale de *Prendre*, c'est-à-dire $u_B((Continuer, Prendre); \beta_B^{Cont})$, se réduit alors à $u_B((Continuer, Prendre); 0) = 4$, comme dans le cas d'un joueur B égoïste. À l'inverse, si B pense que A anticipe qu'il va choisir *Partager* avec une

probabilité égale à 1, alors l'équilibre est (*Continuer, Partager*), c'est-à-dire que $\alpha_R = \beta_B^{Cont} = 1$. Dans le cas où la sensibilité à la culpabilité est suffisamment élevée ($\theta > 1$), il existe également un équilibre mixte, où B choisit *Partager* avec une probabilité de $1/\theta \in (0, 1)$. Notons que, plus θ est grand, plus la probabilité avec laquelle B choisit *Partager* dans l'équilibre mixte est faible, de même pour les anticipations de A concernant l'action *Partager.* Par ailleurs, pour un θ pas trop élevé (entre 1 et 2), le choix qui maximise le profit de A est *Continuer*, c'est-à-dire $\alpha_B = 1$. Pour un θ très élevé (plus grand que 2), le choix qui maximise le profit de A est *Prendre*, c'est-à-dire $\alpha_B = 0$. Pour la valeur intermédiaire de θ ($\theta = 2$), nous avons un continuum d'équilibres séquentiels psychologiques en stratégies mixtes avec A choisissant *Continuer* pour toute probabilité $\alpha_B \in [0, 1]$.

Existe-t-il alors un critère de sélection d'équilibres acceptable? A l'instar de Dufwenberg (2002), considérons l'argument de l'induction psychologique vers l'avant suivant, et qui est assez intuitif : il est rationnel pour A de faire confiance à B seulement si A attribue une probabilité supérieure à 50 % à la stratégie *Partager.* En d'autres termes, la décision optimale de A dépend du choix de B, c'est-à-dire qu'il doit comparer le gain espéré en choisissant *Continuer*, qui est égal à $2\alpha_A$, avec le gain certain en choisissant *Rompre*, qui lui est égal à 1. Par conséquent, un joueur A rationnel choisit de faire confiance à B, et de poursuivre le partenariat, seulement si sa croyance de premier ordre concernant *Partager* est égale à $\alpha_A \geq 1/2$. Par la suite, chaque fois que B observe l'action *Continuer*, et s'il croit que A est rationnel, alors sa croyance de second ordre concernant *Partager* (conditionnellement après avoir observé *Continuer*) doit être égale à $\beta_B^{Cont} \geq 1/2$. La logique de cet argument dépend donc de la croyance β_B^{Cont} (sous réserve que A choisisse *Continuer*). La logique de l'induction vers l'avant implique en quelque sorte une révision de cette croyance, c'est-à-dire une restriction de l'ensemble des croyances de second ordre entrant dans la fonction d'utilité du joueur B lorsque son nœud de décision est atteint. Un tel critère, dans lequel les joueurs influencent de manière stratégique les croyances de leurs partenaires, a un impact encore plus important dans les jeux psychologiques que dans les jeux traditionnels, ceci en raison de l'effet direct des croyances sur les utilités des joueurs.

3. Un résumé des principaux résultats expérimentaux

Après avoir passé en revue les principaux résultats d'études expérimentales dans le jeu du dictateur et le jeu de la confiance (respectivement en 3.1 et 3.2), nous nous demanderons s'ils peuvent être interprétés selon les deux modèles avec préférences sociales proposés dans cet article.

3.1 *Résultats expérimentaux sur le jeu du dictateur*

La première expérience du jeu du dictateur en économie a été réalisée par Kahneman *et al*. (1986). Les sujets se sont vus proposer un choix hypothétique et autoritaire entre une répartition égalitaire de 20 $ (10 $, 10 $) avec un autre

partenaire ou une répartition inégale (18 $, 2 $) en leur faveur.[18] Les résultats obtenus montrent que les trois quarts des participants ont opté pour la répartition égalitaire. Forsythe *et al.* (1994) ont par la suite expérimenté ce jeu du dictateur avec de vraies incitations monétaires et un ensemble de répartition plus étendu et ont observé que les sujets transféraient en moyenne 20 % de la dotation. Comme l'a ensuite souligné Camerer (2003, p. 57, tableau 2.4), une pléthore d'études expérimentales sur le jeu du dictateur a par la suite reproduit ces résultats, en trouvant que généralement plus de 60 % des sujets donnent un montant monétaire non nul en transférant en moyenne presque 20 % de la dotation. Les résultats ci-dessus sont compatibles avec le modèle de Fehr et Schmidt (1999) de l'aversion à l'inégalité. En effet, dans ces expériences, la plupart des proposants ont une sensibilité *k* à l'inégalité *qui est à leur avantage* positive, mais avec une valeur moyenne de *k* faible. Cependant, leur modèle montre qu'il y a également une fraction non négligeable de sujets qui présentent une aversion à l'inégalité non linéaire lorsque cette inégalité est en leur faveur.

On pourrait également expliquer ces résultats par un modèle d'aversion à la culpabilité avec préférences dépendantes des croyances. De plus, des études expérimentales récentes semblent conforter l'hypothèse de l'aversion à la culpabilité du proposant dans le mini-jeu du dictateur.

Bellemare *et al.* (2014) ont par exemple élicité les préférences dépendantes des croyances du proposant dans un mini-jeu du dictateur similaire à celui que nous analysons dans cet article. Ils demandent à chaque receveur dans une session expérimentale de deviner le nombre de proposants présents dans la même session qui choisiraient *Partager* dans un mini-jeu du dictateur stratégiquement équivalent à celui du graphique 1 (croyances de premier ordre du receveur α_R).[19] Ils demandent ensuite à chaque proposant de choisir entre *Prendre* et *Partager* pour chaque croyance de premier ordre que le receveur aurait pu déclarer. Ceci peut être interprété comme le fait de laisser le proposant faire un choix possiblement différent pour chacune de ses propres croyances de second ordre concernant le choix *Partager*, c'est-à-dire pour chaque valeur possible de β_P. Faire le même choix pour tous les β_P révélerait des préférences qui ne sont pas dépendantes des croyances. De même, si son choix était toujours *Prendre*, alors on aurait la preuve de l'hypothèse d'un proposant purement égoïste. Si *a contrario* son choix était toujours *Partager*, on pourrait émettre l'hypothèse d'un proposant averse à l'inégalité avec un *k* assez élevé. Cependant, Bellemare *et al.* (2014) constatent que la moitié des sujets ne font pas les mêmes choix pour tous les β_P. En particulier, quand β_P augmente, ils passent une fois de *Prendre* à *Partager* pour un $\beta_P \in (0, 1)$, fournissant ainsi la preuve d'une corrélation positive entre le choix *Partager* et la croyance de second ordre du proposant pour *Partager*. Ainsi, la moitié des sujets sont sensibles à la

18. Cas très proche du mini-jeu du dictateur analysé théoriquement dans cet article (graphique 1).

19. Il y a 10 proposants dans la session, donc les croyances de premier ordre possibles sont : 0 %, 10 %,..., 100 %.

culpabilité, avec une forte hétérogénéité de la sensibilité à la culpabilité détectée chez les joueurs. Ils constatent également que 35 % de leurs sujets sont égoïstes, et que seuls 15 % ont des préférences distributives à la Fehr et Schmidt (1999). Une autre observation intéressante est que, lorsque les gains monétaires du mini-jeu du dictateur augmentent de manière proportionnelle, la part des sujets égoïstes augmente également.

Khalmetski *et al.* (2015) confirment ces résultats pour un jeu du dictateur avec un ensemble de choix continus pour le proposant (il peut donner au receveur une part de sa dotation de 10€). Ils demandent à chaque receveur de deviner la part de la dotation que le proposant qui lui est apparié de manière aléatoire lui donnerait (croyance de premier ordre α_R). Ils demandent ensuite à chaque proposant de deviner la croyance moyenne des receveurs dans l'expérience (croyance de second ordre du proposant β_P). Leurs résultats montrent que la corrélation entre les transferts des proposants et leurs croyances de second ordre peut être à la fois positive (preuve de leur aversion à la culpabilité) et négative, ce qui brouille un peu l'effet au niveau agrégé. Néanmoins, ils soulignent que les proposants qui montrent cette corrélation négative ont une préférence relativement importante à générer de bonnes surprises[20] chez les receveurs, encore un signe de la présence manifeste des préférences dépendantes des croyances. Tout ceci fournit une possible explication au fait que Ellingsen *et al.* (2010) n'aient trouvé aucune preuve de corrélation entre les croyances de second ordre du proposant et son comportement, dans une étude expérimentale où les croyances de second ordre des proposants ont été induites directement en révélant les croyances de premier ordre des receveurs et sans qu'ils sachent que leurs croyances allaient être révélées.[21]

Enfin, une dernière série d'expériences sur le jeu du dictateur mérite d'être mentionnée ici : celles dont l'impact des préférences sociales est minime lorsque les choix du proposant dans un jeu du dictateur ne sont pas observables (indépendamment du fait que ces préférences sociales soient distributives ou dépendantes des croyances). Hoffman *et al.* (1996) ont par exemple montré qu'un plus grand anonymat du proposant le conduit à se comporter davantage de manière égoïste et réduit la fréquence des choix *Partager.* Dana *et al.* (2006), Dana *et al.* (2007) et Broberg *et al.* (2007) obtiennent le même résultat pour des traitements qui brouillent le rôle du proposant dans la détermination des gains ou qui lui permettent de cacher son rôle. S'appuyant sur ces résultats expérimentaux, Andreoni et Bernheim (2009)

20. Les gens se soucient non seulement des surprises négatives mais aussi des surprises positives induites par leurs choix.

21. Khalmetski *et al.* (2015) ont par la suite répliqué leur expérience mais en recueillant cette fois-ci l'ensemble des vecteurs de stratégies des proposants. Ils ont constaté que les proposants réagissent aux attentes des receveurs de façon hétérogène. Une corrélation positive ainsi qu'une corrélation négative entre les croyances et les transferts du proposant ont été observées, la première étant expliquée par l'aversion à la culpabilité du proposant, la deuxième par l'effet *bonnes surprises* (toujours dépendante des croyances). Les deux motivations, comme elles ne peuvent pas se différencier l'une de l'autre lorsqu'on agrège les données, se neutralisent en quelque sorte, ce qui laisserait à penser de manière erronée que les préférences dépendantes des croyances ne jouent pas un rôle dans le jeu du dictateur.

fournissent une explication théorique alternative à un comportement équitable dans le jeu du dictateur : non seulement les gens se soucient de l'équité, mais ils aiment aussi être perçus comme équitables.

3.2 *Résultats expérimentaux sur le jeu de la confiance*

Dans le jeu de la confiance, l'aversion à l'inégalité a été l'une des premières explications adoptée pour expliquer les résultats expérimentaux qui s'écartent des prédictions de la théorie des jeux standard concernant la défiance du proposant et la défection du receveur. Depuis les premières expériences sur le jeu de la confiance, les résultats diffèrent sensiblement de la prédiction théorique de non-coopération. Par exemple, le résultat de non-coopération ne se produit que 5 fois sur 60 dans Berg *et al.* (1995) et 4 fois sur 97 dans Glaeser *et al.* (2000). De même, si l'on observe que seule une faible proportion de receveurs partagent équitablement le gain dans la première expérience (8 fois sur 55), cette proportion devient très élevée dans la deuxième (55 fois sur 93). Toutefois, étant donné que les croyances n'ont pas été élicitées dans ces expériences, on ne peut pas déterminer si le partage égalitaire est le fruit de la coopération inconditionnelle du receveur (hypothèse de préférences distributives), ou d'une coopération conditionnelle de celui-ci (hypothèse de préférences dépendantes des croyances).

Plusieurs expériences récentes ont apporté des preuves convaincantes en faveur de cette dernière hypothèse. Les études expérimentales du jeu de la confiance montrent une corrélation positive entre l'élicitation des croyances de second ordre et l'option *Partager.* Ceci renforce l'hypothèse que dans ce dilemme social, l'aversion à la culpabilité est la motivation psychologique qui prévaut pour le receveur. Plus précisément, considérons notre mini-jeu de la confiance du graphique 4. Une sensibilité positive (et suffisamment élevée) à la culpabilité impose un comportement particulier au receveur (joueur B). Supposons qu'il soit possible d'obtenir dans une expérience une élicitation correcte de l'anticipation de B concernant l'anticipation de A sur le fait que B choisirait *Partager* après *Continuer* (croyance de second ordre de B pour *Partager* après avoir observé *Continuer*). Dans ce cas, une corrélation positive entre les croyances conditionnelles de second ordre de B concernant *Partager* (après *Continuer*) et la réalisation de cette confiance (B choisissant *Partager* après *Continuer*) devrait être détectée, ainsi qu'une croyance conditionnelle moyenne de second ordre plus élevée de sujets B choisissant *Partager* après *Continuer* que de sujets B choisissant *Prendre* après *Continuer.* C'est exactement ce qui est montré dans la fameuse étude de Charness et Dufwenberg (2006). C'est la première étude expérimentale qui a testé explicitement l'hypothèse de l'aversion à la culpabilité (due à la déception), confirmée par la suite par plusieurs autres études expérimentales (Guerra et Zizzo, 2004, Bacharach *et al.*, 2007, Chang *et al.*, 2011, Charness et Dufwenberg, 2011, Bracht et Regner, 2013 et Regner et Harth, 2014). *A contrario*, Vanberg (2008), Ellingsen *et al.* (2010) et Kawagoe et Narita (2014) n'ont pas trouvé quant à eux de preuves évidentes corroborant l'hypothèse de l'aversion à la culpabilité dans le jeu de la confiance.

Enfin, des preuves évidentes en faveur de l'hypothèse de l'aversion à la culpabilité, par rapport aux autres types de préférences sociales comme l'aversion à l'inégalité (préférence distributive) ou la réciprocité fondée sur les intentions (préférence dépendante des croyances), ont été apportées récemment par Attanasi *et al.* (2013). Ils étudient expérimentalement comment, dans le mini-jeu de la confiance, les préférences sociales des receveurs sur les distributions de gains monétaires dépendent de leurs croyances. L'aversion à la culpabilité des receveurs prévaut pour une large majorité (55 %), tandis que seuls 9 % des sujets peuvent être classés comme étant averses à l'inégalité et seuls 15 % des receveurs peuvent être classés comme égoïstes. En outre, 21 % des receveurs montrent d'autres formes de préférences dépendantes des croyances, 16 % étant guidés par la réciprocité fondée sur les intentions (Rabin, 1993) et 5 % par un mélange d'aversion à la culpabilité et d'aversion à la réciprocité fondées sur l'intention. Au total, ce sont donc les trois quarts de leurs sujets qui montrent une certaine forme de préférence dépendante des croyances.

Conclusion

Les préférences sociales dans les jeux de dilemme social peuvent prendre différentes formes. Dans cet article, nous avons identifié deux grandes familles, selon deux principes simples : i) les sujets choisissent des actions prosociales parce qu'ils se soucient de la répartition des gains monétaires entre eux et les autres sujets; ii) les sujets choisissent des actions prosociales parce qu'ils se soucient de ce que les autres sujets pensent qu'ils feraient ou devraient faire. S'appuyant sur la littérature théorique existante sur ce sujet, nous avons désigné la première famille de préférences sociales comme les préférences distributives, et la deuxième comme les préférences dépendantes des croyances.

Du point de vue théorique, nous avons mis l'accent sur le fait que les modèles avec préférences distributives sont en un sens plus restreints que les modèles avec préférences dépendantes des croyances, qui eux de surcroît prennent en compte les croyances des joueurs. Dans le premier cas, les joueurs ont des préférences fixes, alors que dans le second ils ont des préférences qui évoluent avec les croyances. Cela confère aux modèles avec préférences dépendantes des croyances un plus grand degré de liberté, et donc un avantage en termes de capacité prédictive. Cependant, nous avons également montré que cet avantage est pondéré par un problème de multiplicité d'équilibres.

Pour chacune de ces familles, nous nous sommes concentrés sur une motivation sociale spécifique : l'aversion à l'inégalité à la Fehr et Schmidt (1999) associée aux préférences distributives, et l'aversion à la culpabilité à la Battigalli et Dufwenberg (2009) associée aux préférences dépendantes des croyances.

Nous avons montré que chacune de ces deux motivations, une fois incorporée dans les préférences des joueurs, peut expliquer théoriquement les déviations par rapport au comportement égoïste prédit dans les dilemmes sociaux comme le jeu

du dictateur et le jeu de la confiance. Il s'agit de deux jeux de dilemme social bien connus et largement analysés dans la littérature expérimentale des préférences sociales. Les économistes expérimentaux ont montré que des comportements d'aversion à l'inégalité ainsi que d'aversion à la culpabilité se manifestent dans les expériences économiques pour ces deux jeux. Une autre raison pour laquelle nous nous focalisons sur ces deux jeux est que le rôle du proposant dans le jeu du dictateur peut être considéré comme stratégiquement équivalent à celui du receveur dans le jeu de la confiance, dès lors qu'on ne s'intéresse qu'au sous-jeu où le receveur est le seul joueur actif. En outre, les expériences menées en laboratoire sur le jeu de la confiance ont principalement porté sur l'analyse des préférences sociales des sujets dans le rôle du receveur.

Notre analyse théorique de ces deux jeux soulève deux grandes questions théoriques liées. Premièrement, alors que les outils traditionnels de la théorie des jeux peuvent être appliqués pour résoudre les jeux avec préférences distributives, il n'en est pas de même pour les jeux avec préférences dépendantes des croyances. Une extension de l'approche traditionnelle, définie comme la théorie des jeux psychologiques (Dufwenberg, 2008), est nécessaire. Deuxièmement, la prédiction théorique quant à la coopération dans les jeux avec préférences distributives n'a rien à voir avec les croyances des joueurs sur les croyances des autres joueurs (croyances de second ordre). Ceci n'est plus vrai dans les jeux avec préférences dépendantes des croyances où, par exemple, deux proposants ayant la même sensibilité à la culpabilité dans le jeu du dictateur peuvent faire des choix totalement opposés parce qu'ils ont des croyances différentes sur ce que leur receveur attend d'eux. La même remarque peut être faite pour deux receveurs averses à la culpabilité dans le jeu de la confiance.

Par conséquent, nous soulignons ici un point important à propos de la méthodologie expérimentale utilisée pour mesurer l'impact des préférences sociales dans les jeux de dilemme social. Tant que les croyances de premier et de second ordre des joueurs ne sont pas élicitées au cours de l'expérience, on ne peut pas faire de distinction entre les théories avec préférences distributives et les théories avec préférences dépendantes des croyances. L'impact des préférences distributives peut être également surestimé. Des études expérimentales récentes, où les croyances des joueurs sont élicitées, ont en effet confirmé que tant dans le jeu du dictateur que dans le jeu de la confiance, l'aversion à la culpabilité est le sentiment le plus répandu.

Cependant, dans le jeu du dictateur, chez les sujets ayant des préférences sociales, une proportion importante de sujets averses à l'inégalité est toujours détectée, même en tenant compte des préférences dépendantes des croyances. *A contrario*, la proportion de receveurs averses à l'inégalité détectée dans les jeux de la confiance est assez faible lorsque la méthode expérimentale permet l'élicitation des préférences dépendantes des croyances. Ceci prouve, de manière indirecte, la pertinence des préférences dépendantes des croyances. En fait, la principale différence entre le rôle du proposant dans le jeu du dictateur et le rôle du receveur

dans le jeu de la confiance, est que le choix de ce dernier n'a d'importance que si le proposant a décidé de lui faire confiance. Ceci nous informe sur les intentions du proposant dans le jeu de la confiance, et de la perception de celles-ci par le receveur : c'est lorsque les intentions prennent de l'importance que les préférences dépendantes des croyances se manifestent encore davantage.

Une dernière remarque s'impose : dans la théorie des jeux psychologiques, il est supposé que les préférences des joueurs sur les conséquences monétaires dépendent des croyances endogènes. La plupart des applications entrant dans ce cadre théorique supposent que les fonctions d'utilité psychologique représentant ces préférences soient de connaissance commune pour les joueurs. Mais cette condition est souvent irréaliste. En particulier, cela ne peut être le cas dans les jeux expérimentaux où les joueurs sont des sujets sélectionnés au hasard dans une population. Par conséquent une méthodologie en information incomplète s'impose. Attanasi *et al.* (2016) font un premier pas dans cette direction, en se focalisant sur l'aversion à la culpabilité dans le jeu de la confiance.

ANNEXE

Nous détaillons dans ces annexes le calcul des *équilibres de Nash psychologiques* du mini-jeu du dictateur avec aversion à la culpabilité du proposant (tableau 2) et des *équilibres séquentiels psychologiques* du mini-jeu de la confiance avec aversion à la culpabilité du receveur (tableau 4).

1. Équilibres du mini-jeu du dictateur avec aversion à la culpabilité

Pour $\theta \in [0, 1)$, on a $u_P(Prendre; \beta_P) > u_P(Partager)$, $\forall \beta_P \in [0, 1]$. Le proposant choisit alors *Prendre* et le seul profil d'équilibre de Nash psychologique en croyances est $\alpha_R = \beta_P = 0$.

Pour $\theta \in (1, +\infty)$, selon les valeurs prises par β_P, on pourrait avoir les deux cas suivants :

$u_P(Prendre; \beta_P) \lesseqgtr u_P(Partager)$.

Examinons dans un premier temps les équilibres en stratégies pures.

Pour $\beta_P = 0$, on a $u_P(Prendre; \beta_P) > u_P(Partager)$, le proposant choisit alors *Prendre* et lui et le receveur ont tous les deux des croyances exactes à l'équilibre. Par conséquent, $\alpha_R = \beta_P = 0$ est un équilibre de Nash psychologique en croyances.

Pour $\beta_P = 1$, on a $u_P(Prendre; \beta_P) < u_P(Partager)$, le proposant choisit alors *Partager*, et lui et le receveur ont tous les deux des croyances exactes à l'équilibre. Par conséquent, $\alpha_R = \beta_P = 1$ est un équilibre de Nash psychologique en croyances.

Examinons à présent les équilibres en stratégies mixtes. Le proposant est indifférent et peut jouer de manière optimale une stratégie mixte si et seulement si $4 - 2\theta\beta_P = 2$, c'est-à-dire $\beta_P = 1/\theta$. Suivant la condition de l'exactitude des croyances, on doit obtenir pour chaque équilibre en stratégies mixtes $\alpha_R = \beta_P = 1/\theta$. Notons que pour que l'équilibre soit un équilibre en stratégies mixtes on doit avoir $\theta \neq 1$, sinon le proposant jouerait la stratégie pure *Partager.* Ainsi, les équilibres en stratégiques mixtes ne peuvent exister que pour $\theta \in (1, +\infty)$, dès lors que $0 < 1/\theta < 1$.

Gardant à l'esprit que, pour chaque équilibre, nous avons besoin par définition que $\alpha_R = \beta_P$, nous obtenons les équilibres de Nash psychologiques résumés dans le tableau 2.

2. Équilibres du mini-jeu de la confiance avec aversion à la culpabilité

Pour $\theta \in (0, 1)$, on a $u_B(Continuer, Prendre; \beta_B^{Cont.}) > u_B(Continuer, Partager)$, $\forall \beta_B^{Cont.} \in [0, 1]$. Le joueur B choisit donc toujours *Prendre* et A, par raisonnement à rebours, choisit *Rompre*. L'unique profil d'équilibre séquentiel psychologique en croyances est donc $\alpha_B = 0$, $\alpha_A = \beta_B^{Cont.} = 0$.

Pour $\theta \in [1, +\infty)$, selon les valeurs prises par β_P, on pourrait avoir les deux cas suivants :

$u_B(\textit{Continuer}, \textit{Prendre}; \beta_B^{Cont.}) \lesseqgtr u_B(\textit{Continuer}, \textit{Partager})$.

Examinons dans un premier temps les équilibres en stratégies pures.

Pour $\beta_B^{Cont.} = 0$, on a $u_B(\textit{Continuer}, \textit{Prendre}; \beta_B^{Cont.}) > u_B(\textit{Continuer}, \textit{Partager})$, B choisit alors *Prendre* et comme A a des croyances exactes à l'équilibre ($\alpha_A = 0$), il choisit *Rompre*. On a alors un équilibre séquentiel psychologique en croyances avec $\alpha_B = 0$, $\alpha_A = \beta_B^{Cont.} = 0$.

Pour $\beta_B^{Cont} = 1$, on a $u_B(\textit{Continuer}, \textit{Prendre}; \beta_B^{Cont.}) < u_B(\textit{Continuer}, \textit{Partager})$, B choisit alors *Partager* et comme A a des croyances exactes à l'équilibre ($\alpha_A = 1$), il choisit *Continuer*. On a alors un équilibre séquentiel psychologique en croyances avec $\alpha_B = 1$, $\alpha_A = \beta_B^{Cont} = 1$.

Examinons à présent les équilibres en stratégies mixtes. B est indifférent et il peut joueur une stratégie mixte optimale si et seulement si $4 - 2\theta\beta_B = 2$, c.-à-d. $\beta_B^{Cont.} = 1/\theta$. Suivant la condition de l'exactitude des croyances, on doit avoir pour chaque équilibre en stratégies mixtes $\alpha_A = \beta_B^{Cont.} = 1/\theta$. Quant au joueur A, étant donné que $\alpha_A = 1/\theta$, A choisit *Continuer* si $2 \cdot 1/\theta + 0 \cdot (\theta - 1)/\theta > 1$, c.-à-d. $\theta \in (1, 2)$, il choisit *Rompre* si $\theta \in (2, +\infty)$, et il est indifférent entre *Continuer* et *Rompre* si et seulement si $\theta = 2$. La condition de l'exactitude des croyances de premier ordre nous donne $\alpha_B = 1$ si $\theta \in (1, 2)$, $\alpha_B = 0$ si $\theta \in (2, +\infty)$ et $\alpha_B \in [0, 1]$ si $\theta = 2$.

Gardant à l'esprit que, pour chaque équilibre, nous avons besoin par définition que $\alpha_A = \beta_B^{Cont}$, nous obtenons les équilibres séquentiels psychologiques résumés dans le tableau 4.

BIBLIOGRAPHIE

Andreoni, J. et B.D. Bernheim (2009), « Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects », *Econometrica*, 77 : 1607–1636.

Andreoni, J. et J. Miller (2002), « Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism », *Econometrica*, 70 : 737–753.

Attanasi, G., P. Battigalli et E. Manzoni (2016), « Incomplete Information Models of Guilt Aversion in the Trust Game », *Management Science*, 62 : 648-667.

Attanasi, G., P. Battigalli et R. Nagel (2013), « Disclosure of Belief-Dependent Preferences in the Trust Game », IGIER Working Paper 506, Bocconi University.

Bacharach, M. (1999), « Interactive Team Reasoning: a Contribution to the Theory of Cooperation », *Research in Economics*, 23 : 117–147.

Bacharach, M., G. Guerra et D.J. Zizzo (2007), « The Self–Fulfilling Property of Trust: An Experimental Study », *Theory and Decision*, 63 : 349–388.

Baumeister, R. F., T. F. Heatherton et D. M. Tice, (1994), *Losing Control: How and Why People Fail at Self-Regulation*, San Diego, CA : Academic Press.

BATTIGALLI, P. et M. DUFWENBERG (2007), « Guilt in Games », *American Economic Review: Papers and Proceedings*, 97 : 170–176.

BATTIGALLI, P. et M. DUFWENBERG (2009), « Dynamic Psychological Games », *Journal of Economic Theory*, 144 : 1–35.

BATTIGALLI, P., M. DUFWENBERG et A. SMITH (2015), « Frustration and Anger in Games », IGIER Working Paper 539, Bocconi University.

BELLEMARE, C., A. SEBALD et S. SUETENS (2014), « Heterogeneous Guilt Aversion and Incentive Effects », Working Paper, Tilburg University.

BERG, J., J. DICKHAUT et K. MCCABE (1995), « Trust, Reciprocity and Social History », *Games and Economic Behavior*, 10 : 122–142.

BICCHIERI, C. et A. CHAVEZ (2010), « Behaving as Expected: Public Information and Fairness Norms », *Journal of Behavioral Decision Making*, 23 : 161–178.

BINMORE, K., J. GALE et L. SAMUELSON (1995), « Learning to be Imperfect: The Ultimatum Game », *Games and Economic Behavior*, 8 : 56–90.

BOLTON, G.E. et A. OCKENFELS (2000), « ERC: A Theory of Equity, Reciprocity, and Competition », *American Economic Review*, 90 : 166–193.

BRACHT, J. et T. REGNER (2013), « Moral Emotions and Partnership », *Journal of Economic Psychology*, 39 : 313–326.

BROBERG, T., ELLINGSEN, T. et M. JOHANNESSON (2007), « Is Generosity Involuntary », *Economic Letters*, 94 : 32–37.

BUSKENS, V.et W. RAUB (2013), « Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust », *in* WITTEK R., T. SNIJDERS et V. NEE (éds.), *Handbook of Rational Choice Social Research*, New York: Russell Sage, p. 113–150.

CAMERER, C.F. (2003), *Behavioral Game Theory: Experiments on Strategic Interaction*, Princeton University Press.

CAPLIN, A. et J. LEAHY (2004), « The Supply of Information by a Concerned Expert », *Economic Journal*, 114, 487–505.

CHANG, L.J., A. SMITH, M. DUFWENBERG, et A. SANFEY (2011), « Triangulating the Neural, Psychological and Economic Bases of Guilt Aversion », *Neuron*, 70 : 560–572.

CHARNESS, G. et M. DUFWENBERG (2006), « Promises and Partnership », *Econometrica*, 74 : 1579–1601.

CHARNESS, G. et M. DUFWENBERG (2011), « Participation », *American Economic Review*, 101 : 1213–1239.

CHARNESS, G. et M. RABIN (2002), « Understanding Social Preferences with Simple Tests », *Quarterly Journal of Economics*, 117 : 817–869.

COX, J.C., D. FRIEDMAN et S. GJERSTAD (2007), « A Tractable Model of Reciprocity and Fairness », *Games and Economic Behavior*, 59 : 17–45.

COX, J.C., D. FRIEDMAN ET V. SADIRAJ (2008), « Revealed Altruism », *Econometrica*, 76 : 31–69.

DANA, J., D.M. CAIN et R.M. DAWES (2006), « What you Don't Know Won't Hurt me: Costly (but Quiet) Exit in Dictator Games », *Organizational Behavior and Human Decision Processes*, 100 : 193–201.

DANA, J., R.A. WEBER et X. KUANG (2007), « Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness », *Economic Theory*, 33 : 67–80.

DUFWENBERG, M. (2002), « Marital Investment, Time Consistency and Emotions », *Journal of Economic Behavior and Organization*, 48 : 57–69.

DUFWENBERG, M. (2008), « Psychological Games », *in* DURLAUF, S.N. et L.E. BLUME (éds.), *The New Palgrave Dictionary of Economics*, 6 : 714–718.

DUFWENBERG, M. et G. KIRCHSTEIGER (2004), « A Theory of Sequential Reciprocity », *Games and Economic Behavior*, 47 : 268–298.

ELLINGSEN, T., M. JOHANNESSON, S. TJØ TTA et G. TORSVIK (2010), « Testing Guilt Aversion », *Games and Economic Behavior*, 68 : 95–107.

ELSTER, J. (1998), « Emotions and Economic Theory », *Journal of Economic Literature*, 36 : 47–74.

FALK, A. et U. FISCHBACHER (2006), « A Theory of Reciprocity », *Games and Economic Behavior*, 54 : 293–315.

FEHR, E. et K. SCHMIDT (1999), « A Theory of Fairness, Competition, and Cooperation », *Quarterly Journal of Economics*, 114 : 817–868.

FORSYTHE, R., J.L. HOROWITZ, N.E. SAVIN et M. SEFTON (1994), « Fairness in Simple Bargaining Experiments », *Games and Economic Behavior*, 6 : 347–369.

GEANAKOPLOS, J., D. PEARCE et E. STACCHETTI (1989), « Psychological Games and Sequential Rationality », *Games and Economic Behavior*, 1 : 60–79.

GLAESER, E.L., D.I. LAIBSON, J.A. SCHEINKMAN et C.L. SOUTTER (2000), « Measuring Trust », *Quartery Journal of Economics*, 115 : 811–846.

GUERRA G. et D.J. ZIZZO (2004), « Trust Responsiveness and Beliefs », *Journal of Economic Behavior and Organization*, 55 : 25–30.

HICKS, J. (1939), « The Foundations of Welfare Economics », *Economic Journal*, 49 : 696–712.

HOFFMAN, E., K. MCCABE et V. SMITH (1996), « Social Distance and Other-Regarding Behavior in Dictator Games », *American Economic Review*, 86 : 563–660.

KAHNEMAN, D., J.L. KNETSCH et R.H. THALER (1986), « Fairness and the Assumptions of Economics », *Journal of Business*, 59 : S285–S300.

KAWAGOE, T. et Y. NARITA (2014), « Guilt Aversion Revisited: An Experimental Test of a New Model », *Journal of Economic Behavior and Organization*, 102 : 1–9.

KHALMETSKI, K., A. OCKENFELS et P. WERNER (2015), « Surprising Gifts – Theory and Laboratory Evidence », *Journal of Economic Theory*, 159 : 163–208.

LOEWENSTEIN, G.F., L. THOMPSON et M.H. BAZERMAN (1989), « Social Utility and Decision Making in Interpersonal Contexts », *Journal of Personality and Social Psychology*, 57 : 426–441.

RABIN, M. (1993), « Incorporating Fairness into Game Theory and Economics », *American Economic Review*, 83 : 1281–1302.

REGNER, T. et N. S. HARTH (2014), « Testing Belief-Dependant Models », Working Paper MPI Jena.

SAMUELSON, P. (1947), *Foundations of Economic Analysis*, Harvard University Press.

TADELIS, S. (2011), « The Power of Shame and the Rationality of Trust », Working Paper UC Berkeley, Haas School of Business.

TANGNEY, J.P. (1995), « Recent Advances in the Empirical Study of Shame and Guilt », *American Behavioral Scientist*, 38 : 1132–1145.

TVERSKY, A. et D. KAHNEMAN (1991), « Loss Aversion in Riskless Choice: A Reference-Dependent Model », *Quarterly Journal of Economics*, 106 : 1039–1061.

VANBERG, C. (2008), « Why do People Keep Promises? An Experimental Test of two Explanations », *Econometrica*, 76 : 1467–1480.